DURBAN UNIVERSITY OF TECHNOLOGY
INYUVESI YASETHEKWINI YEZOBUCHWEPHESHE

Qualitative Classification of Sugar Processing Stream Products by Near Infrared Spectroscopy

Submitted in fulfilment of the requirements for the Degree of Master of Applied Science in Chemistry in the Faculty of Applied Sciences at the Durban University of Technology, Durban, South Africa

Rowena Nadar

2022

Supervisor : Professor R M Gengan

Co-Supervisor : Mr S Walford

Declaration

I, Rowena Nadar, hereby declare that this dissertation titled "Qualitative Classification of Sugar Processing Stream Products by Near Infrared Spectroscopy'', submitted to the Durban University of Technology, in fulfilment of the requirements for the award of the Degree of Master of Applied Sciences, Chemistry, in the Faculty of Applied Sciences, is the result of my own work and that all sources used or quoted have been indicated and acknowledged by means of complete references.

Signed: Rowena Nadar (207 111 64)

Date: 23rd March 2022

Signed: Professor R M Gengan (Supervisor)

Date: 23rd March 2022

Signed: Mr S Walford (Co-Supervisor)

Date: 23rd March 2022

Department of Chemistry
Durban University of Technology

Dedication

This dissertation is dedicated to my late grandparents, “Always Together”.

Acknowledgements

First and foremost, all praise and glory go to my Lord and Saviour, Jesus Christ who granted me the wisdom, courage and strength to undertake and persevere through this research.

I would like to express my sincere gratitude to:

My supervisor, Professor R. M. Gengan, for his time, effort, guidance and constant words of encouragement throughout the duration of this research. I am truly grateful for your unwavering support and belief in me.

My industry supervisor and manager, Mr Stephen Walford. I cannot express enough how grateful I am for your time spent, insightful comments, suggestions and assistance at every stage of the research project. All this whilst being extremely busy in your own work capacity for which I am forever grateful.

Lola Naidoo for your constant encouragement and initiative in assisting me with pursuing this degree.

Drs Kim Booysen and Sanet Nel for your technical support and guidance with this study. Thank you both for always being a calming presence and assuring me that I could do this.

Dr Janice Dewar and all my dear friends at the Sugar Milling Research Institute NPC – Thank you for the support and assistance with pursuing this dream. I appreciate the words of encouragement and especially the assistance from the Analytical laboratory, allowing me the time and space needed to conduct my research.

I would also like to acknowledge the late Andy Sachs – A truly remarkable human being, mentor and friend who has left an invaluable mark on my life. Thank you, Andy, for the effort you put into getting me started with this degree, for always lending an ear and offering advice on a personal and professional level.

Finally, I would like to thank my parents, Rajen and Rita, and my brother Joel, for being a support system, for your continuous prayers and for allowing me the space and time needed to complete this degree. The achievements that you all have made have encouraged and motivated me towards my own goals in life.

Abstract

The Sugar Milling Research Institute NPC (SMRI) is an integral and essential part of the sugar industry as it provides a quality control service among other consultation services to sugar mills in South Africa and other parts of Africa. SMRI uses various prediction equations with near infrared spectroscopy (NIRS), in transmission mode, to predict analyte concentrations present in the various sugar stream products. In this study, chemometrics was used to develop a classification model using discriminant analysis, which could be applied to the process analysis to choose the correct prediction equation for a specific sugar stream product. Samples were selected based on various geographical and environmental factors to ensure variability between the samples. Two different types of data sets were explored to determine the best classification model. The first method used the spectral data of absorbance and wavelength of each sample: Pre-processing was carried out to eliminate any scattering effects. Principal component analysis (PCA) was then applied to reduce the data so that only the necessary information remained. Various classification models, namely, K-nearest neighbour (KNN), Classification tree, Support vector machine (SVM), and Logistic regression, were tested and validated by comparing the predicted sample types against actual sample types. Results showed that the KNN (3) model with the Savitzky Golay filter and three principal components (PCs) provided the best separation between the various sugar stream products. The second method used the analyte concentrations for pol (apparent sucrose content), Brix (total dissolved solids), sucrose, fructose, glucose, and ash for the various sugar stream products. These results were standardised before PCA was applied. The same classification models were applied, tested, and validated using actual samples. These results showed that the Logistic regression model with two PCs performed best. The optimum model from each investigation was compared against each other by evaluating the performance measures of the two models. Based on the analyte concentration data, the Logistic regression (lasso) model with two PCs provided the best separation between sugar stream products. The F1 scores and classification accuracies determined this for the calibration and independent validation sample data set, which were 99.4 and 100 %, respectively.

List of Tables

List of Figures

List of Appendices

List of Abbreviations

Brix	-	Measurement of total dissolved solids
Bx	-	Brix
CA	-	Classification accuracy
CTS	-	South African Sugar Association Cane Testing Service
DAC	-	Direct analysis of cane
EMSC	-	Extended multiplicative scatter correction
FDA	-	Factorial discriminant analysis
FIR	-	Far-infrared
FN	-	False negative
FP	-	False positive
FT	-	Fourier transform
g	-	gram
GC	-	Gas chromatography
HPLC	-	High performance liquid chromatography
IR	-	Infrared radiation
KF	-	Karl Fischer
KNN	-	K-Nearest neighbour
LDA	-	Linear discriminant analysis
LED	-	Light-emitting diode
MSC	-	Multiplicative scatter correction
MIR	-	Mid-infrared
MLR	-	Multiple linear regression
NIR	-	Near infrared
NIRS	-	Near Infrared Spectroscopy
nm	-	nanometer
PC	-	Principal component
PCA	-	Principal component analysis
PCR	-	Principal component regression
PET	-	Polyethylene terephthalate
PLS	-	Partial least squares
PLS-DA	-	Partial least squares discriminant analysis
PLSR	-	Partial least squares regression

Pol	-	Apparent sucrose content
PS	-	Polystyrene
PVC	-	Polyvinyl chloride
RBF	-	Radial basis function
RI	-	Refractive index
RMSECV	-	Root mean square error for cross validation
RMSEP	-	Square root of mean standard error of prediction
RPD	-	Residual predictive deviation
RS	-	Reducing sugars
SECV	-	Standard error of prediction cross-validation
SEP	-	Standard error of prediction
SFDA	-	Stepwise forward discriminant analysis
SIMCA	-	Soft independent modeling of class analogy
SMRI	-	Sugar Milling Research Institute NPC
SNV	-	Standard normal variate
SVM	-	Support vector machine
TP	-	True positive
TPD	-	Target purity differences
%	-	percentage
µm	-	micrometer
λ	-	wavelength
$\tilde{\nu}$	-	wavenumber

Table of Contents

Chapter One: Introduction

1.1 Problem statement and purpose of the study

In the South African sugar industry, there are 14 sugar factories that process sugarcane to produce sugar sold for local consumption or export. Sugar factories utilise several different sugar processing steps to obtain the final product. Near Infrared Spectroscopy (NIRS) can be conveniently used to monitor processes rapidly and identify any process issues which may result in an undesirable final product or by-product. This enables management to adjust the processing conditions, thus ensuring a final product or by-product within the standard specifications (Simpson and Naidoo 2010). The Sugar Milling Research Institute NPC (SMRI) currently uses prediction equations based on NIRS calibrations to predict the quantities of various analytes present in sugar stream products (Walford 2019). The sugar stream products used in this study consist of mixed juice and clear juice, grouped as juices; A-, B-, C-massecuite, A-, B- molasses, and syrup, which are grouped as intermediates as well as final molasses. The tested analytes include pol (apparent sucrose), brix (total dissolved solids), ash, dry solids, glucose, fructose, and sucrose.

The South African sugar industry uses NIRS to quantify sugar parameters; however, the protocol requires the operator to choose the correct sugar stream product. Although the correct information is input in most instances, sometimes the operator inadvertently supplies the incorrect product, resulting in incorrect usage of the prediction equations. Hence incorrect results have been recorded, making it difficult for factory managers to implement correct decisions with processing problems. Furthermore, a culture of supplying reliable results to factory personnel, a prerequisite for the smooth running of the factory, is compromised.

Aim of the study

This study aimed to use chemometric methods to develop a qualitative classification method to select sugar stream products in the NIRS system correctly.

Objectives of the study

The objectives of the study were to:

- Use chemometric methods to develop a qualitative classification method to overcome the problem of incorrect sugar stream products selection. The industry covers a diverse range of geographical, environmental, seasonal and mill variations. These factors will

be included in the development of the chemometric models by using samples from across the whole industry and season,

- Test and validate models by comparing predicted sample types against actual sample types.

1.2 Structure of the Dissertation

This dissertation is presented in four chapters.

Chapter One discusses the problem and states the aim and objectives.

Chapter Two contains a review of the literature on the application of NIRS in the sugar industry, a review of studies in which NIRS is used to distinguish between various products and chemometrics analysis in NIRS. Current literature, as well as prior research in this area, are discussed. The basic theory of spectroscopy, infra-red and NIRS are presented. An overview of the South African sugar industry, its various processes, and the parameters used to run an efficient factory for producing a good quality product is explained. The conventional methods used to determine pol, brix, glucose, fructose, sucrose, conductivity ash, and dry solids are also explained.

Chapter Three presents the research methodology, materials, methods, and instruments used and the spectroscopic and analytical data obtained for each sample.

Chapter Four details the results, discussion, and conclusion. This chapter discusses the statistical analysis of the scans and analyte concentrations of each sample and the results obtained from the discriminate model; the discussion includes the validation of the results and any problems encountered. Finally, this chapter ends with a conclusion and suggestions for further research.

1.5 References

Simpson, R. and Naidoo, Y. (2010). Using near infrared spectroscopy for rapid quantification of intermediate sugar factory products. Proceedings of South African Sugar Technologists' Association. 83, p. 382.

Walford, S. (2019). Near infrared spectroscopy: rethinking the analysis of sugarcane factory streams. Proceedings of International Conference Near Infrared Spectroscopy. 18, p.129-133.

Chapter Two: Literature Review

2.1 South African sugar industry

The South African sugar industry comprises 14 sugarcane processing factories across two provinces in South Africa, namely KwaZulu-Natal and Mpumalanga. It is one of the world's leading producers of high-quality sugar (South African Sugar Association 2020b). The industry comprises the agricultural sector, which grows the cane, and the industrial factory sector, which processes the cane into raw and refined sugar (Sugar Milling Research Institute NPC, no date), sold locally or exported. The farmers are paid based on pol, brix and sucrose results of mixed juice. The results produced in the laboratory must be accurate and reliable, primarily for factory control purposes. The results generated by the laboratory are a decision-making tool used to ensure the profitability and sustainability of the factory. Slight variations in process stream products can be attributed to (a) using sugarcane from various geographical regions, (b) using sugarcane grown under different environmental conditions, (c) seasonal changes, and (d) mill variations. Each factory has unique processes that make their products' characteristics slightly different from each other in terms of the analytical results produced.

2.1.1 General process of sugar production

Sugar is produced in two forms, namely raw and refined sugar. In South Africa, all 14 factories produce raw (brown) sugar. Of these 14 factories, at least four further processes and decolourises the sugar into refined (white) sugar (Sugar Milling Research Institute NPC 2020). A typical flow diagram of a sugar mill processing raw sugar is shown in Figure 2.1.

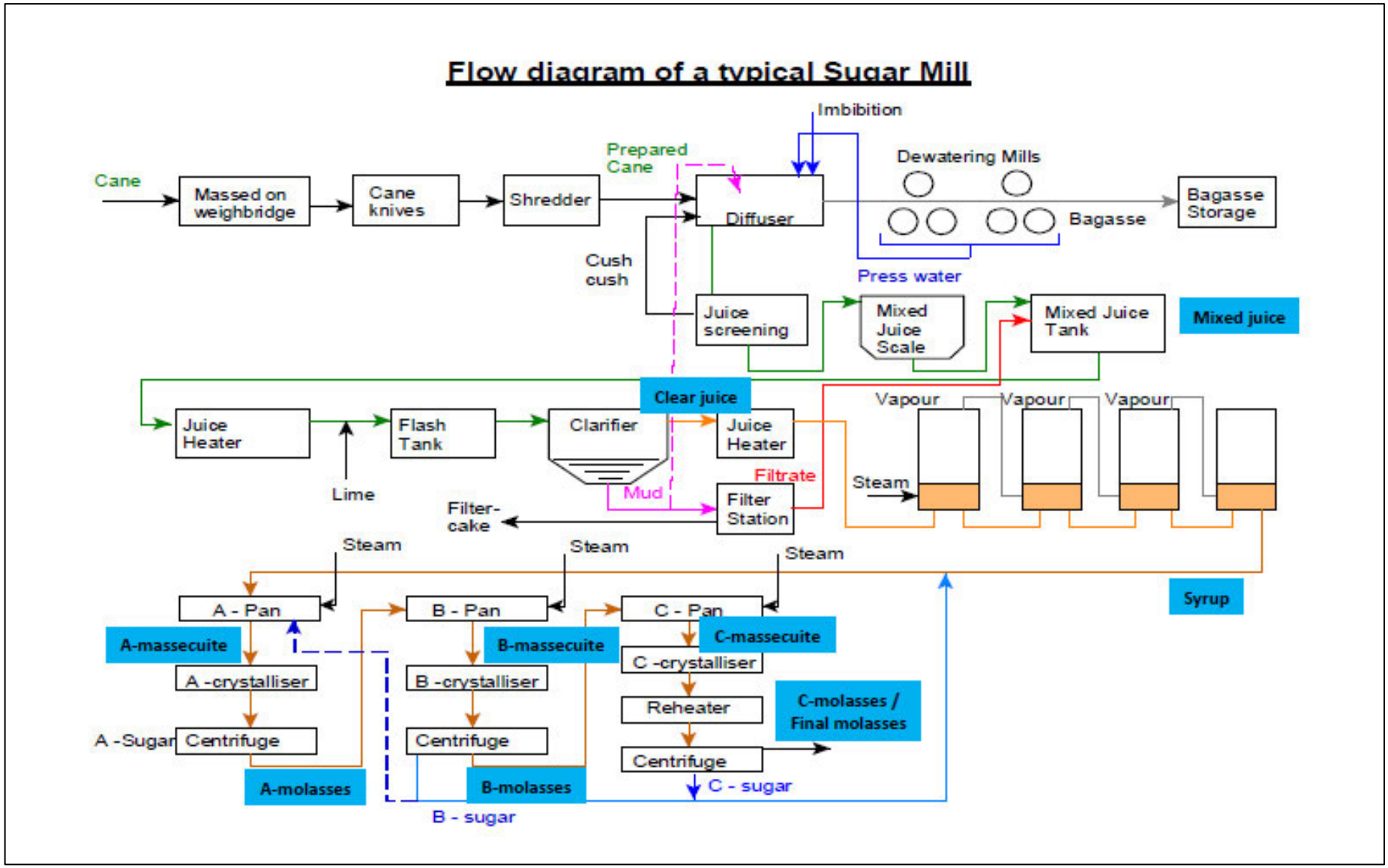

Figure 2.1: A flow diagram of a typical raw house sugar mill (Sugar Milling Research Institute NPC 2020)

Part 1: The process of manufacturing raw sugar

A. Cane preparation

The sugarcane stalks are cut and shredded using heavy preparation equipment, reducing it to a fine consistency.

B. Extraction of juice

The juice is extracted from the cane by washing it out in diffusers or squeezing and washing in the mill. The juice extract has a 10 to 15 % sucrose concentration which is sugar (Honiron 2017). The leftover cane is called bagasse, and this is usually burned in the factory boilers to generate steam and electricity.

C. Juice clarification

The extracted juice is purified by heating the juice and then adding the chemical lime to the process. This causes mud to be formed, which is left to settle in vessels called clarifiers. The mud is removed from the bottom of the clarifier and filtered to extract whatever juice is remaining. This juice is returned to the process.

D. Evaporation

The juice that comes out of the clarifier is termed clear juice and consists of approximately 90% water (Sugar Milling Research Institute NPC 2020). The evaporator's task is to remove most of the water. This creates a high viscosity syrup which has approximately 35% of water.

E. Crystallisation of sucrose into sugar crystals

The syrup is further concentrated in vacuum pans into which seed crystals are added. Sucrose molecules from the syrup grow onto these seed crystals until they have reached the required size. The vacuum pan is then discharged, and the mixture of the large sugar crystals and syrup is sent to the centrifuge. This mixture is now known as A-massecuite. The A-massecuite is spun in a centrifuge at a very high rate until the sugar crystals (A-sugar) separate from the mixture. The remaining liquid is termed A-molasses since it contains less sucrose. The A-sugar is then dried and despatched or kept for further processing into refined sugar. The A-molasses still hold a large amount of sucrose; therefore, they are sent back to the vacuum pans, where they undergo the same process to produce B-massecuite. The B-massecuite is sent to the centrifuge, where it is separated into

B-sugar and B-molasses. Similarly, the B-molasses still contain a portion of sucrose and are sent back into the process to produce C-sugar and final molasses. Final molasses cannot be reprocessed as the recovery of any more sugar is not feasible. B- and C- sugar is then remelted and sent back to the boiling house to start the process again.

Part 2: The process of manufacturing refined sugar

The goal of a refinery is to remove soluble, insoluble and colloidal impurities from raw sugar resulting in very high purity of refined sugar (Rein 2017c). These impurities are caused by brown colour, ash, and dextran, among others. This is done by melting the raw sugar and then removing the melt colour using various clarification methods. A standard method is carbonatation followed by sulphitation (Sugar Milling Research Institute NPC 2020; Tongaat Hulett 2016), while some refineries prefer phosphatation and ion exchange. Finally, water is evaporated from the resulting clarified and decolourised melt, leading to a refined sugar product.

2.1.2 Analytes of importance relating to sugar quality

There are multiple process streams involved in the manufacturing of raw and refined sugar. The in-house factory laboratory tests these streams to ensure that the final product is of a good standard. The various process streams and their analytes that are required for process control are shown in Table 2.1. For this thesis, mixed juice and clear juice are grouped as juices. A-, B-, C– massecuite, A-, B- molasses and syrup are grouped as intermediates. Results of intermediate products and final molasses are generally used for factory control, while juices are commonly used to determine payment to farmers.

Table 2.1: Process streams and analytes determined

Process stream	Analyte						
	Brix	Pol	Sucrose	Glucose	Fructose	Ash	Dry solids
Mixed Juice	✓	✓	✓	✓	✓	✓	
Clear Juice	✓	✓	✓	✓	✓	✓	
Syrup	✓	✓	✓	✓	✓	✓	
A-massecuite	✓	✓	✓	✓	✓	✓	
B-massecuite	✓	✓	✓	✓	✓	✓	
C-massecuite	✓	✓	✓	✓	✓	✓	
A-molasses	✓	✓	✓	✓	✓	✓	
B-molasses	✓	✓	✓	✓	✓	✓	
Final molasses	✓	✓	✓	✓	✓	✓	✓

2.1.2.1 Definition of analytes

- Brix is defined as the measurement of total dissolved solids in juices, syrups, massecuites, molasses, or sugar using a refractometer (Rein 2017a, Rein 2017d). The solids concentration by mass of a sucrose-containing solution is mathematically determined based on the relationship between the refractive index and the percentage by mass of total soluble solids of a pure aqueous sucrose solution at 20 °C (South African Sugar Technologists' Association 2005).
- Pol is the apparent sucrose content of the sample determined by a polarisation method. It is expressed as a mass percent measured by the optical rotation of polarised light passing through a sugar solution. This value will be an accurate measure of sucrose for pure sucrose solutions. For a solution containing sucrose and other optically active substances, the pol denotes the algebraic sum of the rotations of the components present (Rein 2017a, Rein 2017d).
- Sucrose is a pure chemical compound, a disaccharide comprising glucose and fructose, known as $C_{12}H_{22}O_{11}$. It is commonly known as white sugar and is determined by polarisation (if in a pure form) (Rein 2017b) or by gas chromatography (GC) or high-performance liquid chromatography (HPLC) in an impure solution.

- Reducing sugars, also known as invert sugars, usually refer to glucose and fructose in the sugar milling industry. Glucose and fructose are monosaccharides. (Rein 2017d)
- Conductivity ash measures the concentration of ionised soluble salt in samples by measuring the conductivity of the solution (Rein 2017a).
- Dry solids is also known as total solids, dry substance or solids determined by evaporating the sample or solution until dryness is reached. The mass of the total moisture deducted from the mass of the product gives the mass of a dry substance (South African Sugar Technologists' Association 2005).

2.2 Conventional methods of analyses

An overview is given on different selected analytical methods that have been conventionally used to obtain results for sugar stream products. The general information of all the methods reviewed is presented in Table 2.2. All methods are written up as controlled documents requested from the Sugar Milling Research Institute NPC (SMRI). The methods are assigned an SMRI reference number.

Table 2.2: Conventional methods used for the analysis of sugarcane processing stream analytes

Process stream	Analyte	Name of method	Instrument used	SMRI Reference
Mixed / Clear Juice	Brix	Determination of the refractometer brix in juice (First expressed juice, first mill juice, mixed juice)	Refractometer	TM005
	Pol	Determination of the polarisation (pol) of juice with lead	Polarimeter	TM042
	Fructose Glucose Sucrose (F, G & S)	The determination of glucose, fructose, and sucrose in cane mixed juice by gas chromatography	Gas chromatograph (GC)	TM300
	Ash	The determination of conductivity ash in juices and molasses	Conductivity meter	TM066

Table 2.2: (Continued)

Process stream	Analyte	Name of method	Instrument used	SMRI Reference
Intermediates (Syrup A-massecuite B-massecuite C-massecuite A-molasses B-molasses)	Pol and brix	Determination of the polarisation (pol) and brix of intermediate products	Polarimeter and refractometer	TM312
	F, G & S	Determination of sucrose, glucose, and fructose by High-performance Liquid Chromatography on intermediate products *TM300 is used for the F, G and S analysis of syrup	HPLC	TM311
Syrup	Ash	Determination of the conductivity ash in syrup	Conductivity meter	TM012
Final molasses	Pol	Determination of the polarisation (pol) of molasses	Polarimeter	TM043
	Brix	Determination of the refractometer brix of molasses	Refractometer	TM007
	F, G & S	Determination of sucrose, glucose and fructose by HPLC	HPLC	TM310
	Ash	The determination of conductivity ash in juices and molasses	Conductivity meter	Refer to TM066
	Dry Solids	Determination of the moisture/dry solids in final molasses using the Karl Fischer titration method	Karl Fischer titrating apparatus	TM030

Description of selected analytical methods

1. Determination of Conductivity ash: The sample is prepared for a concentration of 5 g/100 cm^3 or less. The specific conductivity of this solution is then measured and calculated using a conventional factor.

2. Brix: A solution of the sample and distilled water is made. This is then filtered and the filtrate measured on a refractometer using a brix scale. This scale measures the density of a pure sucrose solution with 0 °Bx on the brix scale referring to pure water and 100 °Bx referring to pure sucrose. The drawback of the brix scale is the lack of accuracy due to being greatly affected by the number of impurities and other non-sucrose components. In most cases, the brix scale has been replaced by the refractive index (RI) measurement, which is related to the pure sucrose content. Although the RI is also accurate for pure sucrose content, the non-sucrose components influence the RI similarly to sucrose. The RI is thus a better measurement of total dissolved solids containing sucrose (Rein 2017d). It is referred to as refractometer brix.

3. Pol: This method measures the apparent sucrose content in juices, molasses, and intermediates. Sucrose is optically active, and the concentration is determined by measuring the angle of rotation of polarised light when passed through the solution. Sucrose solutions rotate the light to the right and are therefore dextrorotatory. Glucose is also dextrorotatory, while fructose is levorotatory (Kvittingen and Sjursnes 2020). Glucose thus adds to the dextrorotatory rotation created by sucrose. Fructose will decrease this value. Pol is thus defined as apparent sucrose as the glucose and fructose (and other optically active compounds) can cause a change in the actual polarimeter reading and thus sucrose value.

4. Fructose, Glucose and Sucrose (F, G and S) in mixed juice, clear juice and syrup using Gas Chromatograph (GC): Processing stream analytes are first made volatile by a technique known as silylation. Then, a flame ionisation detector is used to detect ions in a gas stream. Trehalose is also added as an internal standard for F, G and S. An electronic integration system estimates peak areas of F, G and S.

5. F, G and S in intermediate products and final molasses using HPLC: The chromatographic separation of F, G and S occurs by HPLC using a sodium or calcium-based cation exchange resin as the separation column. The diluted, filtered sample is injected onto the column using a sample loop injector system. The sugars elute at varying times and this, together with its peak height, is used for the final calculation.

6. Dry solids in final molasses determined using the Karl Fischer (KF) method: Moisture determination is done electrochemically using a KF reagent. This is conducted using an amperometric titration method.

2.3 Near infrared spectroscopy

NIRS is a technique whereby a multivariate calibration based on NIRS scans is developed to predict analytes relatively quickly and requires little or no sample preparation or reagents (Blanco and Villarroya 2002). It is also accurate and can analyse various types of matrices. In order to understand the principle of NIRS, it is necessary to discuss its background and evolution as well.

2.3.1 Background to NIRS

Electromagnetic radiation is described as the electromagnetic field waves and is defined by the wavelength or frequency of the waves (Petrucci *et al.* 2007). The wavelength (λ), expressed in nanometers, is the distance between two crests or troughs of a wave. In infrared spectroscopy, the wavelength is converted to a wavenumber ($\widetilde{v}$) expressed in cm^{-1}. The wavenumber is the number of wavelengths per unit distance. Mathematically the wavelength is converted to wavenumber by multiplying the reciprocal of the wavelength in nm by 10^7. The frequency is defined as the number of wavelengths per unit time and is calculated as:

$$\boldsymbol{v = {}^{c}/_{\lambda}} \qquad \textbf{… Eqn 1}$$

The energy of a single photon (electromagnetic radiation) is calculated as (Van der Merwe 2018):

$$\boldsymbol{E = h \times v = \frac{h \times c}{\lambda} = h \times c \times \tilde{v}} \qquad \textbf{…Eqn 2}$$

Where: $\boldsymbol{E}$ **= energy** $\boldsymbol{h}$ **= Planck constant in** $\mathbf{m^2.kg.s^{-1}}$

$\tilde{\boldsymbol{v}}$ **= frequency of the light in** $\mathbf{s^{-1}}$ $\boldsymbol{c}$ **= speed of light in** $\mathbf{m.s^{-1}}$

$\boldsymbol{\lambda}$ **= wavelength of the light in m** $\tilde{\boldsymbol{v}}$ **= wavenumber of the light in** $\mathbf{m^{-1}}$

According to Equation 2, it can be deduced that the shorter the wavelength, the higher the wavenumber and the higher the energy of the photon (Metrohm 2013). The electromagnetic

spectrum, shown in Figure 2.2 comprises of waves of varying frequencies and thus divided into different regions.

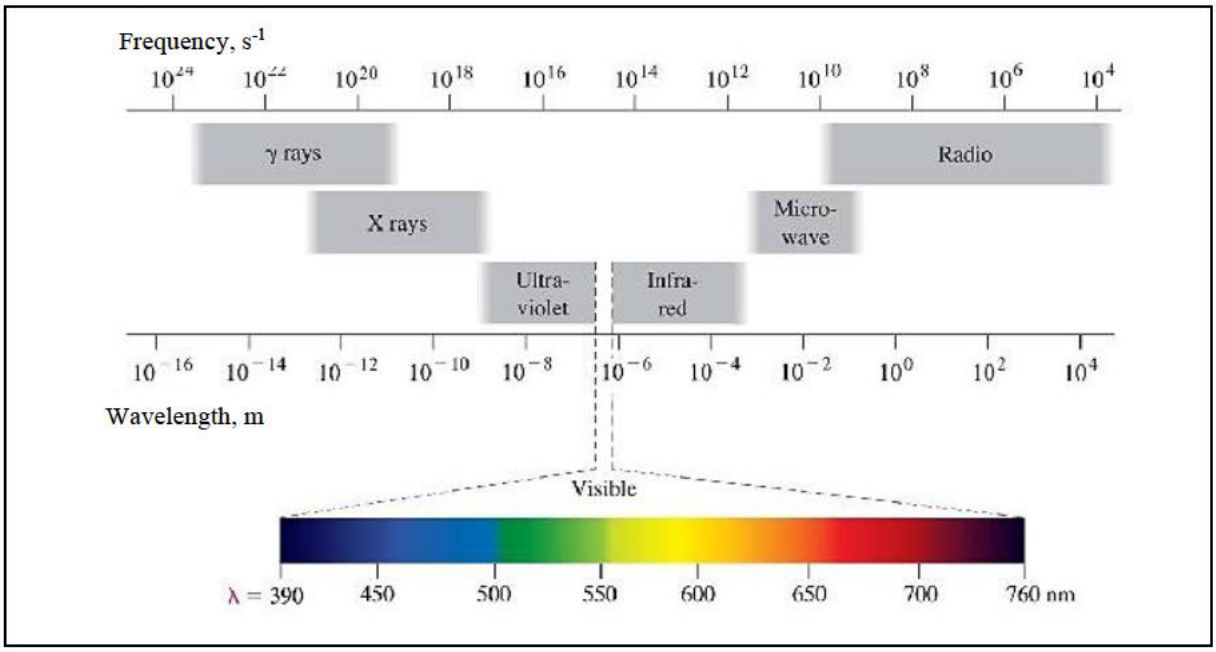

Figure 2.2: Electromagnetic spectrum (Petrucci *et al.* 2007)

Each region has a unique molecular or atomic transition making it applicable to specific spectroscopic techniques as shown in Figure 2.3.

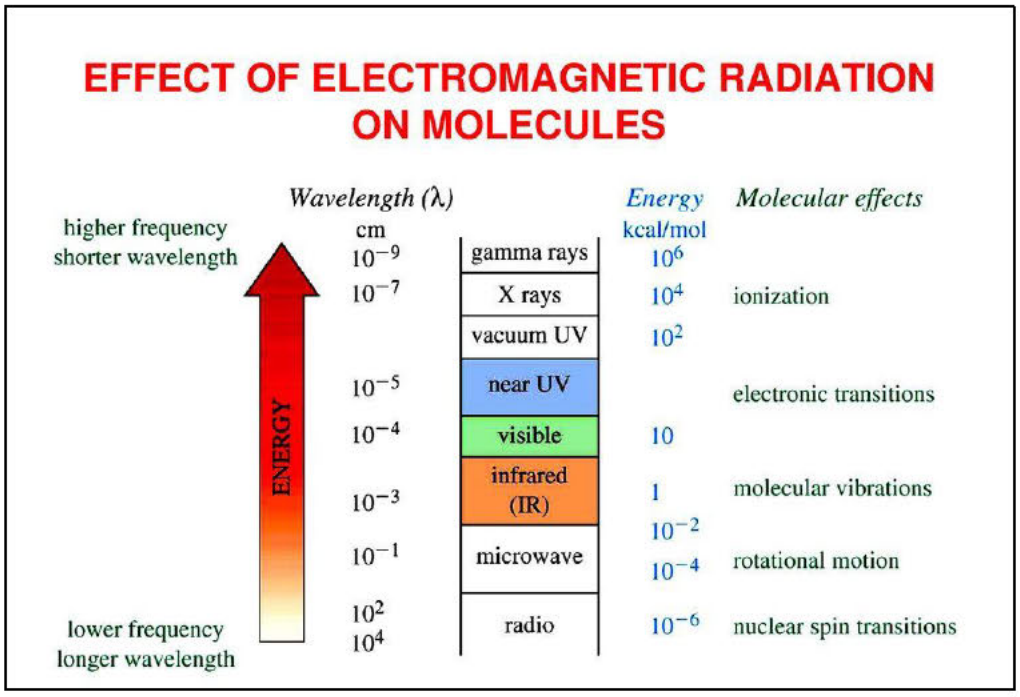

Figure 2.3: Electromagnetic spectrum showing molecular effects (Shukla 2018)

Spectroscopy is based on the interaction between light, also known as electromagnetic radiation and matter, depicted in Figure 2.4. When this interaction occurs, the radiation can be absorbed, reflected, scattered, transmitted, or undergo photoluminescence (Chong 2020). This energy absorption is detected and recorded as a spectrum, enabling the user to identify further and analyse the matter and its constituents (Metrohm 2013).

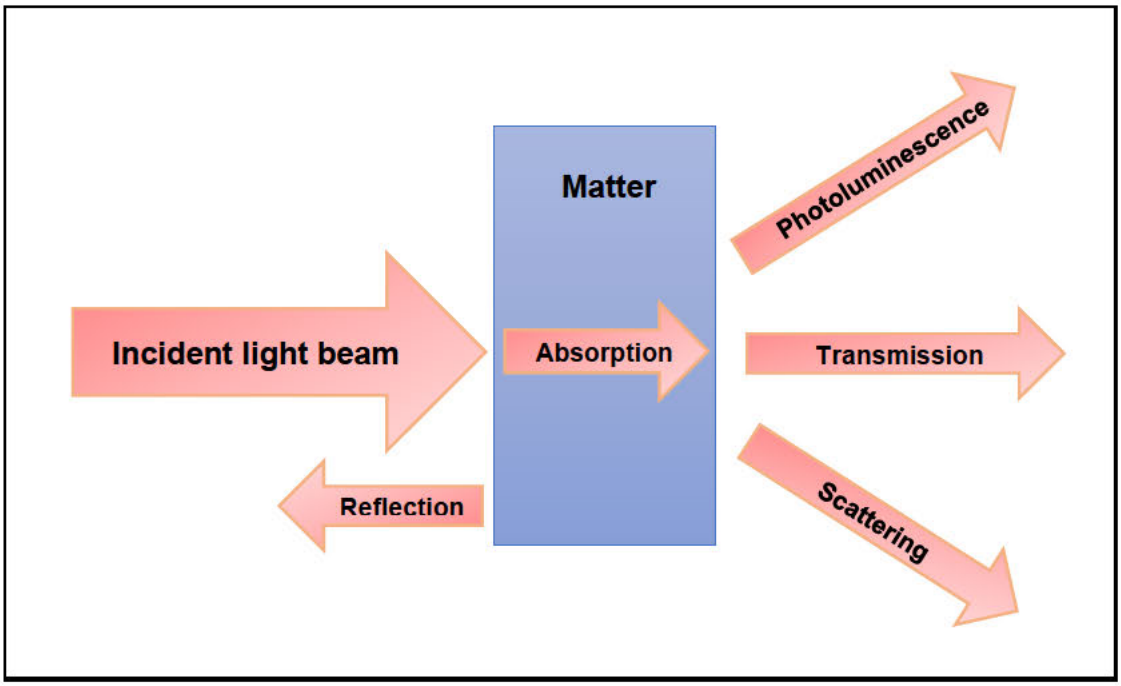

Figure 2.4: Schematic representation of the interaction of radiation and matter

Frederick William Herschel first discovered infrared radiation (IR) in the year 1800 (Pasquini 2003). Herschel had created a rainbow of colours by passing sunlight through a glass prism. This produced a spectrum. When checked with a blackened bulb thermometer, he noted that each colour corresponded to a different temperature. Of more incredible surprise, a higher temperature was detected when Herschel placed the thermometer in a non-coloured region adjacent to the red coloured part of the spectrum. This region then became IR, separated into three bands based on their wavelength range and different transitions. These bands are near-infrared (NIR: 750 to 2500 nm); mid-infrared (MIR: 2500 to 50000 nm), and far-infrared (FIR: 50- 1000 μm). IR spectroscopy is a technique based on the absorption of IR light by the substance to be measured. This absorption causes the molecules to vibrate and rotate (Patel 2017) resulting in an IR spectrum. The spectrum is then capable of providing information about the sample and its constituents. NIRS can also be termed vibrational spectroscopy as the vibrational frequencies play a significant role in sample information (Blanco and

Villarroya 2002). There are different types of vibrations observed in molecules. These are termed stretching, scissoring, rocking, twisting, and wagging (Nawrocka and Lamorska 2013). Figure 2.5 depicts the different vibration modes.

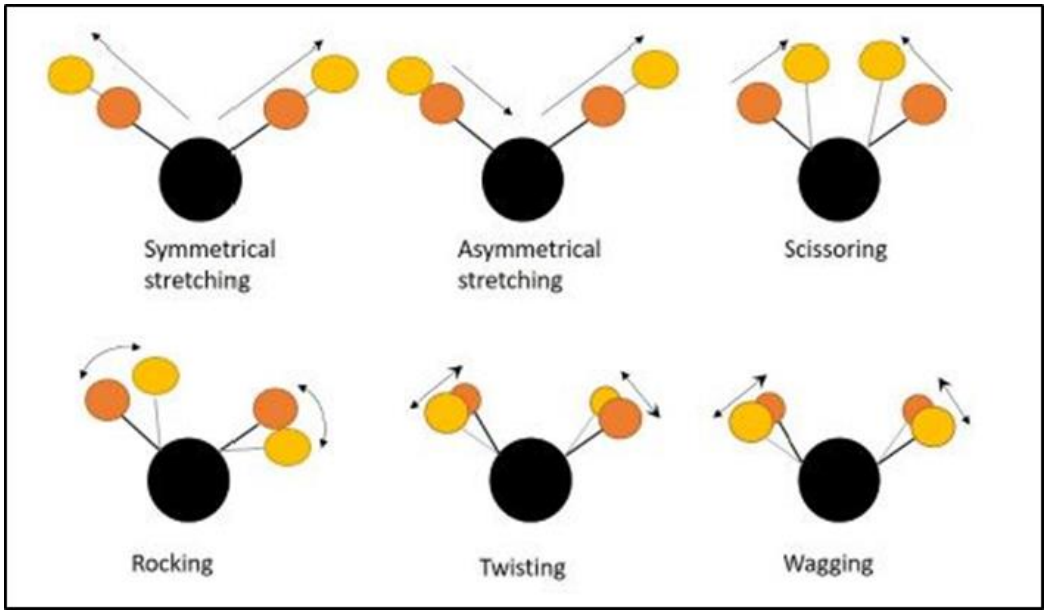

Figure 2.5: Different modes of molecular vibrations [Adapted from (Nawrocka and Lamorska 2013)

When molecules with covalent bonds respond to radiation in the IR range, the absorbed energy causes a change in the electric dipole moment (Blanco and Villarroya 2002). This allows molecules to transition from one vibrational level to another since the molecule will only absorb wavelengths with energies corresponding to the energy difference of the present transition. Figure 2.6 shows the different vibrational levels a molecule can transition.

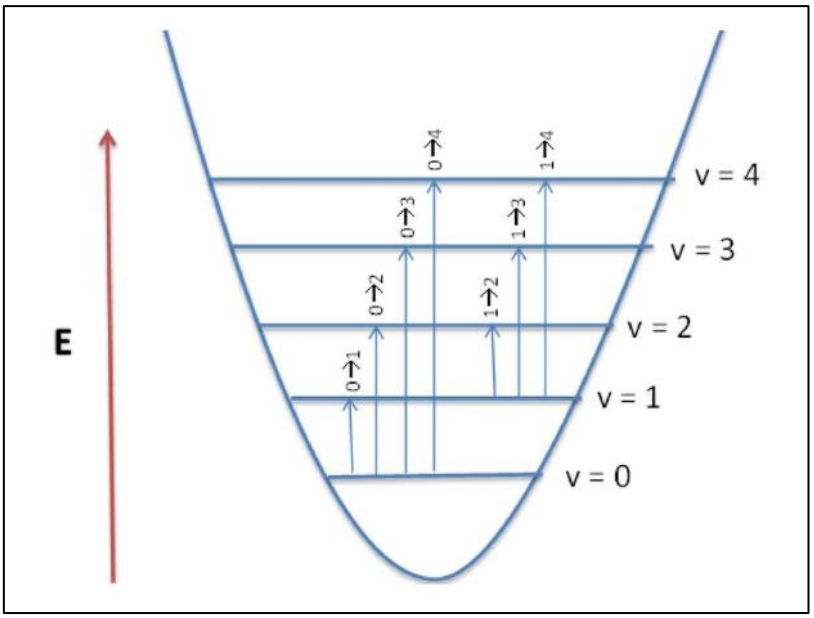

Figure 2.6: Transitions between different vibrational levels (Shukla 2018)

A spectrum is observed when molecules absorb energy. The absorptivity of the matrix and the concentration of the sample influences the spectral response. The Beer-Lambert law shown in Equation 3 explains this relationship. The law states that the absorbance of a compound as measured by a spectrophotometer is equivalent to the absorptivity of the compound (ε), the concentration of the compound (C) and the path length (L) of the sample container.

$$A = \varepsilon \times C \times L \quad \text{…Eqn 3}$$

2.3.2 Principle of NIRS

A transition from the ground state (v0) to the first vibrational state (v1) is called a fundamental transition. Overtones are identified to transition from the ground state to the second, third, or fourth excited state. The intensity of an overtone from v0 to v2 is stronger than for a transition from v0 to v3. Thus, the response band for the overtone from v0 to v2 will appear at twice the energy for a fundamental band (OPUS 2016). The subsequent overtones will appear at multiples of the energy (wavenumber) for the fundamental band. This is represented in Figure 2.7. A combination band is what occurs when more than one chemical bond in the molecule is excited simultaneously.

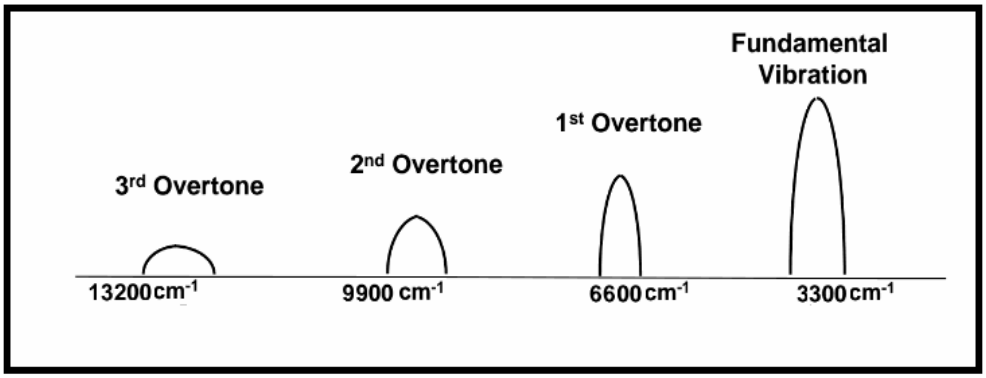

Figure 2.7: Fundamental and overtone transitions. A compound with the fundamental transition at 3300 cm^{-1} has overtones at multiples of 3300 cm^{-1}. For example, the first overtone occurs at 6600 cm^{-1} (2x 3300 cm^{-1}), the second overtone occurs at 9900 cm^{-1} (3 x 3300 cm^{-1}) (OPUS 2016)

The MIR range carries fundamental vibrations. In the NIR range, the overtone and combination vibrations are excited. The FIR region has low energy and can be used for low-frequency vibrations (OPUS 2016).

NIR spectroscopy is a technique based on the interaction of NIR light with a sample containing the functional groups: O-H, C-H, C=O and N-H. The vibrations of these bonds are observed as absorption bands in the NIR region (Nawrocka and Lamorska 2013). Due to the chemical bonds in these functional groups, NIRS can be used to study and measure organic samples. Thus, instead of individual compounds, major functional groups are allotted to the NIRS region. This is depicted in Figure 2.8.

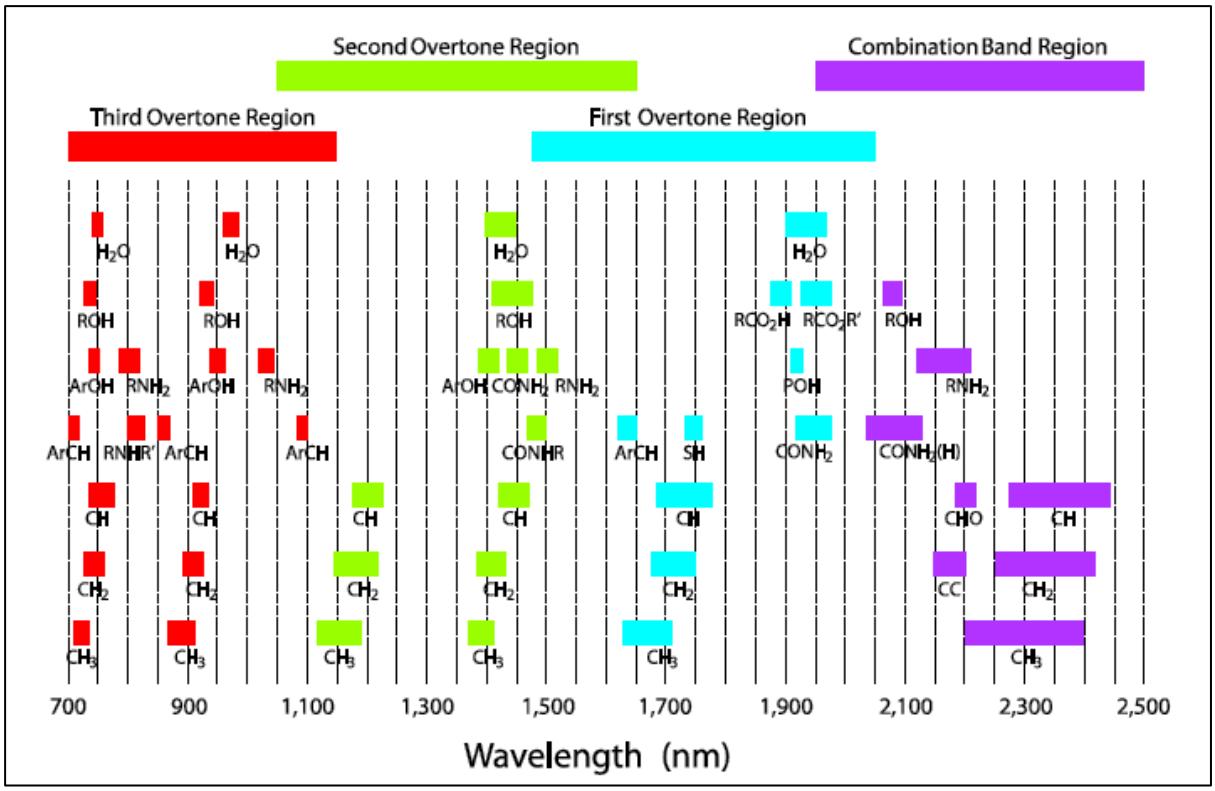

Figure 2.8: Diagram depicting overtone and combination bands for different chemical bonds in the NIR range (Metrohm 2013)

2.3.3 NIRS measurement modes

When a sample interacts with NIR, the radiation is reflected, absorbed, or transmitted by the molecular bonds. The resultant NIR spectrum is then recorded in transmission, reflection or transflection modes. The spectra then allow for a qualitative and quantitative investigation of the sample.

The type of mode used is dependent on the optical properties of the sample. Each mode also has unique instrument modifications to suit the application. The reflectance mode is usually used for solid samples (Blanco and Villarroya 2002). The radiation is reflected from the sample surface towards the detector for measurement. Figure 2.9 shows a sample in reflectance mode.

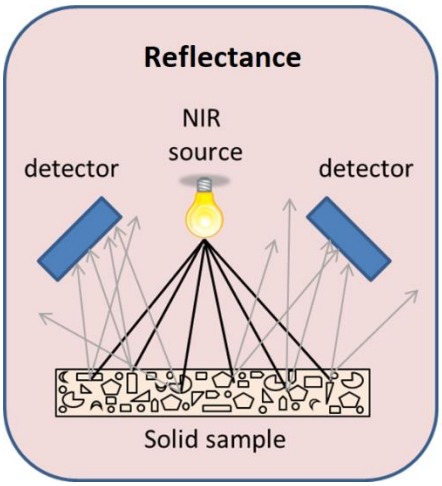

Figure 2.9: A sample in reflectance mode (Walford 2013)

Transmission mode is for liquid samples (Blanco and Villarroya 2002). Radiation passes through a sample and is collected by the detector to determine how much light was absorbed by the sample. Figure 2.10 illustrates a sample in transmission mode.

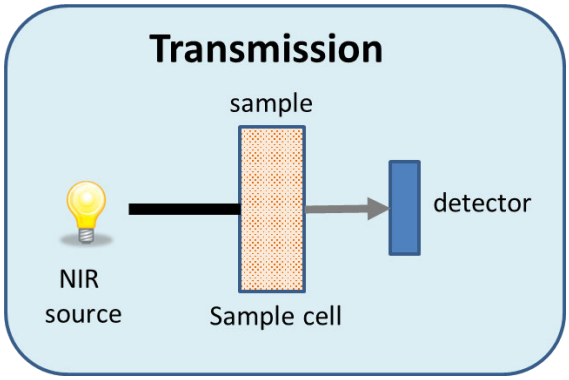

Figure 2.10: A sample in transmission mode (Walford 2013)

A sample analysed in transflectance mode is usually an emulsion or turbid liquid type (Blanco and Villarroya 2002). Transflection works by transmitting the radiation through a sample; however, the radiation gets reflected by an object such as a mirror or metal plate and is transmitted back through the sample onto the detector. This is depicted in Figure 2.11.

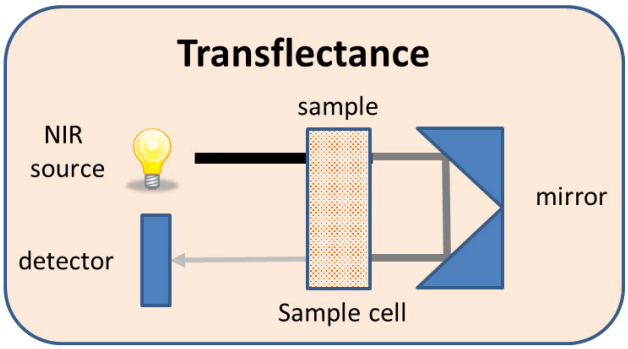

Figure 2.11: A sample in transflectance mode (Walford 2013)

2.3.4 NIRS instrumentation

A typical NIRS instrument consists of a light source, monochromator, and detector, which creates a spectrum on the attached computer (Patel 2017). The spectrum is analysed by various means of statistical software. This is depicted in Figure 2.12.

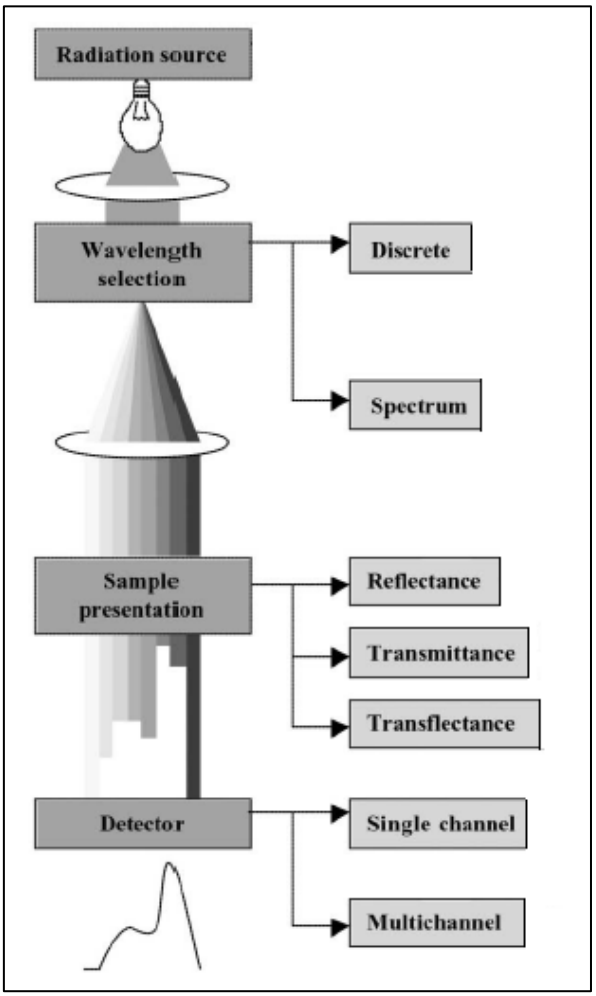

Figure 2.12: Components of a typical NIRS instrument (Blanco and Villarroya 2002)

The light source is usually a single polychromatic thermal type such as a tungsten halogen lamp, or in some cases, a light-emitting diode (LED) can be used.

A NIR spectrophotometer can be categorised as a discrete wavelength or whole spectrum (dispersive) instrument. Discrete wavelength systems pass light through a sample with only a few wavelengths. This limits its application to analytes which will only absorb in certain spectral regions. A whole spectrum instrument is based on diffraction grating and Fourier transform (FT) instruments. This system is more commonly used as it displays good signal-to-noise ratios and can be used in various applications. In diffraction grating instruments, the light passes through a sample and is then split into its component frequencies (Payne 2019). FT-NIR instruments operate by using the interferometer to split the NIR beam into two. Each beam is reflected onto a fixed and moving mirror and then recombined to produce an interferogram signal, measured on a detector. The most common detectors used in NIRS make use of silicon, lead sulfide (PbS), and indium gallium arsenide (InGaAs) photoconductive materials (Pasquini 2003). A mathematical transformation of the interferogram is achieved using FT to produce the NIRS absorption spectrum.

2.3.5 Advantages and disadvantages of NIRS

Advantages

a) The technique is non-destructive and straightforward to operate (Patel 2017),
b) It is rapid, efficient, and requires little or no sample preparation,
c) Avoids the need for potentially harmful reagents, hence an environmentally friendly technique,
d) Multiple components can be analysed simultaneously,
e) Cost-effective as there are few to no reagents required,
f) Calibrations can be reliable and precise and
g) It enables many samples to be analysed in less time.

Disadvantages

a) NIRS is a secondary method that is dependent on the results of the primary analysis method in order to develop a calibration for the quantification of analytes of interest,
b) Requires separate calibrations for structures of compounds (Patel 2017) and
c) It is not suitable for analysis of very dark samples.

2.4 Chemometrics

The complex NIR spectra usually contain interferences such as background data and noise together with the sample data. A technique is thus required to extract all valuable information about the chemical properties of the sample. This technique is called chemometrics which can be regarded as a subset of machine learning or pattern recognition. Machine learning aims to extract the important information from raw data and develop algorithms, using mathematical or statistical methods, that can predict information about a sample from new data points (Torrione, Collins and Morton 2014). In Barra *et al*. (2021), the application of chemometrics and machine learning techniques coupled with infrared spectroscopy was used to determine the physical and chemical properties of soil. There are many different levels of data analysis techniques available. These include univariate, bivariate, and multivariate analysis. Univariate analysis refers to analysing one spectral data point to predict one variable (Conzen 2006). Bivariate data contains two variables that explain the causes and relationships between them. Chemometrics is used to analyse multivariate data which is the most complex form. It is used when more than two sets of data may not necessarily have a relationship between them. Instead, a cause or relationship is investigated amongst many variables. Due to the complexity of multivariate analysis, statistical techniques such as factor analysis, cluster analysis, discriminant analysis, principal component analysis (PCA), and multiple logistic regression are used to depict the relationships between the variables. NIRS data is multivariate. Multivariate analysis is performed on multiple measurements, wavelengths, samples, and data sets to predict multiple variables. An example of this is shown in Table 2.6.

Table 2.3: Example of NIRS absorbance vs. wavelength data in the form of a multivariate table for multiple samples scanned by an NIRS instrument

Absorbances						
	Wavelength (cm^{-1})					
Sample ID	**12493**	**12489**	**12485**	**12481**	**12478**	…
1	0.26162	0.26153	0.26139	0.26118	0.26103	…
2	0.18799	0.18794	0.18782	0.1877	0.18767	…
3	0.20537	0.20543	0.20537	0.20522	0.20512	…
4	0.29346	0.29326	0.29307	0.2928	0.29255	…
5	0.31512	0.3149	0.31474	0.31456	0.31427	…
…	…	…	…	…	…	…

The objective of chemometrics is to: (a) identify patterns or trends in the data, (b) design and use multivariate classification models, and (c) analyse, model, classify and predict data.

There are various models or multivariate techniques used to achieve the objectives laid out above. The multivariate techniques can be grouped under the following categories (Chemometrics understand chemical data 2020):

a) Exploratory Analysis – Identifies any outliers and determines patterns or trends in the data. PCA is a typical example of reducing sizeable complex data sets into a series of optimised and interpretable data sizes (Geladi 2003).

b) Regression analysis – This model is used for quantitative analysis whereby a model is created which correlates the information in the set of known measurements to the desired property. Common techniques include partial least squares regression (PLSR), principal component regression (PCR), and multiple linear regression (MLR).

c) Classification analysis – Used to assign predefined categories to samples and predict an unknown sample belonging to one of these distinct groups (Biancolillo and Marini 2018). Common methods include Soft Independent Modeling by Class Analogy (SIMCA), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Partial Least Squares Discriminant Analysis (PLS-DA).

2.4.1 Spectral pre-processing

In order to identify the correct model, a spectral pre-processing event needs to take place. This involves weeding out the redundant information within the data matrices so that only the relevant data remains. This will then improve the subsequent classification models. There are one of three objectives for the pre-processing step. These are to improve (a) a subsequent exploratory analysis, (b) a subsequent bi-linear calibration model or c) a subsequent classification model. Pre-processing allows the multivariate signals to conform to the Beer-Lambert law. The physical and chemical phenomena that cause deviations from this rule include light scatter from particulates in the sample, interferents, molecular interactions, shifts in the chemical equilibrium as a function of concentration, stray light, and changes in sample size or path length. Pre-processing techniques have been designed to counteract these deviations, thus improving the linear relationship between the spectral signals and the analyte

concentrations. There are two types of pre-processing techniques: Scatter-correction and spectral derivatives (Rinnan, van den Berg and Engelsen 2009). These methods provide a smoothing of the spectra before calculating the derivative, which will decrease the signal-to-noise ratio. Scatter correction methods includes Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC (EMSC), Extended Inverse MSC, de-trending, Standard Normal Variate (SNV), and normalisation. The most commonly used methods, MSC; SNV and normalisation, are designed to decrease the physical variability between samples due to scattering. Spectral derivatives comprise of two methods: Norris-Williams and Savitzky-Golay derivation. These derivatives remove both additive and multiplicative effects in the spectra (Mark and Workman 2003).

2.4.2 Exploratory analysis

Exploratory analysis analyses a group of samples and depicts their similarities or differences by looking at their chemical or physical composition. A commonly used technique is PCA.

2.4.2.1 Principal component analysis (PCA)

PCA is the workhorse of multivariate techniques as its generally the basis for other classification and regression methods (Munck and Møller 2005). It is a variable-reduction technique whose output is a visual representation of the information in a data table. PCA identifies the differences between each sample, which variable is responsible for this difference, and if these variables act in the same way or are independent of each other. Based on these variances, similar groupings or patterns can then be observed (Biancolillo and Marini 2018). It is also valuable in quantifying the valuable data and separating it from the meaningless data.

The new variables that are produced are known as principal components (PCs). These PCs are ordered so that the most data variance present in the original variables is retained within the first few PCs. The PCA results are visualised as loadings and scores, which allows the data space to be correctly interpreted. Loading refers to the contribution of each variable to a PC. A score is the amount of PC present in a sample and is visualised as the distance from the origin to the projection of a point to the PC. Figure 2.13 and 2.14 depicts a score and loading, respectively.

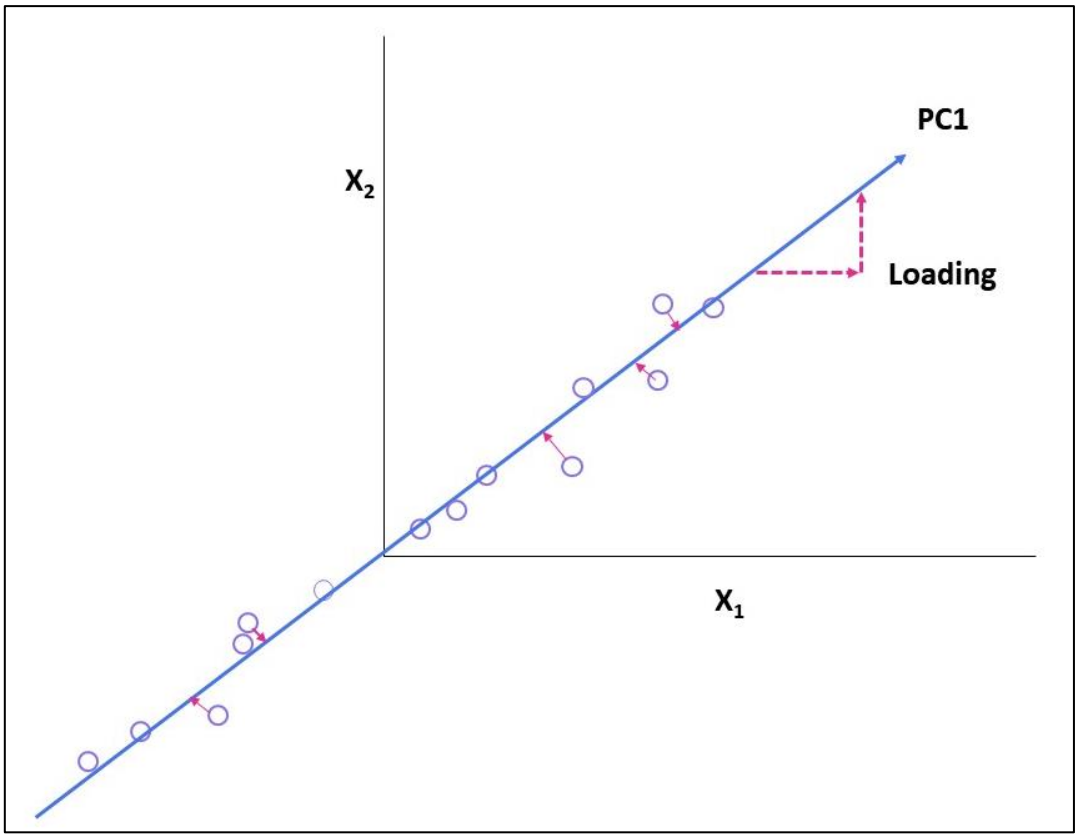

Figure 2.13: A loadings plot

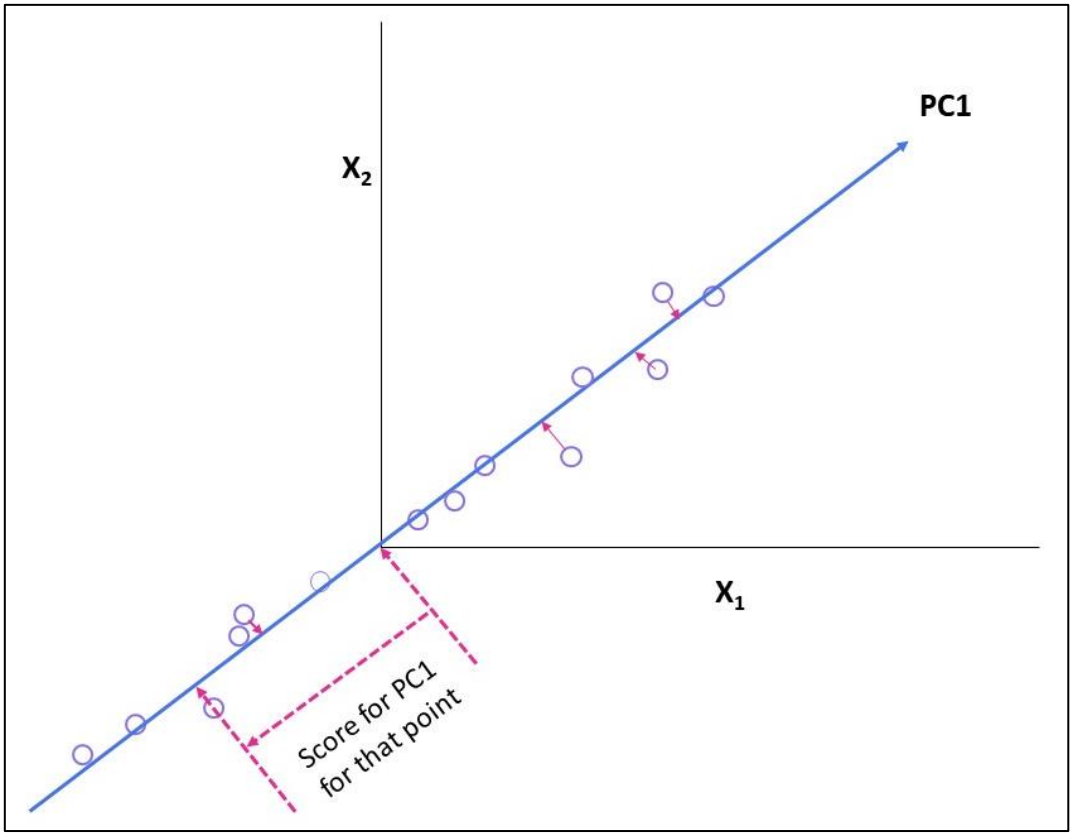

Figure 2.14: A scores plot

A scores plot generally illustrates the correlation between two types of samples, while the loadings plot depicts the variation within the variables. These two plots simultaneously give analytical insight into the data present by depicting the presence of a variation and then identifying the variable responsible for that variation (Payne 2019).

2.4.3 Regression analysis

The principle of quantitative analysis is that the amount of analyte present can be determined from a signal or response of the instrument. This signal should change based on the quantity of the analyte. Regression analysis is used to design a calibration model that will correlate data between two groups of variables. The model can either quantify analytes by way of prediction or describe the relationship between the two groups of variables. To ensure a robust model that can accurately predict future unknown samples, the model should be developed and validated on a wide range of samples that cover a variety of variables.

2.4.3.1 Partial least square regression (PLSR)

PLSR is the most common regression technique as it deals with multicollinearity; noisy variables allow for interpretation of results by depicting results visually and can model several variables simultaneously. The objective of PLSR is to predict Y values from X values, with the general idea being to calculate the PCs of the X and Y data separately. The regression model will then be based on the scores and not the original data.

2.4.4 Classification analysis

When interpreting data for a qualitative output, it is necessary to define unsupervised and supervised methods. Samples that are classified without prior information except for the spectra obtained fall under the unsupervised methods. PCA falls under the unsupervised technique. A supervised method classifies data according to known information. The most common techniques are SIMCA, SVM, LDA, and PLS-DA. Classification analysis can also be differentiated into two sub-sections: Discriminant and class-modeling. The discriminant technique operates by segregating the hyperspace into as many regions as the number of investigated categories. For example, if there are three categories of samples, the hyperspace will be divided into three regions. Whichever region a new sample falls in, it will then be allocated to the corresponding category.

On the contrary, class-modeling techniques create a separate model for each category or class that is being investigated. Thus, some regions can have more than one class or no class at all.

Figure 2.15 illustrates the difference between discriminant (A) and modeling (B) classification techniques.

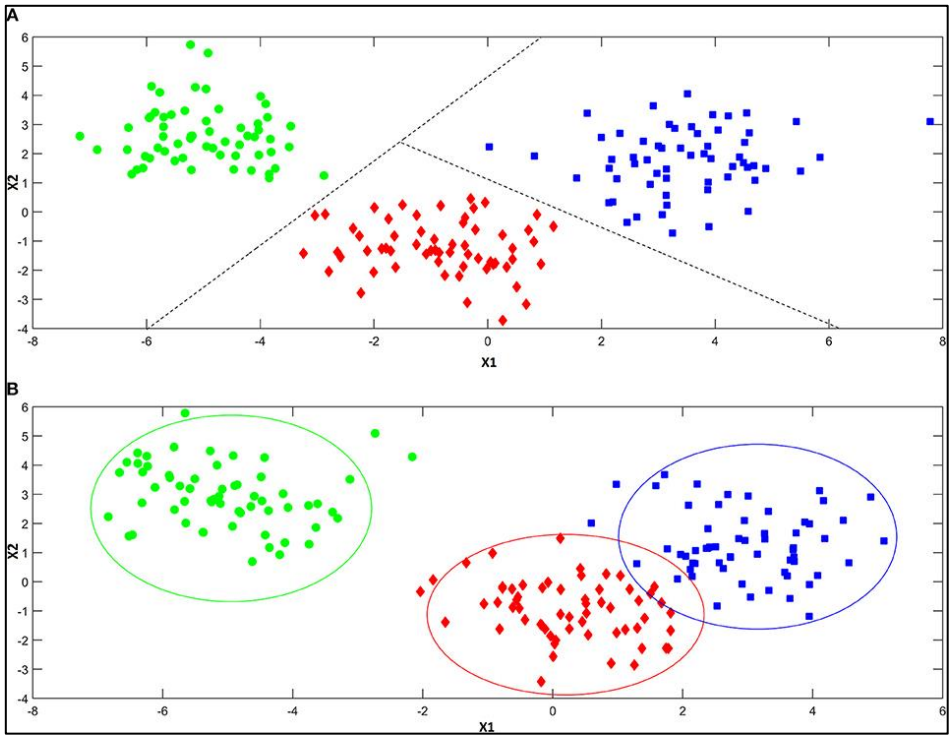

Figure 2.15: A depiction of the difference between discriminant (A) and modeling (B) classification techniques (Biancolillo and Marini 2018)

2.4.4.1 Discriminant methods

As mentioned above, each sample can be assigned to a specific region in the application of discriminant analysis. This is done by defining decision surfaces that allow for clear demarcation of the boundaries. The boundaries can be linear or more complex. The linear models are usually more common as they have fewer parameters to tune, require a smaller number of samples for the model, and are more robust in predicting new samples.

2.4.4.1.1 Linear discriminant analysis (LDA)

LDA is based on the postulation that samples from each class are distributed around their respective mean value with the same variance/co-variance matrix. Thus, an LDA model can predict which class the sample is assigned to by whichever class has the highest probability (Williams 2007).

2.4.4.2 Class modeling

Class modeling creates models where it is likely to find samples with similar characteristics. The class space is defined by identifying a normal variability expected amongst samples of the same category. There is also a "distance-to-the-model" standard that describes the degree to which a new sample lies in the model.

2.4.4.2.1 Soft independent modeling of class analogies (SIMCA)

A typical modeling technique is SIMCA, as it can work with poorly conditioned experimental data matrices. SIMCA operates by developing PCA models for each class. Then, when a new sample is suspected of falling into one of these classes, two measurements are used to decide which class the sample falls into. These are Euclidean distance, which is the distance of the sample to its projection onto the PC space, and Mahalanobis distance, which is the distance of spectral data of the individual sample from the population mean (Williams 2007).

2.4.5.1 General NIRS validation statistics

The following provides an overview of the usual information used to provide a sound statistical evaluation of the model quality (Agelet and Hurburgh 2010):

- The coefficient of determination (R^2) estimates how much variance between reference and predicted values is explained versus the total variance.

- The standard error of prediction (SEP, or SECV for cross-validation results) reports on the calibration precision. Cross-validation is performed when, due to a small sample range, the same samples are used for validation and calibration.

- The square root of the mean, standard error of prediction (RMSEP) for test sample validation, or root mean square error for cross-validation (RMSECV) when cross-validation is used: These describe the prediction error of a calibration model by providing an approximate average uncertainty expected for predictions of future samples.

- Residual predictive deviation (RPD): This refers to the ability of the model to predict future samples based on the initial variability of the calibration data. The RPD value is the ratio of the standard deviation of the response variable to the RMSEP or RMSECV.

- F-test: An estimate of the goodness of fit of spectral and constituent data. It is also helpful to determine how many factors should be used during regression and choose which samples to eliminate as outliers.

- The slope and bias are used to determine the accuracy and linearity of the calibration model. The bias is calculated as a systematic error between the predicted NIRS results and the results of the reference conventional methods. If the slope value is close to one and the bias is close to zero, this indicates that the deviations are distributed normally (Metrohm 2013).

2.5 Application of NIRS and chemometrics

In 1968 Karl Norris and his team first applied NIRS technology to agricultural products. Grains were being adsorbed in specific NIR regions. This information was then used to measure the protein, oil, and moisture of the grains. Computers were also used in conjunction with the NIRS system. The NIRS system was a rapid, non-destructive analysis across various industries, as shown in Table 2.7.

Table 2.4: Major fields of application of NIRS and selected examples of the parameters determined (Blanco and Villarroya 2002)

Applications of NIRS	
Agricultural food sector	**Petrochemical sector**
❖ Composition	❖ Petroleum fractions
➢ Grain: % moisture and protein	➢ Light alkenes mixtures
➢ Milk and dairy products	➢ Physical properties of bitumen
➢ Fruits and vegetables	❖ Fuels
❖ Adulteration	➢ Chemical composition and additives
➢ Beef hamburgers containing other types of meat	➢ Gasoline octane index, Diesel cetane index
➢ Orange juice with sugar	➢ Physical properties
❖ Quality	❖ Polymers
➢ Industrial vs slow-growing chicken	➢ Physical properties
➢ Fruit ripeness	➢ Polymerization monitoring
➢ Crop infections	

Table 2.4: (Continued)

Applications of NIRS	
Pharmaceutical sector ❖ Manufacturing steps ➢ Raw materials ➢ Intermediate products ➢ End products • Active principles • Excipients ❖ Intact tablets ❖ Packaged products	**Clinical sector** ❖ Human serum ➢ Glucose ➢ Proteins ❖ Tumours ❖ Tissue oxygenation
Environmental sector ❖ Recycling: Characterisation of plastics (Polyvinyl chloride (PVC), Polyethylene terephthalate (PET), Polystyrene (PS)) ❖ Soil contamination ➢ Motor oils ➢ Fuels	**Miscellaneous** ❖ Process control ❖ Textiles ➢ Cotton fibres ➢ Finishing oils in fibres ❖ Leather tanning

2.5.1 NIRS application in the sugar industry

NIRS technology with chemometrics is a widely used application in the sugar industry worldwide. In Mauritius, a reliable and rapid NIRS method was developed to predict cane quality (Koonjah *et al*. 2019). The sucrose and fibre content analysis via NIRS was now less than a minute with better analytical precision and less sample preparation. Three consecutive models were developed. Each one proved better than the previous model as the number of samples in the calibration set increased. The final model was developed using 3000 samples for Brix % cane, Pol % cane, and fibre % cane. SNV and Savitzky – Golay were used for spectral pre-treatment and calibration was done using PLS (Koonjah *et al*. 2019). The first model resulted in 17 % outliers. The robustness of the final model was shown in the decrease of outliers to only 5 %. Sample selection is critical when developing models for NIRS application. The samples need to reflect the various agro-climatic zones for sugar cane production. The large the sample size in a calibration set, the better chance of encompassing all sample varieties.

In Brazil, a model was created to classify four Brazilian sugarcane varieties (Steidle Neto 2018). Previous studies had used NIRS and hyperspectral data to differentiate cane varieties based on the leaves of the sugar cane stalk. This did not prove successful. Steidle Neto investigated the use of cane stalks instead of the leaves. This model was developed using PCA, factorial discriminant analysis (FDA), stepwise forward discriminant analysis (SFDA) and PLS-DA based on the reflectance mode (Steidle Neto 2018). The spectra were pre-treated by centering, normalisation, and second order derivatives. This improved the accuracy of the models. PCA showed that of the four cane varieties, two had overlapped resulting in three distinct groups. The model was thus based on the three groups. PLS-DA provided to be the most effective classification method with a correct classification 82 % (Steidle Neto 2018). The wavelength region of 600 to 750 nm was examined and further work is to be undertaken by investigating other regions as well as increasing the sample size. An increase in sample size may result in a distinction between the two overlapped groups as more variety will be included.

In another case, NIRS, in conjunction with PLSR was used to predict ethanol content in Aguardente, a Brazilian spirit produced from fermented sugarcane juice (De Carvalho *et al.* 2016).

In Indonesia, the cane payment system is directly based on cane juice quality. NIRS was investigated as an alternative to the laborious conventional methods for pol and brix content (Kuswurjanto and Triantarti 2019). A PLS calibration model, using spectra that was pre-treated by SNV, was developed. This resulted in a R^2 of 98.60 % for pol and 98.80 % for brix. The results proved that NIRS has great potential to be an appropriate alternative to conventional analysis.

Currently, the South African industry uses NIRS to predict the processing stream parameters. Calibration equations have been determined for final molasses and mixed juice (Schaffler and De Gaye 1997). Approximately 550 mixed juice and 900 final molasses samples, encompassing four seasons of sugar production, were used to in the development and validation of NIRS equations for pol, brix, fructose, glucose, sucrose and ash for both products. An equation for dry solids was also developed for final molasses samples. PCA was applied to the spectra of the samples and thereafter PLSR was used to develop the calibrations. Results show that the NIRS is capable of predicting the aforementioned analytes. Further work was done to determine the accuracy of these calibrations. Simpson and Oxley (2008) focused on using NIRS, in a sugar factory environment, as an alternative to conventional methods for factory control purposes. The calibration equations developed by Schaffler and De Gaye (1997) were used to predict

results for mixed juice and final molasses samples. Approximately 430 mixed juice samples were analysed conventionally by the SMRI for pol, brix, fructose, glucose, sucrose and conductivity ash. The same samples were also analysed by NIRS. These samples were sent by 22 different South African and African sugar mill factories over the course of the sugar production season, on a weekly basis. This ensures variability within the samples encompassing general weather conditions, geographical areas and different factory processes. Likewise, approximately 430 molasses samples were received from 25 South African and African mills over the course of the sugar season. The samples were analysed conventionally for pol, brix, fructose, glucose, sucrose, conductivity ash and dry solids. The spectra and conventional results were added to the calibration equations on a weekly basis. This ensures robustness of the equations. For mixed juice, the NIRS results were within 0.06 % of the conventional results. Final molasses comparisons between NIRS and conventional results predicted well except for the sucrose calibration equation. Equations for both mixed juice and final molasses samples improved over time as more spectra were added to the equations. A quantitative study (Simpson and Naidoo 2010) was conducted to determine analytes of intermediate sugar stream products. Approximately 300 clear juice samples and sample sets of syrups, A-, B-, C-massecuites, A- and B molasses were received from five different geographically located sugar factories over three seasons. Variations due to geographical, environmental, and other conditions were included in the prediction sets. These samples were analysed using the final molasses calibration equations developed by Shaffler and De Gaye (1997) and Simpson and Oxley (2008). PLSR was then employed to develop intermediate equations for pol, brix, fructose, glucose and sucrose. Spectra was continually added to the equations making them more robust. The SEP for glucose results of A-massecuite products was 0.15 % and 0.19 % for the fructose results of C-massecuites (Simpson and Naidoo 2010). This value can be decreased by adding more samples to improve the correlation between NIRS and conventional results. The SEP for pol in B-mol products was high at 1.55 %. This was attributed to analytical sampling errors. Crystals found in the B-molasses samples caused sample inhomogeneity. Overall, the correlation co-efficient for most of the products for most of the analytes was greater than 0.9 (Simpson and Naidoo 2010). This proved that NIRS could give reliable results for A-, B- and C-massecuites, A- and B- molasses, clear juice, and syrup. Rapidly receiving fructose and glucose results for juices and syrups can quickly expose potential inversion problems across evapourators. Calibrations for juices, intermediate and final molasses products will benefit the sugar factories as the reliable data produced can be used to combat problems in a fraction of the time. In addition, research has been conducted on raw and refined sugar (O'Shea *et al.* 2011),

resulting in the potential to be quickly developed into a routine technique for sugar mill applications. These studies, as mentioned earlier, confirm that NIRS calibrations and adequate equations have been developed for quantification. However, the current system requires the correct process stream analyte to be manually chosen before analysis in order for the instrument to use the correct prediction equation. This manual operation is a potential source of error by the operator, which may lead to incorrect results being used in factory personnel's decision-making.

2.5.1.1 SMRI-NIRS verification scheme

The SMRI has been providing the various South African sugar industry mills with annually updated prediction equations for juices, intermediates, and final molasses. The SMRI-NIRS technology, based on transmission mode, has already been widely accepted in the sugar industry. A proficiency scheme has been developed to uphold the SMRI confidence in the prediction equations' ability to improve factory performance, thus producing a quality product. The proficiency scheme was created to ensure the instrument is functioning correctly, that the analysts are using the technology correctly, and continually update the existing SMRI-NIRS prediction equations (Nadar and Walford 2019).

2.5.1.2 SMRI-NIRS Toolkits

Many mills across South Africa use the NIRS results to report factory performance parameters identified for factory recovery improvements. These include the daily reporting of sucrose inversion losses and target purity differences (TPD) on final molasses and molasses from individual C-centrifuges (Barker and Madho 2019; Gounden and Walthew 2018), profiles of pH; colour; ash and reducing sugars, and various ratios. NIRS can be conveniently used to monitor the quality and quantity of C-massecuite, final molasses and C-sugar. The amount of sucrose lost in these areas can be ascertained and adjustments made to the process to minimise this loss. This ensures a factory operating with optimum efficiency.

2.5.2 Applications in other industries

Literature shows that NIRS can be used to distinguish between various products. For example, NIRS application in the cotton industry (Gaitan-Jurado *et al.* 2008) and in coal blends (He *et al*. 2011) show that it is an excellent classification method and is feasible for industrial applications. Gaitan-Jurado explained the importance of sampling by taking cotton sub-samples from various parts of the truck in which the cotton was transported. The sub-samples were then

mixed to form one sample. 100 cotton samples were taken over thirty eight days. This allows for variability within the samples and ensures a robust calibration model. In straw-coal blends, the NIRS was shown to be able to distinguish between coals and straw and blends of this nature. Chemometric analyses such as PCA was used to isolate different copper ore content (Gaydon, Glass and Pascoe. 2009) and differentiate between types of beer (Li *et al.* 2009). SNV was applied to the spectra of beer samples before PCA. This provided an improved differentiation between the beers when MLR was applied (Li *et al.* 2009). PCA determines the components of variance in a system. The use of NIRS coupled with PCA was used in a study of gelatinised starch. Using PCA, an outlying sample was found and identifying the importance of spectral variations due to the effect of scattering. The scatter variations were thus eliminated (Bertrand and Scotter 1992). A partial least squares (PLS) discriminant method was used as a classification method in the distinguishing of flexo printed and offset printed newspapers (Pigorsch 2010). PLS finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Discriminant analysis was used to develop a method to distinguish tissues paper made from virgin fibre and a mixture of recycled and virgin fibre (Xin *et al.* 2014). Thirty eight samples were used to develop a calibration model. Twenty nine samples formed the prediction set. PLS-DA gave a 100 % discrimination rate in both the calibration and prediction set. NIRS was also used in classifying olive oils from different geographical regions (Sinelli *et al*. 2008). Chemometrics analysis and NIRS were used for the correct prediction of the geographical region of each olive oil. This information is vital for our investigation since we will use a similar classification model for samples in various regions in South Africa.

2.6 References

Agelet, L.E. and Hurburgh, C. (2010). A Tutorial on Near Infrared Spectroscopy and Its Calibration, Critical Reviews in Analytical Chemistry, 40(4), p246-260.

Barker, B. and Madho, S. (2019). Monitoring and managing losses across the C-station using the SMRI- NIRS Technology. Proceedings of South African Sugar Technology Association. 92 (1), p28.

Barra, I.; Haefele, S.M.; Sakrabani, R.; Kebede, F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. Trends in Analytical Chemistry. 135 (116166).

Bertrand, D. and Scotter, C.N.G. (1992). Application of Multivariate Analyses to NIR Spectra of Gelatinized Starch. Applied Spectroscopy. 46 (9), p1420-1425.

Biancolillo, A. and Marini, F. (2018). Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. Frontiers in Chemistry. 6 (576).

Blanco, M. and Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. TrAC Trends in Analytical Chemistry, 21(4): p240-250.

Chong, N. (2020). Principles of Spectroscopy. Available: https://www.ugpti.org/smartse citations/downloads/Chong-Principlesz_of_Spectroscopy_CHEM6230-2013.pdf.
(Accessed: 9 November 2020).

Conzen, J. (2006). Multivariate Calibration. 2nd ed. Germany: Bruker Optik GmbH. p1.

De Carvalho, L. C.; Morais, C. L. M.; De Lima, K. M. G.; Junior, L.; Nascimento, P. a. M.; Faria, J. B. and Teixeira, G. (2016). Determination of geographical origin and ethanol content in Brazilian sugarcane spirit using near-infrared spectroscopy coupled with discriminant analysis. Analytical Methods, 8(28), p5658-5666.

Gaitan-Jurado, A. J.; Garcia-Molina, M.; Pena-Rodriguez, F. and Ortiz-Somovilla, V. (2008). Near Infrared applications in the quality of seed control. Journal of Near Infrared Spectroscopy. 16 (1), p421-429.

Gaydon, J.; Glass, H. and Pascoe, R. (2009). Method for Near Infrared sensor-based sorting of a copper ore. Journal of Near Infrared Spectroscopy. 17 (1), p177-194.

Geladi, P. (2003). Chemometrics in spectroscopy. Part 1. Classical chemometrics. Spectrochimica Acta Part B. 58 (1), 767–782

Gounden, T. and Walthew, D. (2018). NIRS as a tool for improved process monitoring. Proceedings of South African Sugar Technology Association. 91 (1), p350- 356.

Honiron. (2017). Sugar Processing – Juice Extraction, Clarification and Concentration. Available: https://www.honiron.com/sugar-processing-juice-extraction-clarification concentration/#:~:text=From%20a%20consistency%20of%2010,containing%2055%2D59%20percent%20sucrose. (Accessed: 10th October 2020).

He, C.; Yang, Z.; Huang, G.; Chen, L. and Han, L. (2011). A feasibility study on using Near Infrared Spectroscopy to classify straw-coal blends. Journal of Near Infrared Spectroscopy. 19 (1), p277-284.

Koonjah, S.; Beekharry, A.; Badaloo, M.; Henderson, C. and Saumtally, A. (2019). Evaluation of Near Infrared Spectroscopy for the Direct Analysis of Cane Quality Characters. Universal Journal of Agricultural Research. 7 (5), p169-176.

Kuswurjanto, R and Triantarti. (2019). Study on application of near infrared (nir) spectroscopy for sugar cane juice analysis to replace conventional analysis methods. IOP Conference Series: Earth and Environmental Science. 355 (012059), 1-7.

Kvittingen, L. and Sjursnes, B. (2020). Demonstrating Basic Properties and Application of Polarimetry Using a Self-Constructed Polarimeter. Journal of Chemical Education. 97, p2196 - 2202.

Li, H.; Takahashi, Y.; Kumagai, M.; Fujiwara, K.; Kikuchi, R.; Yoshimura, M.; Amano, T.; Lin, J. and Ogawa, N. (2009). A Chemometrics approach for distinguishing between beers using Near Infrared Spectroscopy. Journal of Near Infrared Spectroscopy. 17, p69-76.

Mark, H. and Workman, J. (2003). Derivatives in Spectroscopy. Spectroscopy. 18 (4), p32-37.

Metrohm. (2013). Monograph - NIR Spectroscopy. A guide to near-infrared spectroscopic analysis of industrial manufacturing processes. Herisau, Switzerland: Metrohm AG.

Munck, L. and Møller, B. (2005). Principal Component Analysis of Near Infrared Spectra as a Tool of Endosperm Mutant Characterisation and in Barley Breeding for Quality. Czech Journal of Genetics and Plant Breeding. 41 (3), p89-95.

Nadar, R. and Walford, SN. (2019). The SMRI-NIRS Technology: Proficiency and quality assurance scheme. Proceedings of South African Sugar Technology Association. 92 (1), 35.

Nawrocka, A. and Lamorska, J. (2013). Infrared spectroscopy. In: Grundas, S and Stepniewski, Advances in Agrophysical Research. London: IntechOpen. p24-26.

OPUS. (2016). Germany: Bruker Companies.

O'Shea, M. G.; Staunton S. P.; Donald, D. and Simpson, J. (2011). Developing Laboratory Near Infra-Red (NIR) Instruments for the Analysis of Sugar Factory Products. Proceedings of the Australian Society of Sugar Cane Technology. 33 (1), p1-8.

Pasquini, C. (2003). 'Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications', Journal of the Brazilian Chemical Society, 14, p. 198- 219.

Patel, H. (2017). 'Near Infrared Spectroscopy: Basic principles and use in tablet evaluation', International Journal of Chemical and Life Sciences, 6.02, p. 2006.

Payne, K. (2019). 'Rapid differentiation of South African game meat using portable near-infrared (NIR) spectroscopy', MSc Thesis, Stellenbosch University, Stellenbosch.

Petrucci, R; Harwood, W; Herring, F and Madura, J (2007). General Chemistry Principles and Modern Applications. 9th ed. New Jersey: Pearson Education, Inc. p278-280.

Pigorsch, E. (2010). Classification of offset and flexo printed newspapers by near infrared spectroscopy. Journal of Near Infrared Spectroscopy. 18, p225-229.

Rein, P. (2017a). Cane Sugar Engineering. 2nd ed. Berlin, Germany: Bartens. p32.

Rein, P. (2017b). Cane Sugar Engineering. 2nd ed. Berlin, Germany: Bartens. p35.

Rein, P. (2017c). Cane Sugar Engineering. 2nd ed. Berlin, Germany: Bartens. p549.

Rein, P. (2017d). Cane Sugar Engineering. 2nd ed. Berlin, Germany: Bartens. p714.

Rinnan, Å.; van den Berg, F. and Engelsen, S. (2009). Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry. 28 (10), p1201-1222.

Schaffler KJ and De Gaye MTD. (1997). Rapid estimation of multi-components in mixed juice and molasses: The possibility of day-today control of raw sugar factories using NIR. Proceeding of the South African Sugar Technology Association 71: 153-60.

Shukla, N. K. (2018). Rotational-Vibrational Spectroscopy. Available: https://slideplayer.com /slide/12663774/. (Accessed: 5th December 2020).

Simpson, R. and Naidoo, Y. (2010). Using Near Infra-Red Spectroscopy for Rapid Quantification of Intermediate Sugar Factory Products. Proceedings of the South African Sugar Technologists Association. 83 (1), p382-391.

Simpson, R. and Oxley, J. (2008). Routine analysis of Molasses and Mixed Juice by NIR Spectroscopy. Proceedings of the South African Sugar Technologists Association. 81 (1), p245- 265.

Sinelli, N.; Casiraghi, E.; Tura, D. and Downey, G. (2008). Characterisation and Classification of Italian virgin olive oils by near- and mid- infrared spectroscopy. Journal of Near Infrared Spectroscopy. 16 (1), p335-342.

South African Sugar Association. (2020). World's Top Competitive Producers of High Quality Sugar. Available: https://sasa.org.za/market-competitiveness/. (Accessed: 9 September 2020).

South African Sugar Technologists' Association. (2005). CHAPTER 1 Definitions and important formulae used in sugar factories. Available: https://sasta.co.za/laboratory-manual/. (Accessed: 12th February 2020).

Steidle Neto, A. J.; Lopes, D. C.; Toledo, J. V.; Zolnier, S. and Silva, T. G. F. (2018). Classification of sugarcane varieties using visible/near infrared spectral reflectance of stalks and multivariate methods. J. Agric. Sci., 1-10.

Sugar Milling Research Institute NPC. (2020). The Manufacture of Raw and Refined Sugar in SA. Available: http://www.smri.org/rawsugarfacts.php. (Accessed: 9th September 2020).

Sugar Milling Research Institute NPC (no date). The South African Sugar Industry. Available: http://www.smri.org/include/sugarfacts/sasugarindustry.htm. (Accessed: 5th September 2020).

Tongaat Hulett. (2016). Sugar Manufacture Process. Available: http://www.huletts.co.za/car/sm_process.asp. (Accessed: 20th June 2020).

Torrione, P.; Collins, L. M. and Morton, K. D. (2014). Multivariate analysis, chemometrics, and machine learning in laser spectroscopy. In: Baudelet, M. Laser Spectroscopy for Sensing. USA: Woodhead Publishing. p125–164.

Van der Merwe, S. (2018). 'Application of Near Infrared Spectroscopy and chemometrics for the analysis of nutraceuticals in South Africa', MSc thesis, Cape Peninsula University of Technology, Cape Town.

Walford, SN. (2013). 'NIR - A technique that has come to stay'. Sugar Milling Research Institute NPC.

Williams, P. (2007). Near Infrared Technology - Getting the best out of light. 5th ed. Canada: PDK Projects. p6-16.

Xin, L.; Chai, X.; Barnes, D.; Chen, C. and Chen, R. (2014). Rapid identification of tissue paper made from blended recycled fibre by Fourier transform near infrared spectroscopy. Journal of Near Infrared Spectroscopy. 22, p347-355.

Chapter Three: Experimental

This thesis discusses the classification of sugar factory stream products using existing SMRI-NIRS prediction equations (Walford 2018). The data obtained from the prediction equations were used to develop the classification model for this thesis. The development of the prediction equations is briefly explained in subdivision 3.1 to provide a background for the classification model.

3.1 Quantitative analysis

SMRI has previously developed NIRS prediction equations to analyse brix, pol, sucrose, glucose, fructose, ash, and dry solids for sugar factory processing streams. NIRS allows for rapid, routine analysis used for factory control. In addition, value-added troubleshooting toolkits have been added to the technology to allow routine factory-scale sucrose inversion studies and measurement of losses across centrifuges (Walford 2019). Table 3.1 gives NIRS and conventional results for a mixed juice sample. These are typical of the values used for the development of the NIRS prediction equations but are proprietary information of the SMRI.

Table 3.1: An example of typical NIRS and conventional results for a mixed juice sample

	NIR	**Conventional**
Ash %	0.61	0.57
Brix °Bx	14.02	13.97
Fructose %	0.26	0.25
Glucose %	0.24	0.25
Pol °Z	11.94	12.04
Sucrose %	12.09	12.13

3.1.1 Instrumentation used

Sample sets of mixed juice, clear juice, syrups, A-, B- and C-massecuites and A-, B- and C-molasses were analysed by conventional standard reference methods for brix, pol, sucrose, glucose, fructose, ash, and dry solids (South African Sugar Technologists Association Laboratory Manual 2009). The results thereof were used to develop NIRS calibrations by Partial least squares regressions (PLSR). Figure 3.1 is a schematic of the sugar processing steps in a typical factory and indicates the process areas where the SMRI-NIRS prediction equations are applicable (Simpson and Oxley 2008; Simpson and Naidoo 2010; Walford and Naidoo 2015).

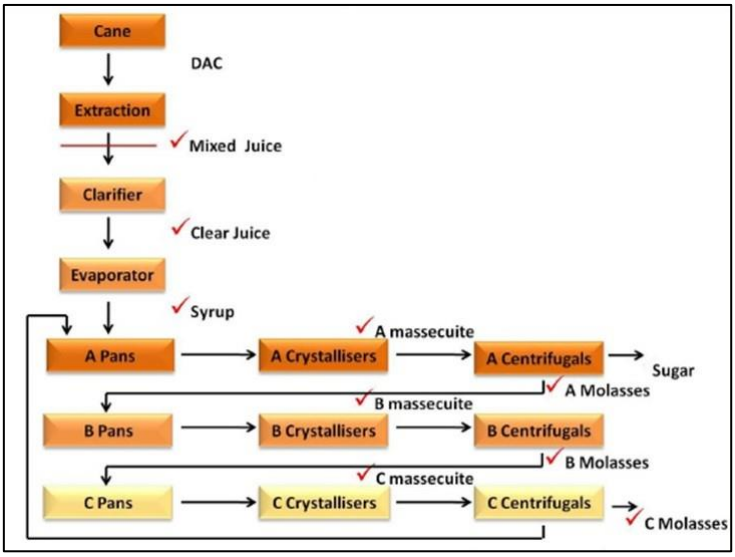

Figure 3.1: Schematic of sugar processing steps in a factory with ticks indicating areas applicable to SMRI-NIRS prediction equations

3.1.1.1 Analytical instruments

Polarimeter: Pol was measured on a Schmidt & Haensch Universal Polartronic at 589 nm.

Refractometer*:* Brix was determined using the Bellingham and Stanley RFM 500 refractometer.

Gas chromatography (GC): Fructose, glucose and sucrose for mixed juice, clear juice, and syrup were quantified on a Varian 3900 GC.

High-Performance Liquid Chromatography (HPLC): A Perkin Elmer 200 series was used to quantify molasses and massecuite fructose, glucose and sucrose, using a PE200 pump, PE 200 autosampler and a Dionex II Pulsed Amperometric Detector connected to Peak, Simple chromatography software for data acquisition.

Conductivity ash*:* A Metrohm 856 conductivity meter was used to determine the conductivity ash.

Dry solids*:* A Karl Fischer titrating apparatus (Metrohm Titrando 890 with a Ti Stand 803), operated by Tiamo software, measured dry solids in molasses samples.

3.1.1.2 NIRS system

A Bruker Multi-purpose Analyser (MPA), in the transmission mode, was used to scan sugar processing samples using a 1 mm QX quartz Hellma flow-through cell (Walford 2019). The absorbance mode was used in the scanning range of 800 to 2500 nm (12500 to 4000 cm^{-1}). OPUS is the Bruker NIRS software used for spectral processing and the development of calibrations, among other chemometric uses (Simpson and Naidoo 2010). Its subset, Opus Lab, provided a simple interface for analysing samples and storing data.

3.1.2 Development of equations

3.1.2.1 Sampling

Weekly composite samples of A, B- molasses, final molasses, juices, massecuites, syrups and other product streams were collected over a season from all 14 mills across South Africa. These samples were chosen to be representative of the various environmental and geographical conditions.

3.1.2.2 NIR calibrations

The samples were subsampled into two portions. The first portion was analysed using conventional methods, and the second was scanned by the NIR method (Walford 2016a, 2016c). For intermediates and final molasses, a Schmidt & Haensch Autodilutor was used to accurately weigh 15 g of sample and add 85 g of distilled water. The solution was then dissolved, injected into the flow-through cell, and scanned using the relevant product to produce a spectrum. For juices, the sample was filtered through postlip medium white filter paper or an equivalent such as Whatman No.6. It was then injected into the flow-through cell and scanned using the relevant product to produce a spectrum (Walford 2016b). Each spectrum was correlated with its respective conventional result, thus producing calibration equations for pol, brix, sucrose, fructose, glucose, ash and dry solids. These calibrations exist for front end (juices), intermediates (Syrup, A-, B-molasses, A-, B- and C-massecuites) and final molasses. The calibration equations are known as prediction equations, as they were then used to predict future samples. The equations need to be robust to accommodate samples of different geographical, seasonal, varietal, and operational effects. To improve the robustness, the equation was updated thrice annually and over many seasons. Figure 3.2 is a schematic depicting the NIR equation development.

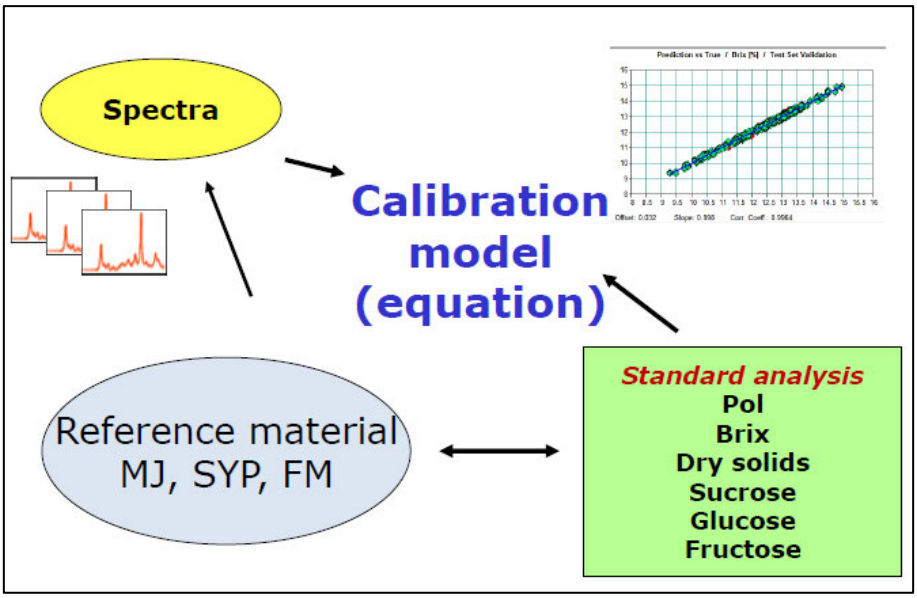

Figure 3.2: NIRS equation development

Once the equations were developed, they were validated to ensure accuracy. Figure 3.3 is a depiction of how an equation is validated. First, a sample of known constituents was prepared using the NIRS method (Walford 2016a, 2016c). Next, the sample was analysed using the SMRI-NIRS prediction equation. The results obtained were then compared to the conventional results. If they were within the measurement uncertainties of the conventional methods, then the prediction equation is deemed successful. If the results were out of uncertainty, the accuracy of the equation was investigated.

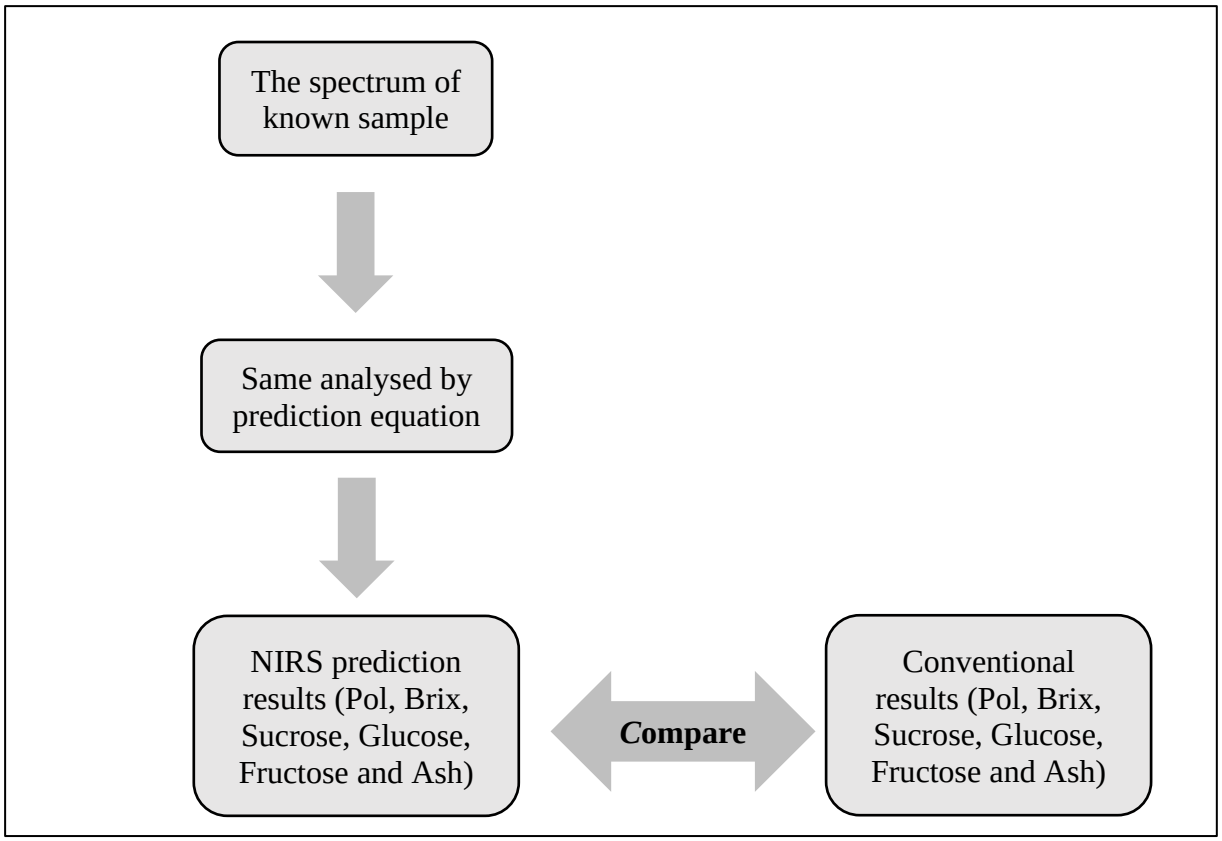

Figure 3.3: Equation validation

The equations were developed using a test set, as the number of samples analysed was large, and validated using an independent set of samples. The robustness of the equations was evaluated using various statistical measures and are tabulated in Tables 3.1, 3.2 and 3.3 (Walford 2021). The statistical measures are outlined below:

- RMSEP: Provided a measure of the average uncertainty that is expected when predicting new samples (Walford 2021),
- Bias: The average difference between the NIRS predicted value and the value obtained from conventional methods. A positive value indicates that the model is overestimating the amount, whilst a negative value indicates an underestimation (NIR Model Development At Celignis 2021),
- SEP: Refers to the standard deviation of differences between the conventional and predicted results (Walford 2021),
- Slope: Rate of change of prediction of the result as a function of the rate of change of conventional result (Walford 2021) and

- Correlation coefficient: A measure of the agreement between the conventional and NIRS predicted values (Walford 2021).

Table 3.2: Statistical measures determining the robustness of juice equations

Analyte	Number of samples	RMSEP	Bias	SEP	Slope	Correlation coefficient
Brix	530	0.084	0.018	0.082	0.998	0.998
Pol	530	0.081	0.031	0.075	1.008	0.998
Sucrose	530	0.097	0.019	0.096	1.006	0.997
Glucose	530	0.040	0.015	0.037	0.927	0.845
Fructose	530	0.028	-0.006	0.027	0.928	0.914
Ash	530	0.063	-0.025	0.058	0.805	0.650

Table 3.3: Statistical measures determining robustness of intermediate equations

Analyte	Number of samples	RMSEP	Bias	SEP	Slope	Correlation coefficient
Brix	811	0.716	0.028	0.715	0.998	0.999
Pol	608	0.703	-0.047	0.702	1.000	0.999
Sucrose	455	1.000	-0.016	1.000	0.994	0.997
Glucose	212	0.341	-0.038	0.339	0.971	0.987
Fructose	542	0.319	0.051	0.315	0.994	0.989

Table 3.4: Statistical measures determining the robustness of final molasses equations

Analyte	Number of samples	RMSEP	Bias	SEP	Slope	Correlation coefficient
Brix	249	0.697	0.013	0.697	0.922	0.947
Pol	248	0.708	-0.163	0.689	0.888	0.944
Sucrose	152	0.767	-0.272	0.717	0.978	0.935
Glucose	152	0.469	0.082	0.462	0.943	0.907
Fructose	152	0.590	-0.113	0.580	0.909	0.778
Ash	181	1.113	-0.252	1.084	0.891	0.896
Dry Solids	229	1.803	0.185	1.794	0.681	0.701

3.2 Qualitative analysis

Samples from various geographical and environmental conditions of the sugar stream products were collected. First, these samples were scanned by NIRS and the correct equation was

selected; this information was used to obtain the correct reference NIRS spectra. After that, an appropriate classification model was developed using discriminant analysis. The model's validity was then determined by comparing the classification predictions to the true identities of the validation samples.

3.2.1 Instrumentation and data processing software used

A Bruker MPA and the Opus Lab software were used to scan sugar stream products as per routine SMRI methods (Walford, SN. 2016a, 2016b, 2016c). The data from these scans were then processed using Orange: Data Mining Toolbox in Python (version 3.27.1) (Demsar *et al.* 2013). This is a data mining software that allows for data visualisation and classification of groups and sub-groups.

3.2.2 Sampling

One hundred and seventy-two samples were chosen across several variables such as geographical, seasonal, varietal and operational. These samples comprised of mixed juice, clear juice, syrup, A-, B- molasses, A-, B-, C-massecuite, and final molasses. Twenty seven mixed juice and clear juice samples were clustered into one group called juices. One hundred and twenty five syrup, A-, B- molasses, A-, B- and C-massecuites all formed intermediates. The remaining twenty samples were final molasses. A successful classification model should therefore be able to define three separate groups.

3.2.3 Development of classification model

Two classification models were explored. One uses the analyte absorbances and wavenumbers of these absorbances obtained from the spectra of all the samples. The other uses the analyte concentrations obtained when the samples were analysed with the SMRI-NIRS prediction equations. A general procedure was followed when developing the models. The first step would be to pre-process the data based on the type of data used, such as NIR spectra. This eliminates any physical phenomena in the spectra (Rinnan, Van den Berg and Engelsen 2009). The next step was to reduce the data using principal component analysis (PCA). The resulting data was then tested on a variety of classifier models. Each model was then put through an evaluation scheme of the prediction performance. Based on these results, the optimum classification model was chosen (Mushtaq and Mellouk 2017). A flow chart of this procedure is shown below in Figure 3.4.

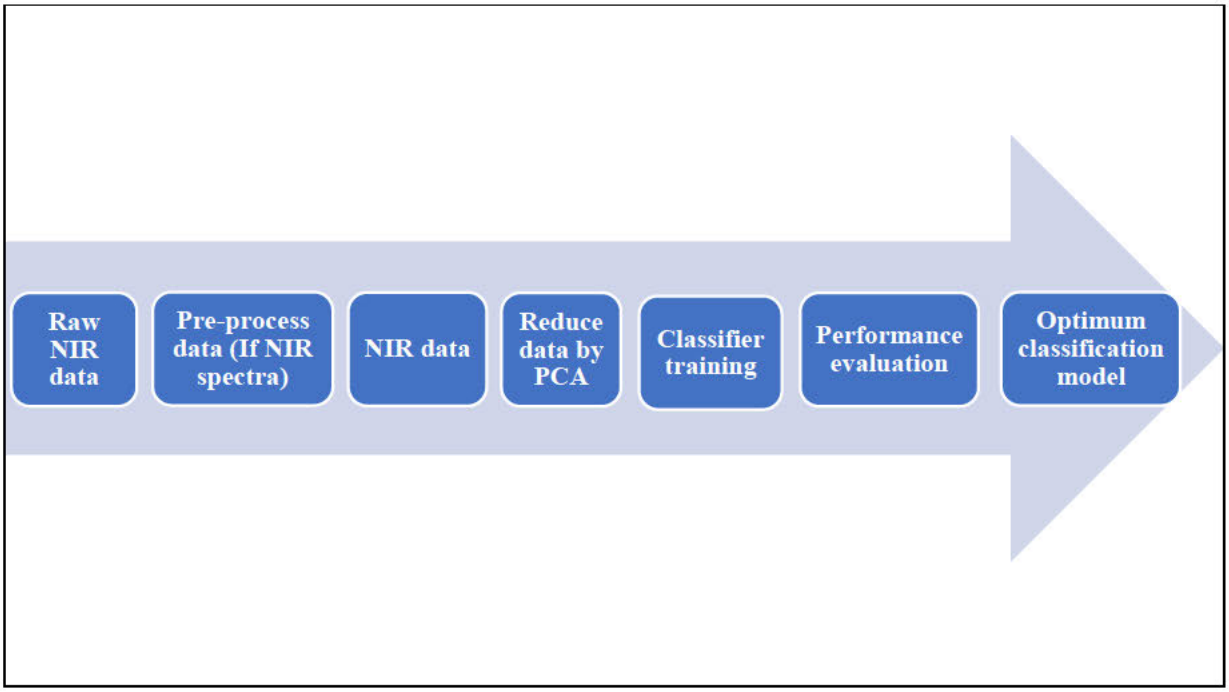

Figure 3.4: General procedure to follow when developing a classification model

3.2.3.1 Classification model using sample absorbances

A model was developed using the analyte absorbances. (Appendix 1)

3.2.3.1.1 Spectra acquisition

Final molasses and intermediate samples such as A-, B-molasses, A-, B-, C-massecuite and syrup were all prepared in the same manner. First, each sample was diluted by mass/mass, by adding 15 g of sample and 85 g of distilled water. Next, the solution was dissolved and then injected into the NIR instrument to be scanned in triplicate. The average of the three scans gave the final NIRS spectra and predicted results using OPUS. The spectra obtained were a plot of absorbance versus wavenumber. These two variables were then used to develop the model.

3.2.3.1.2 Pre-processing

Pre-processing was performed to eliminate any physical phenomena from the spectra which could negatively impact the next step, such as exploratory data or development of the classification model (Rinnan, Van den Berg and Engelsen 2009). Various pre-processing techniques such as Savitzky-Golay (Rinnan, Van den Berg and Engelsen 2009), baseline correction (Rinnan, Van den Berg and Engelsen 2009), Gaussian smoothing (Rinnan, Van den Berg and Engelsen 2009) and normalising spectra (Demsar *et al.* 2013) were evaluated to see which would yield the best principal components (PC) for the next step. Figure 3.5 shows the

PC plots of the various pre-processing techniques. Gaussian smoothing and Savitzky -Golay presented as the best techniques.

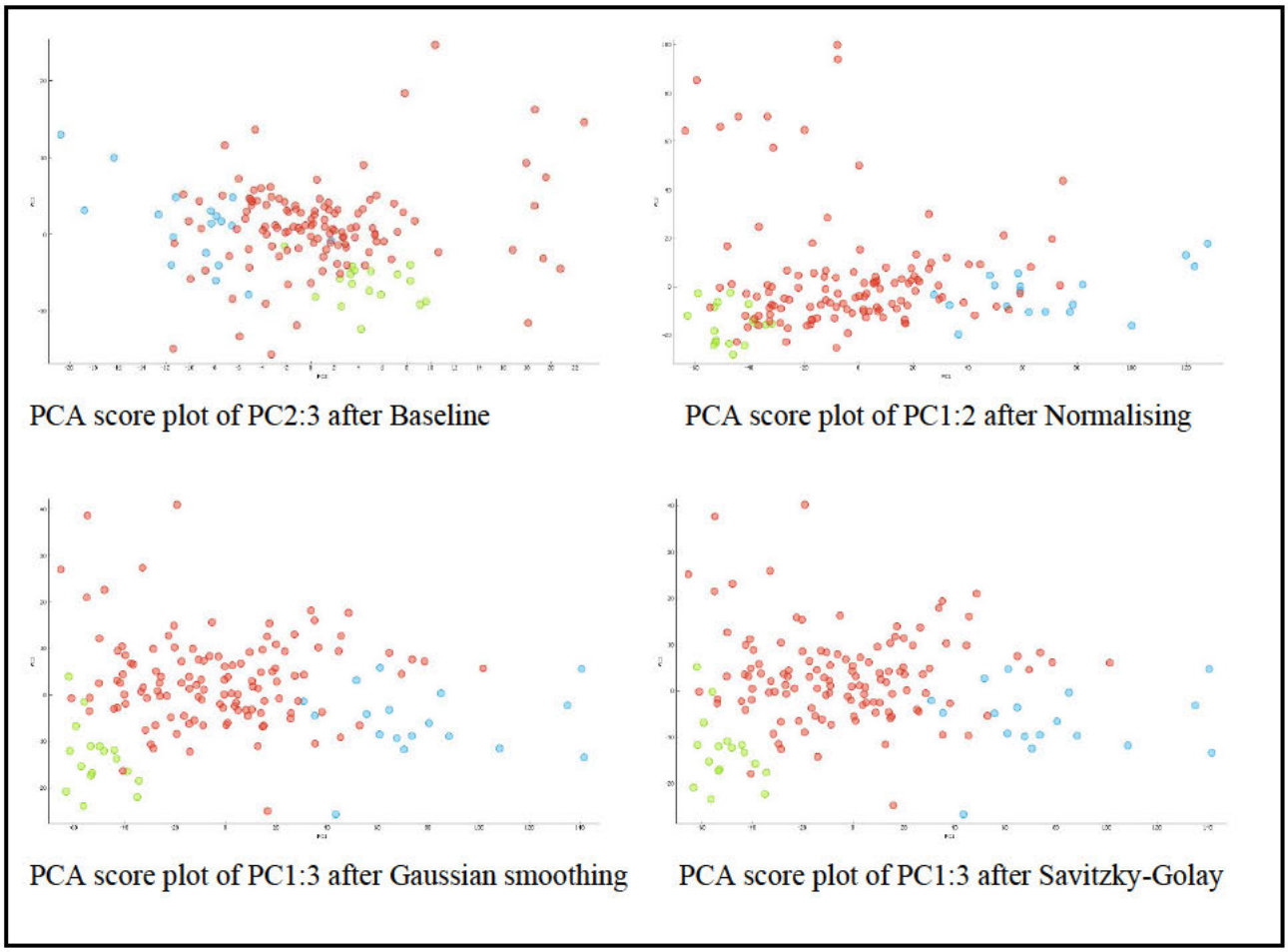

Figure 3.5: PCA score plots after various pre-processing techniques

The spectra after pre-processing with Savitzky-Golay were compared against the spectra after Gaussian smoothing. Gaussian smoothing was then chosen as the preferred pre-processing technique as the spectra showed that more noise was removed. This is seen in Figure 3.6.

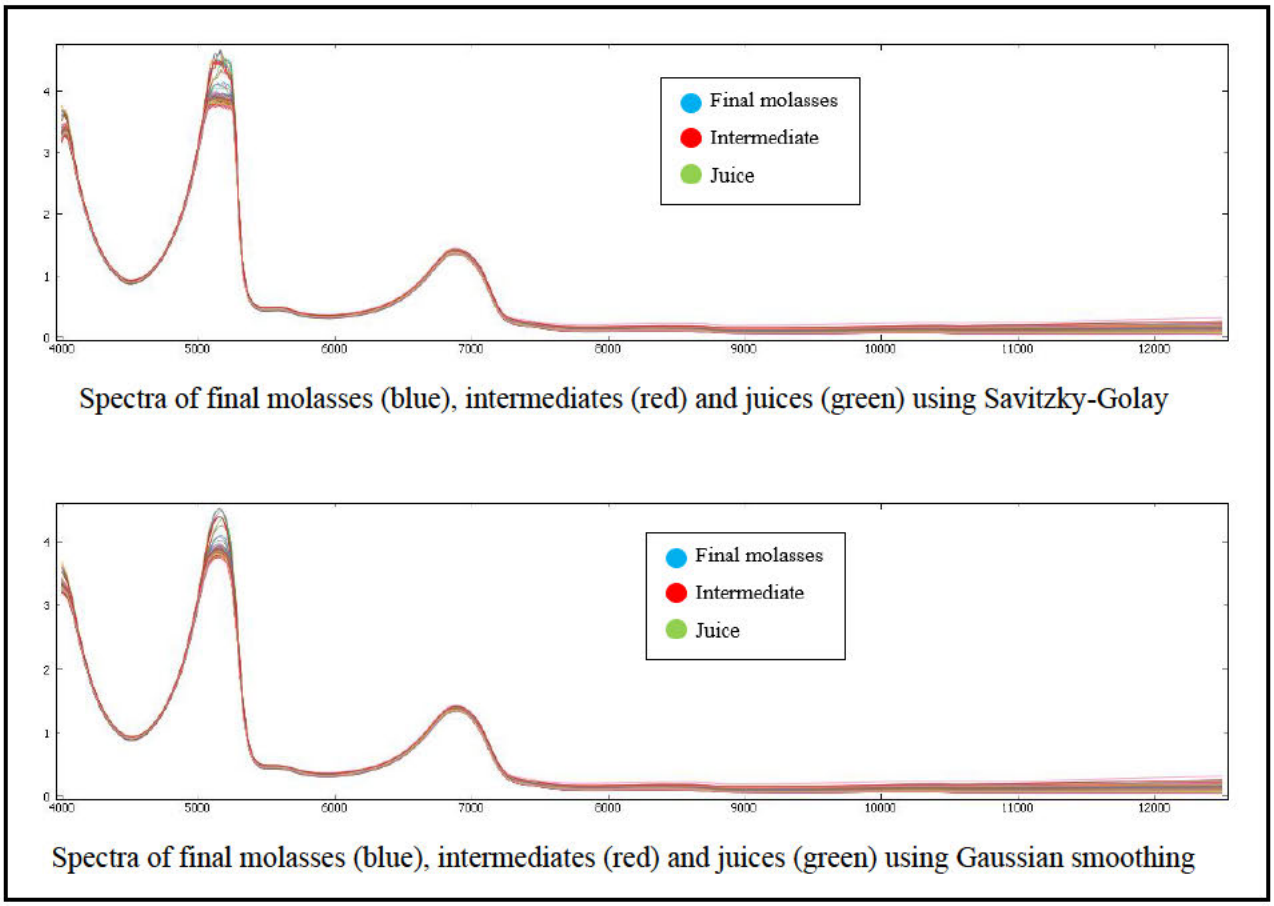

Figure 3.6: Comparison of spectra after Savitzky-Golay and Gaussian smoothing

3.2.3.1.3 Exploratory data analysis - Principal component analysis

PCA was applied to the pre-processed absorbance spectra using three PCs and an explained variance of 95 %. Any potential outliers were identified and investigated to determine the reason for the deviation. True outliers were removed if they resulted from a spectral irregularity due to incorrect dilutions or incorrect sample processing in the factory. (Nadar and Walford 2019). Fifteen outlier samples were removed. The PCA model was recalculated, and the subsequent PCA score plot of PC 1:3 was used to classify the different sugar streams using wavenumbers and absorbances of interest.

3.2.3.1.4 Classifier training

Supervised classification models were developed to differentiate between the various sugar stream products: juices, intermediates, and final molasses. The samples spectra were randomly separated into a calibration set and a test set. The calibration set was used to assess the various classification techniques such as KNN, classification tree, SVM, and Logistic regression (Mushtaq and Mellouk 2017).

3.2.3.1.4.1 K-nearest neighbour (KNN)

The KNN algorithm works on the assumption that similar objects will exist close to each other (Zhang 2016). First, the distance between the unknown sample point and sets of samples with known classes was calculated. Then, KNN was used to determine the similarity based on the Euclidean distance between the points on a graph. Next, classification was determined based on the closest K samples. Next, cross-validation was performed on a calibration data set to determine the optimum K value. K values of 1, 3 and 4 were evaluated based on classification accuracy, precision, recall, F1 score and specificity. A confusion matrix was also used to show the performance of the calibration model. Finally, the model was applied to an independent validation set, and the results were investigated based on the performance evaluation measures such as classification accuracy, misclassification rate and the F1 score.

3.2.3.1.4.2 Classification tree

A classification tree is a method used to split the data into nodes depending on class type. The structure of the tree typically consists of the following components:

a) Internal nodes, which test a feature
b) branches, which correspond to feature values
c) leaf nodes, which assign a classification

Instances start at the root node and are based on the feature values. The tree was sorted down to a leaf node (Mushtaq and Mellouk 2017). This is depicted in Figure 3.7 (Sá et al. 2011).

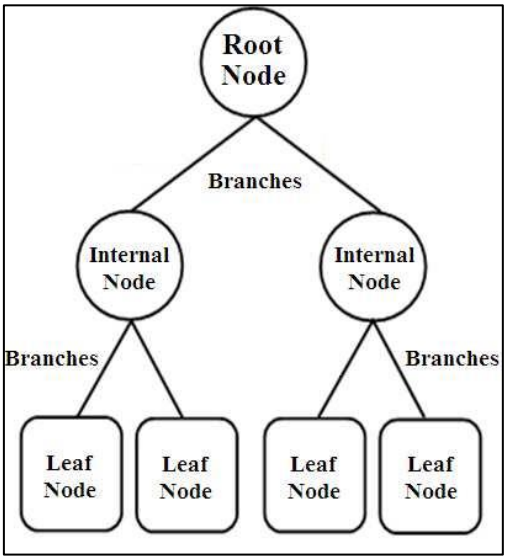

Figure 3.7: The structure of a classification tree (Mushtaq and Mellouk 2017)

Cross-validation was performed on a calibration data set. The results were then evaluated based on classification accuracy, precision, recall, F1 score and specificity. A confusion matrix was also used to show the performance of the calibration model. Finally, the model was applied to an independent validation set, and the results were investigated based on the performance evaluation measures such as classification accuracy, misclassification rate and the F1 score.

3.2.3.1.4.3 Support vector machine (SVM)

SVM operates by analysing the available data and deciphering the patterns in the data to make a classification model (Mushtaq and Mellouk 2017). SVM is best used for binary classification but can be used for multiclass classification. First, data points were separated into two classes. In this case, multiclass classification was performed by creating multiple binary classifications and then separating each one further into the two classes (Baeldung 2020). Next, cross-validation was performed on a calibration data set. The results were then evaluated based on classification accuracy, precision, recall, F1 score and specificity. A confusion matrix was also used to show the performance of the calibration model. Finally, the model was applied to an independent validation set, and the results were investigated based on the performance evaluation measures such as classification accuracy, misclassification rate and the F1 score.

3.2.3.1.4.4 Logistic regression

Logistic regression is a predictive type of analysis that analysed the data and then explained the relationship between one dependent variable and two independent variables (Statistics Solutions 2020). Next, cross-validation was performed on a calibration data set. The results were then evaluated based on classification accuracy, precision, recall, F1 score and specificity. A confusion matrix was also used to show the performance of the calibration model. Finally, the model was applied to an independent validation set, and the results were investigated based on the performance evaluation measures such as classification accuracy, misclassification rate and the F1 score.

3.2.3.1.5 Performance evaluation

Classification accuracy (CA), which indicates how many samples were correctly classified, was calculated using Equation 5. CA is the most important measurement of performance as it looks at the efficacy of the model.

$$\boldsymbol{Classification\ Accuracy = \frac{Number\ of\ correct\ predictions}{number\ of\ all\ predictions}} \qquad \textbf{...Eqn 5}$$

Other performance measures were also taken into consideration when choosing the best model. These included the confusion matrix, precision, recall, F1 score, and specificity (Hossin 2015). The confusion matrix further expounds on the predictions given. It shows which predictions were true or false. Precision measures how good the model is at positive predictions and was calculated as per Equation 6.

$$\boldsymbol{Precision = \frac{TP}{TP+FP}} \qquad \textbf{... Eqn 6}$$

Where: TP = True positive and FP = False positive

Recall describes the efficacy of the model with regards to the probability that positive classes will be correctly predicted (Payne 2019). This was calculated using Equation 7.

$$\boldsymbol{Recall = \frac{TP}{TP+FN}} \qquad \textbf{...Eqn 7}$$

Where: TP = True positive and FN = False negative

The F1 score was calculated as the weighted average of precision and recall as shown in Equation 8.

$$\boldsymbol{F1\ score\ = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall}} \qquad \textbf{...Eqn 8}$$

Specificity was measured as a proportion of the negative class that was correctly classified as negative (Hossin 2015). This was accounted for in Equation 9.

$$\boldsymbol{Specificity = \frac{TN}{TN+FP}} \qquad \textbf{...Eqn 9}$$

Based on these performance measures, the best model was optimised by selecting the optimum number of PCs and pre-processing technique. The resulting optimum model was then evaluated using the same performance measures.

3.2.3.2 Development of classification model using sample analyte concentrations

A second model was developed using the analyte concentrations (Appendix 2). The Orange software automatically normalises the values to ensure all variables are on a standard scale.

3.2.3.2.1 Ranking system

Using the rank scoring system, which scores variables according to their correlation with discrete or numeric target variables (Demsar *et al.* 2013), glucose was a negligible variable and was thus excluded from the development of the model. Several scoring methods were evaluated. These included information gain, information gain ratio and relief. Information gain describes the expected amount of information (Demsar *et al.* 2013). Information gain ratio was determined as the ratio of information gain to the intrinsic information (Demsar *et al*. 2013). Relief refers to an attribute's ability to distinguish between classes on similar data instances (Demsar *et al*. 2013).

3.2.3.2.2 Exploratory data analysis - Principal component analysis

PCA was applied to the reduced data using only 2 PCs and a 98% explained variance. Outliers were identified, investigated and removed as per **3.2.3.1.3**. Twelve outliers were removed. The only resulting score plot of PC 1:2 was used to classify the sugar stream products based on their analyte concentrations.

3.2.3.2.3 Classifier training

The same models used in **3.2.3.1.4** were used to classify the sugar stream products. This was done to ensure consistency between the two models.

3.2.3.2.4 Performance evaluation

This was performed as per **3.2.3.1.5** to ensure consistency between the two models.

3.3 References

Baeldung. (2020). Multiclass Classification Using Support Vector Machines, Available at: https://www.baeldung.com/cs/svm-multiclass-classification (Accessed: 2nd September 2020).

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

Hossin, M and Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5 (2), p.1-11

Mushtaq, M and Mellouk, A. (2017). Quality of Experience Paradigm in Multimedia Services, in Mushtaq, M and Mellouk, A (ed.) Methodologies for Subjective Video Streaming QoE Assessment. Great Britain: Elsevier, p. 27-57

Nadar, R. and Walford, S. (2019). The SMRI-NIRS Technology: Proficiency and Quality Assurance Scheme. [Poster]. South African Sugar Technologists' Association, 20 – 22 August, International Convention Centre Durban.

NIR Model Development At Celignis. (2021). Available at: https://www.celignis.com/NIR.php (Accessed: 25th April 2021).

Payne, K. (2019). Rapid differentiation of South African game meat using portable near-infrared (NIR) spectroscopy. MSc Thesis. Stellenbosch University, Stellenbosch.

Rinnan, A.; Van den Berg, F. and Engelsen, SB. (2009). Review of the most common pre-processing techniques for near-infrared spectra. Trends in Analytical Chemistry. 28 (10). p.1201-1222.

Sá, J. A. S.; Almeida, A. C.; Rocha, B. R. P.; Mota, M. A. S.; Souza, J. R. S. and Dentel, L.M. (2011). 'Lightning forecast using data mining techniques on hourly evolution of the convective available potential energy', 10th Brazilian Congress on Computational Intelligence, p. 1-5.

Simpson, R. and Naidoo, Y. (2010). Using Near Infra-Red Spectroscopy for Rapid Quantification of Intermediate Sugar Factory Products. Proceedings of the South African Sugar Technologists Association. 83 (1), p382-391.

Simpson, R. and Oxley, J. (2008). Routine analysis of Molasses and Mixed Juice by NIR Spectroscopy. Proceedings of the South African Sugar Technologists Association. 81 (1), p245 - 265.

South African Sugar Technologists Association Laboratory Manual. (2009). 5th Edition, South African Sugar Technologist Association. South Africa. Methods 6.1.7, 6.1.8, 6.1.9, 6.3.6, 6.5.1, 6.5.8, 6.6.1, 6.6.2, 6.6.3, 6.6.4, 6.6.6.

Statistics Solutions. (2020). What is Logistic Regression?, Available at: https://www.statisticssolutions.com/what-is-logistic-regression/
(Accessed: 3rd September 2020).

Walford, S.N. and Naidoo, S. (2015). Light at the end: A season of composite MJ NIRS Analysis. Proceedings of South African Sugar Technology Association. 88, p90-101.

Walford, S.N. (2016a). Standard operating procedure - The analysis of Factory Streams by NIRS, 3rd edition, South Africa, Method NIR – FS1.

Walford, S.N. (2016b). Standard operating procedure - The analysis of filtered Mixed Juice by NIR, 2nd edition, South Africa, Method NIR – MJ1.

Walford, S.N. (2016c). Standard operating procedure - The dilution of Factory Stream samples for analysis by NIRS, 4th edition, South Africa, Method NIR – FS2.

Walford, S. (2018). The SMRI-NIRS Technology: Development, validation and application. [Poster]. South African Sugar Technologists' Association, 14 – 16 August, International Convention Centre Durban.

Walford, S. (2019) Near infrared spectroscopy: rethinking the analysis of sugarcane factory streams. Proceedings of the 18th International Conference on Near Infrared Spectroscopy. p. 129-133

Walford, S. (2021). 'NIRS Steering Committee 21-02-25.' [PowerPoint presentation]25 February. Sugar Milling Research Institute NPC, Durban.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine. 4 (11), 218

Chapter Four: Results, Discussion and Conclusion

4.1 Sugar stream product classification based on absorbances at respective wavenumber

4.1.1 Spectral analysis

The samples analysed comprised juices (mixed juice and clear juice), intermediates (Syrup, A-, B- molasses, A-, B- and C-massecuites) and final molasses. Each sample was scanned between a wavenumber of 12500 cm^{-1} and 4000 cm^{-1}. The spectrum of each sample was plotted with absorbance as a function of wavenumber shown in Figure 4.1. The wavenumber axis has been automatically transposed by the data processing software Orange 3.27.1 so that the wavenumbers increase left to right.

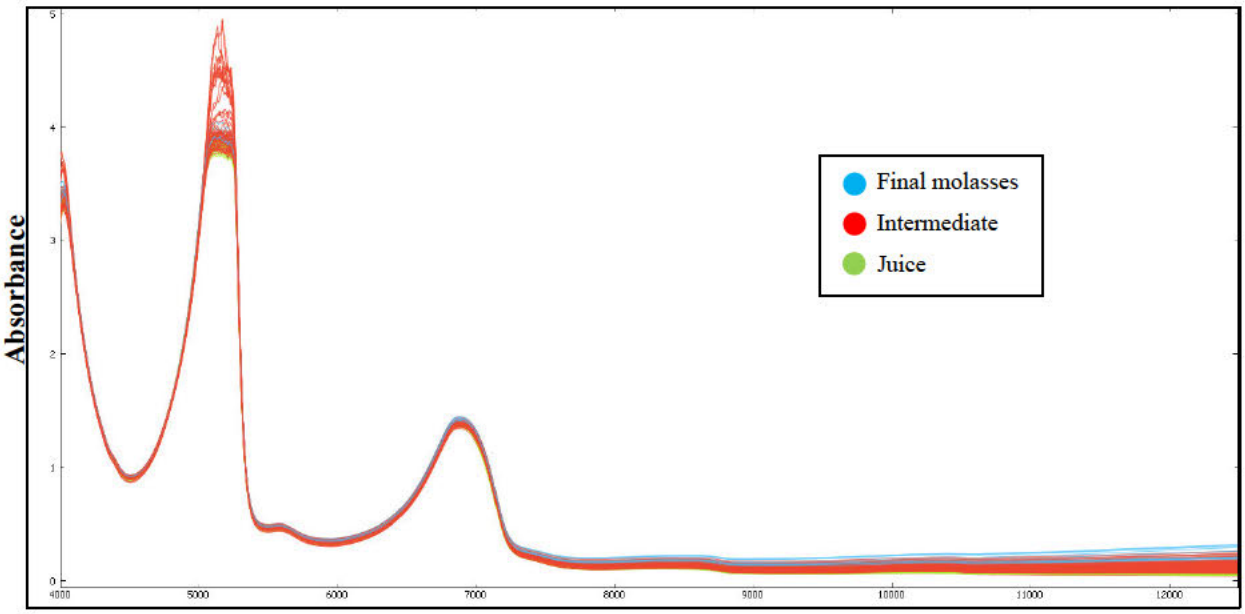

Figure 4.1: Unprocessed spectra of final molasses (blue), intermediates (red) and juices (green)

The unprocessed spectra show a similar trend for all products. All samples except juices were diluted in the same manner, that is, 15 g of the sample plus 85 g of distilled water (100 g of solution). This created a solution that was similar in concentration to juices. For this reason, the spectra were all found to be similar; however, a difference in intensity bands between 5000 and 5268.70 cm^{-1} was observed. As the spectra are unprocessed, the difference can be due to chemical and physical phenomena such as light scattering. There are three prominent peaks in the NIR region, cresting at 4000, 5200 and 6890 cm^{-1}. This coincides with the peaks observed

in the spectra. All samples, except the juices, are dissolved in distilled water before being scanned. The juice products have a substantial amount of water which occurs during the production of sugar. The distilled water thus exhibits three typical peaks in all spectra. It is, therefore, necessary to use chemometrics to differentiate between spectra of sugar stream products.

4.1.2 Principal component analysis (PCA)

The raw spectra were pre-processed using Gaussian smoothing, shown in Figure 4.2. PCA was then calculated on the resultant data using the first five principal components (PCs) to explain 99 % of the variability. PC1, PC2, PC3, PC4 and PC5 account for 84.3 %, 6.7 %, 4.5 %, 2.9 % and 1.1 % of the variation, respectively. The best separation was explained using PC1:3 score plots, as shown in Figure 4.3. PC1:2 (Figure 4.4), PC1:4 (Figure 4.5), PC1:5 (Figure 4.6), PC2:3 (Figure 4.7), PC2:4 (Figure 4.8), PC2:5 (Figure 4.9), PC3:4 (Figure 4.10), PC3:5 (Figure 4.11) and PC4:5 (Figure 4.12) all show poor separation between sugar stream products.

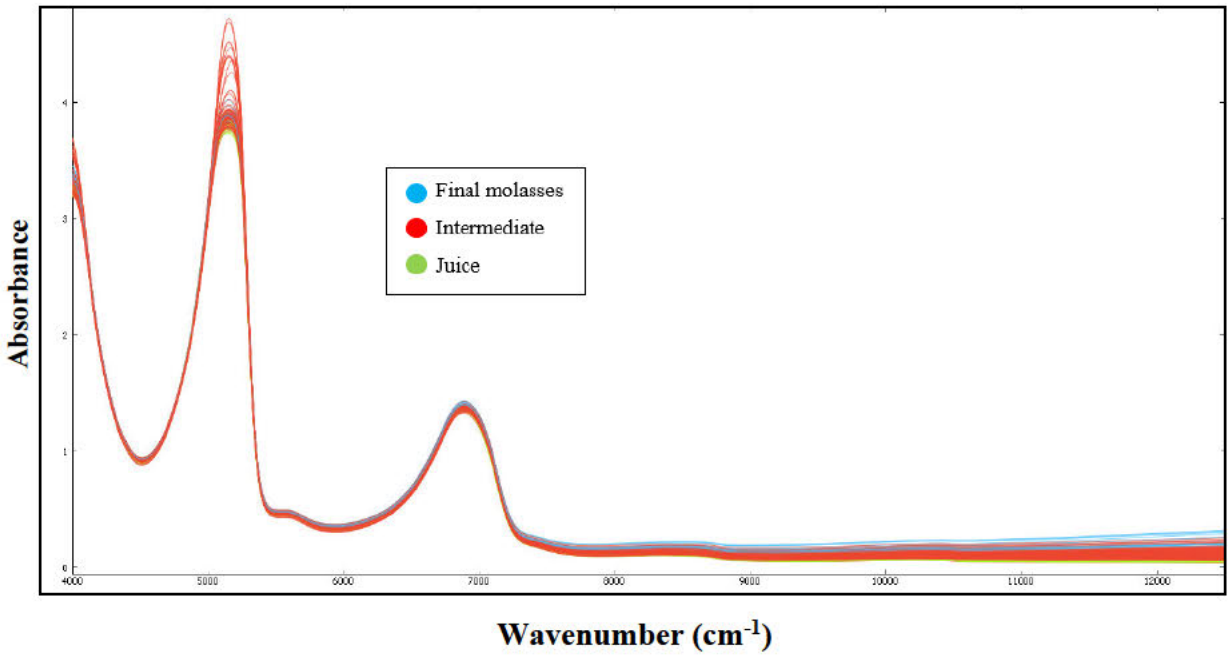

Figure 4.2: Pre-processed spectra of final molasses (blue), intermediates (red) and juices (green) using Gaussian smoothing

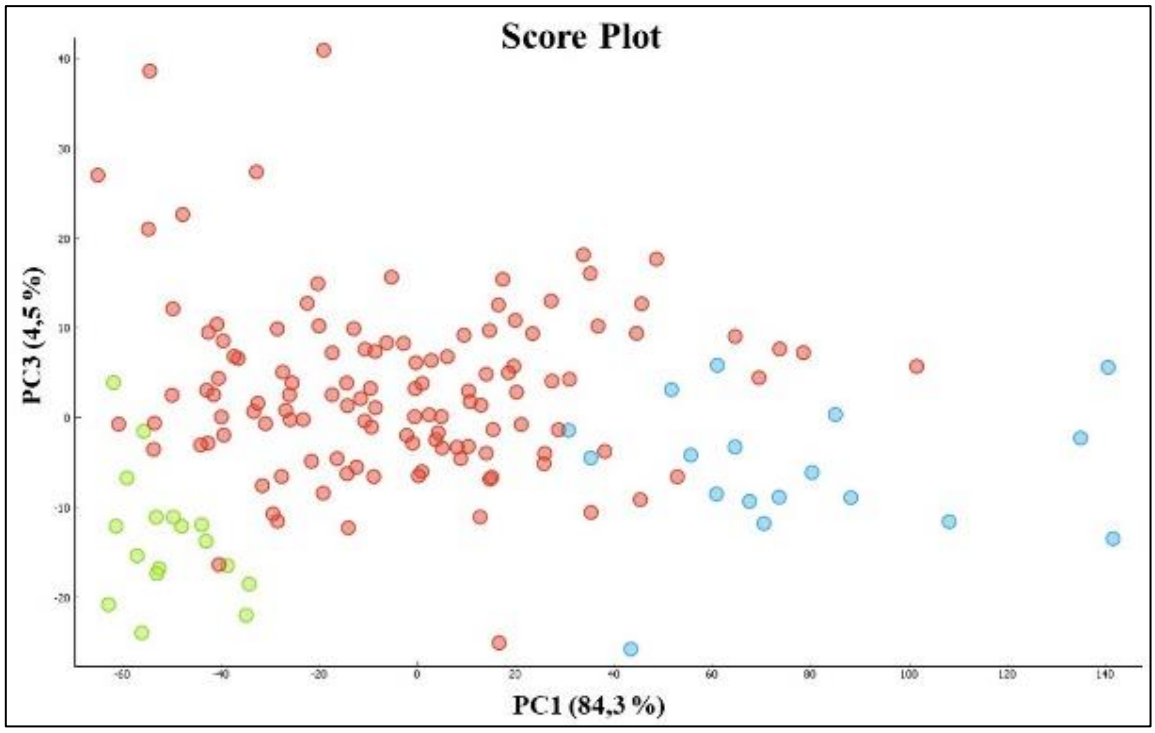

Figure 4.3: PCA score plot of PC1:3

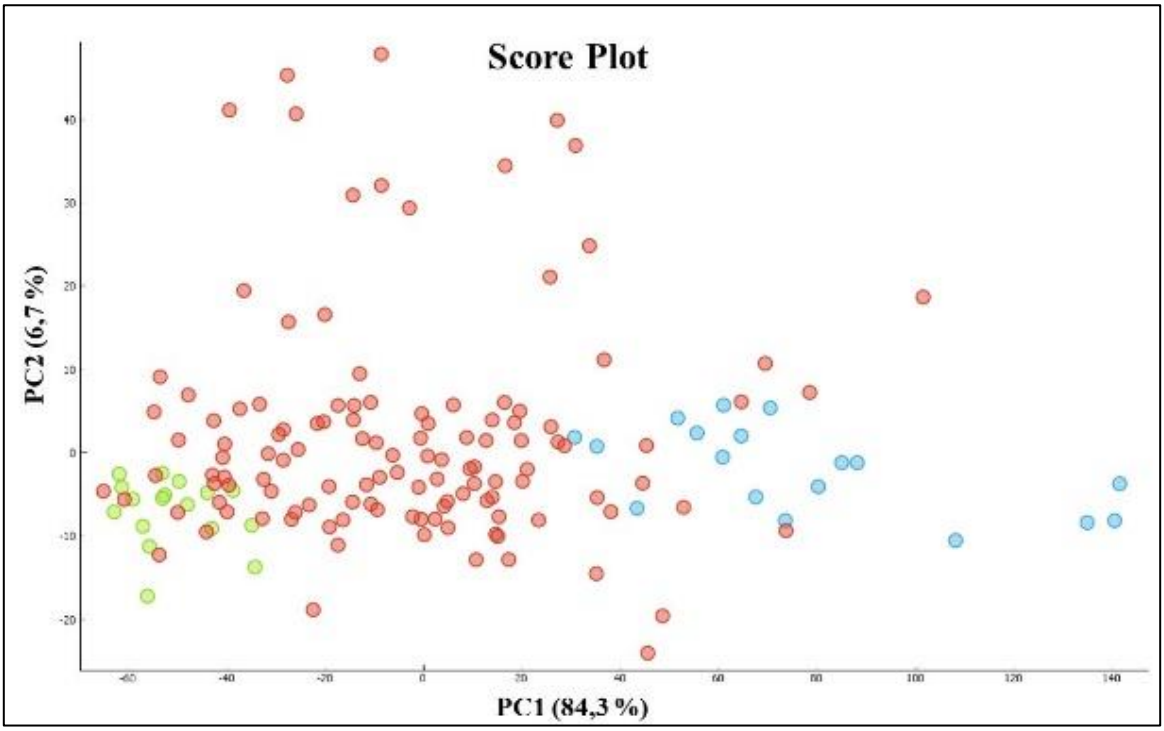

Figure 4.4: PCA score plot of PC1:2

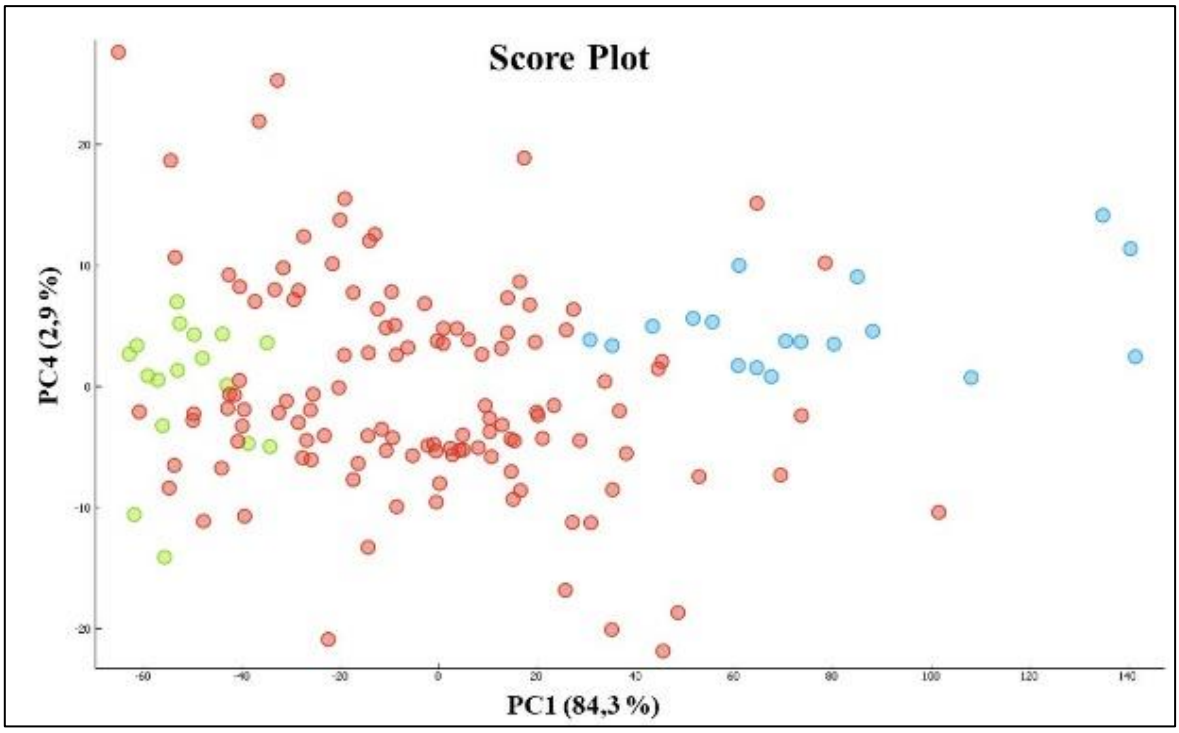

Figure 4.5: PCA score plot of PC1:4

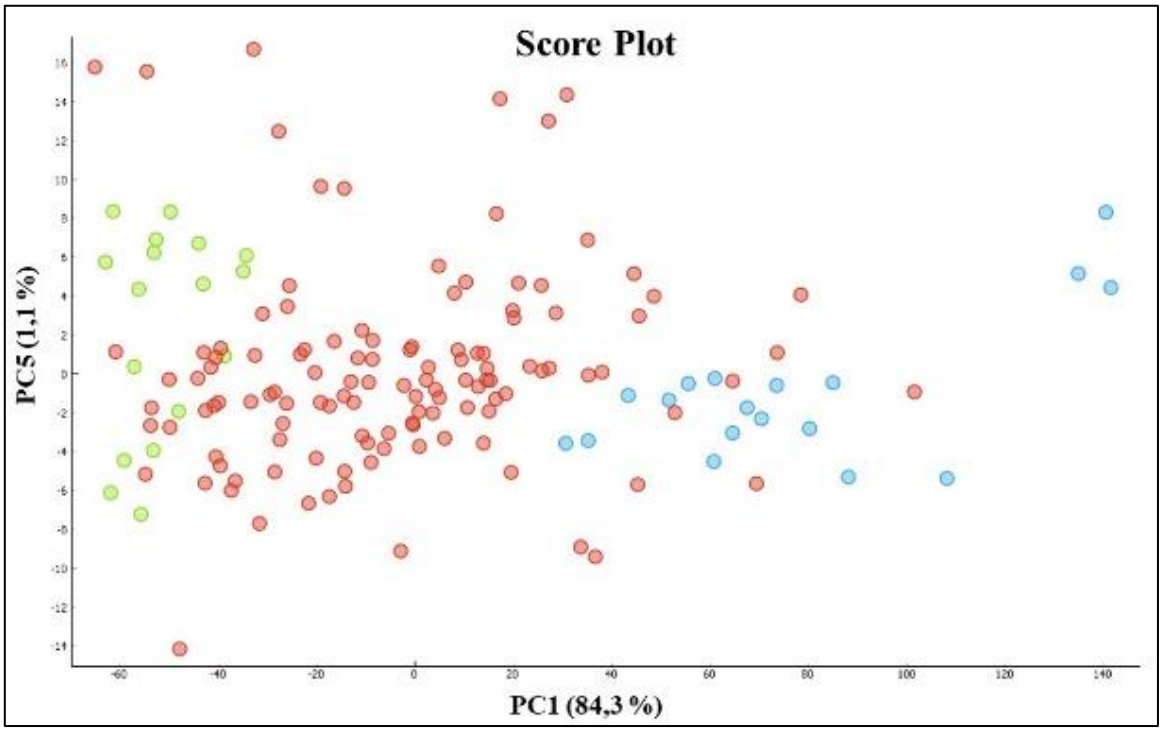

Figure 4.6: PCA score plot of PC1:5

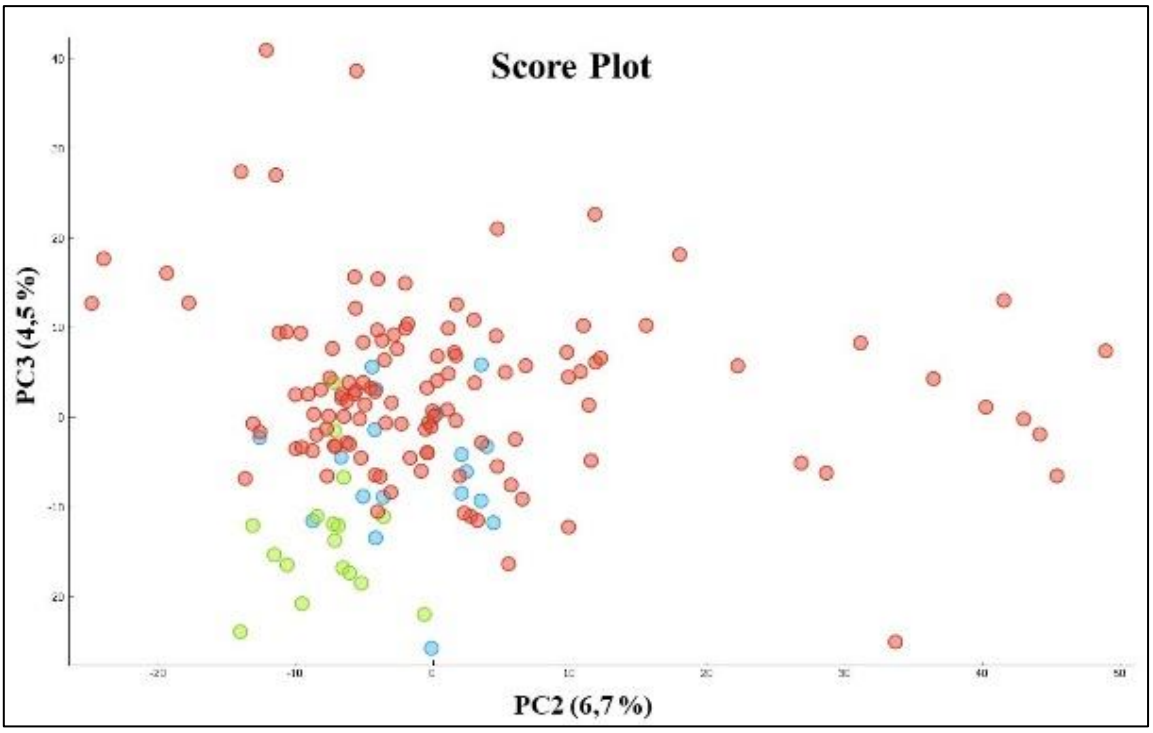

Figure 4.7: PCA score plot of PC2:3

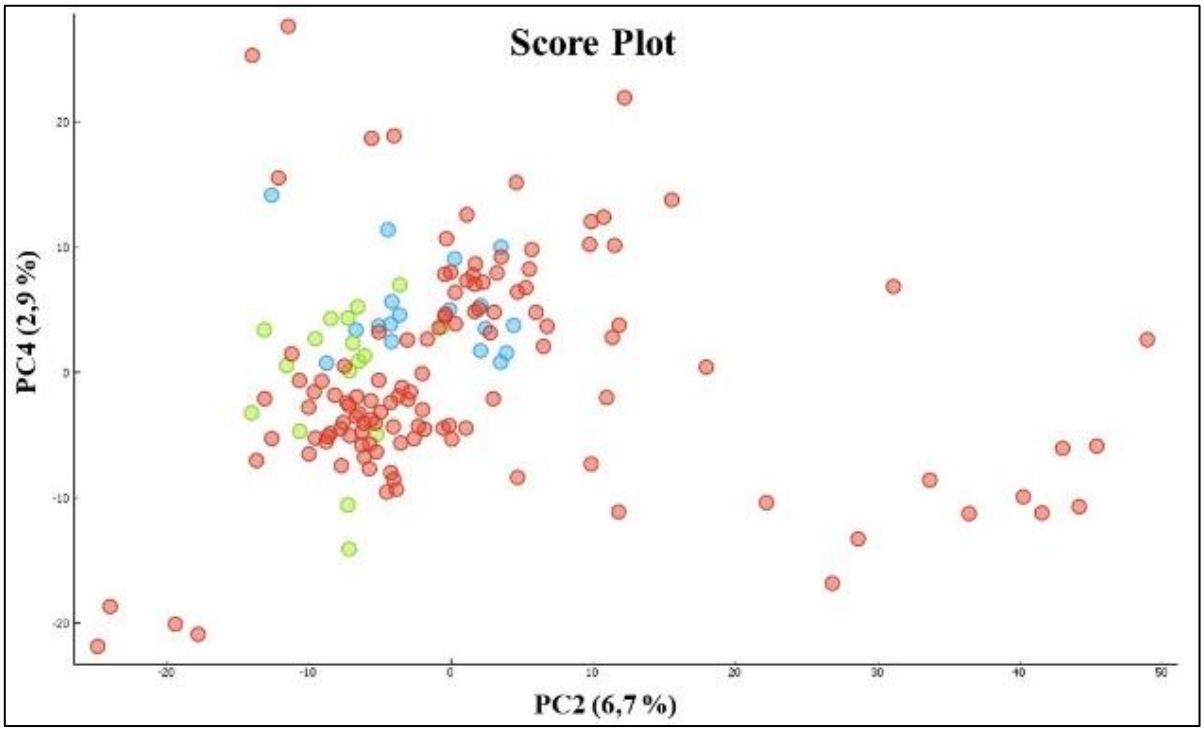

Figure 4.8: PCA score plot of PC2:4

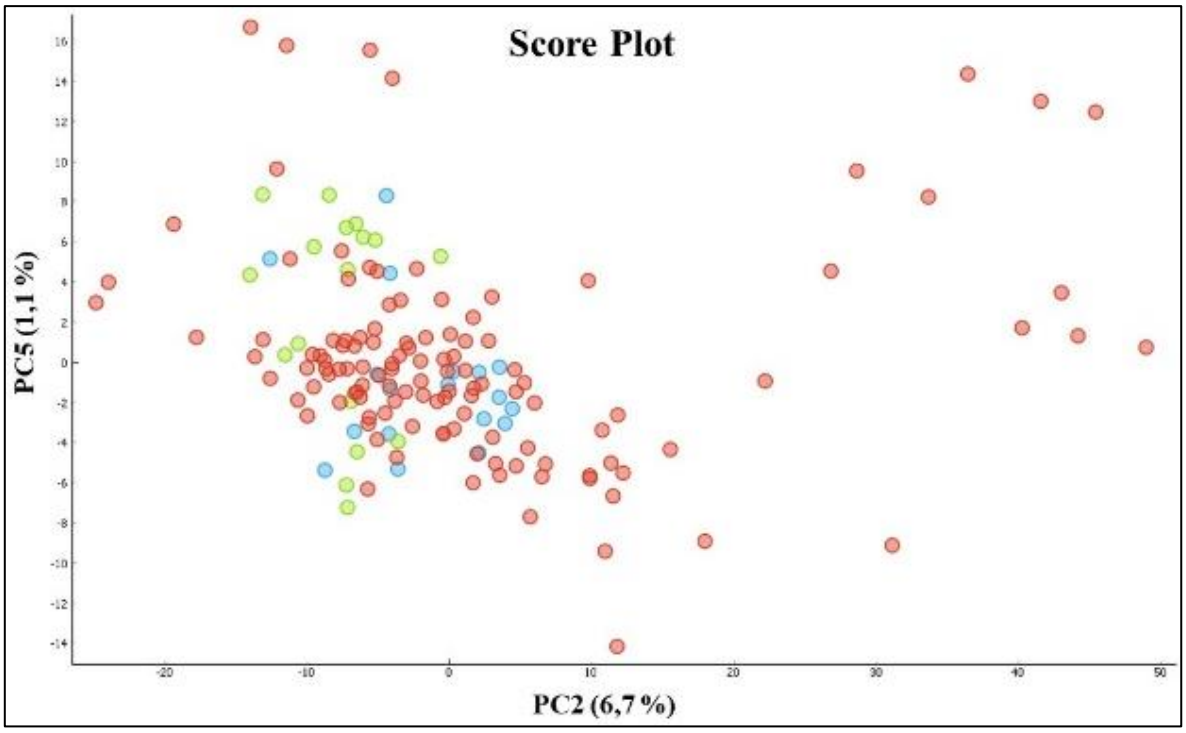

Figure 4.9: PCA score plot of PC2:5

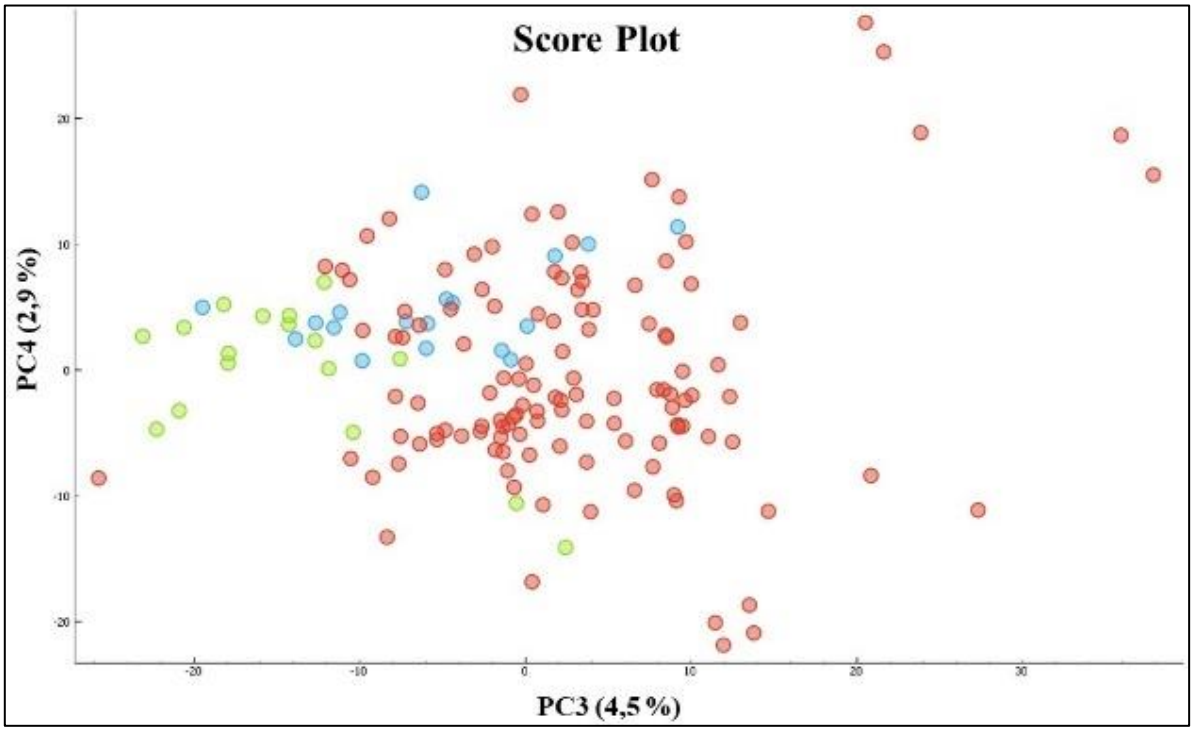

Figure 4.10: PCA score plot of PC 3:4

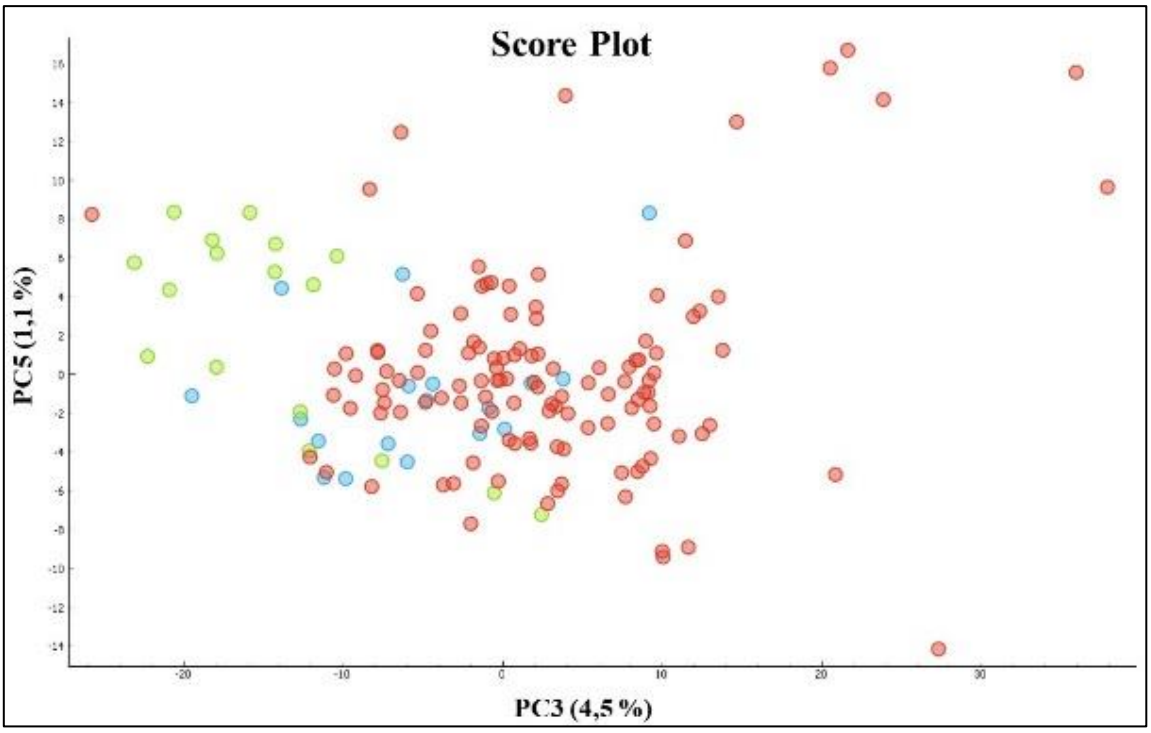

Figure 4.11: PCA score plot of PC3:5

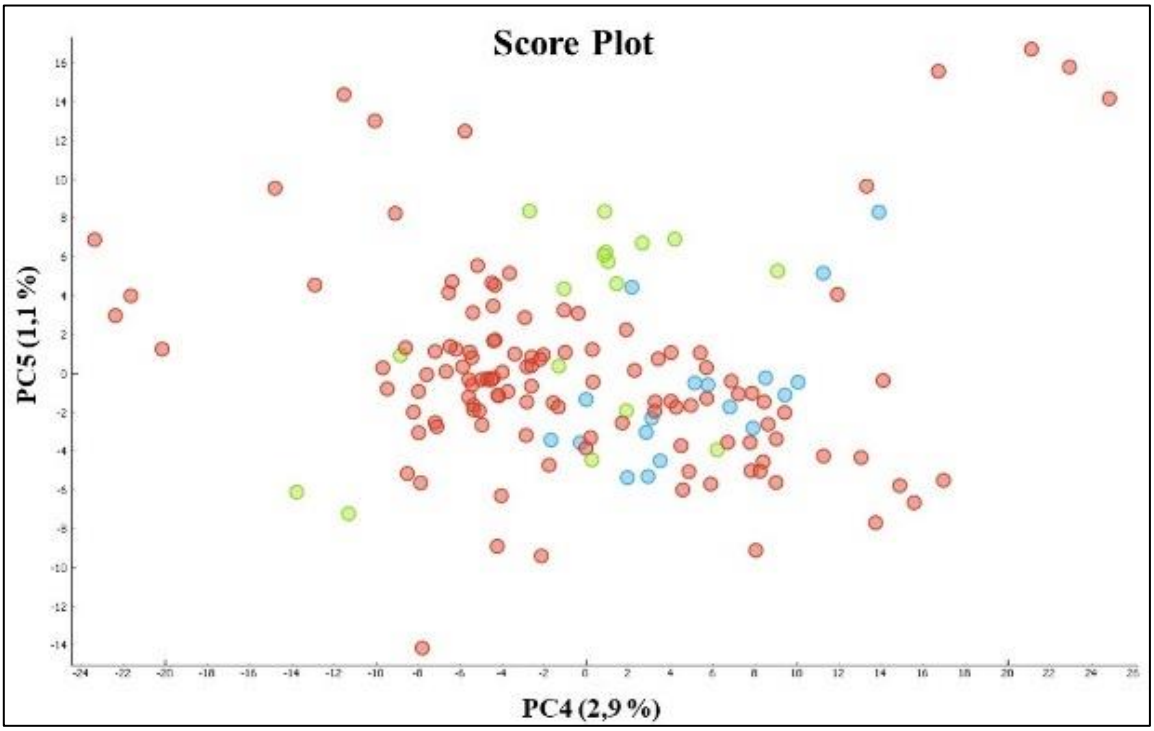

Figure 4.12: PCA score plot of PC4:5

4.1.3 Development of a classification model

All models were pre-processed with Gaussian smoothing and 5 PCs. A score plot of PC1:3 was used to develop subsequent models. A cross-validation sampling method was also applied to all models for consistency. F1 score is used as a measure of accuracy across all models as imbalanced data was used to develop the model.

4.1.3.1 K-nearest neighbour (KNN)

KNN was applied to the PC scores, the optimum K value was determined, and the evaluation results of the calibration and independent validation sets are tabulated below in Table 4.1. Based on the results of Table 4.1, the classification accuracy of the three nearest neighbours (93.0 %) proved to give the best classification model. This value also gave a 92.5 % accuracy for the independent validation set. An overall calibration misclassification rate of 7.0 % was determined. This was observed in a scatter plot of PC1:3 shown in Figure 4.13. Figure 4.13 shows the majority of the misclassification occurred between the intermediates and final molasses region, where there is a clear overlap between the two types of sugar stream samples. Although the KNN model with one nearest neighbour gave a higher classification accuracy of 96.2 %, the lower validation result of 85.0 % indicates that the model was over-fitted. This may have occurred because the calibration training set did not have a good representation of sample data. The independent validation set consisted of samples that the model did not recognise well. The F1 scores are used as a measure of performance for imbalanced data. These scores coincide with the calibration and validation classification accuracy scores, confirming the models' efficiency. The results of the KNN (3) model suggest the model, although over-fitted, is more robust with better predictive powers than the others. Literature also recommends using an odd-numbered nearest neighbour as it provides an unbiased evaluation (Shahraki 2017).

Table 4.1: Evaluation of KNN model with various KNN values

K value	KNN Model	Classification accuracy (%)	Misclassification rate (%)	F1 Score (%)
1	Calibration	96.2	3.8	96.2
	Independent validation	85.0	15.0	85.1
3	Calibration	93.0	7.0	93.1
	Independent validation	92.5	7.5	92.7
4	Calibration	92.4	7.6	92.5
	Independent validation	90.0	10.0	89.9

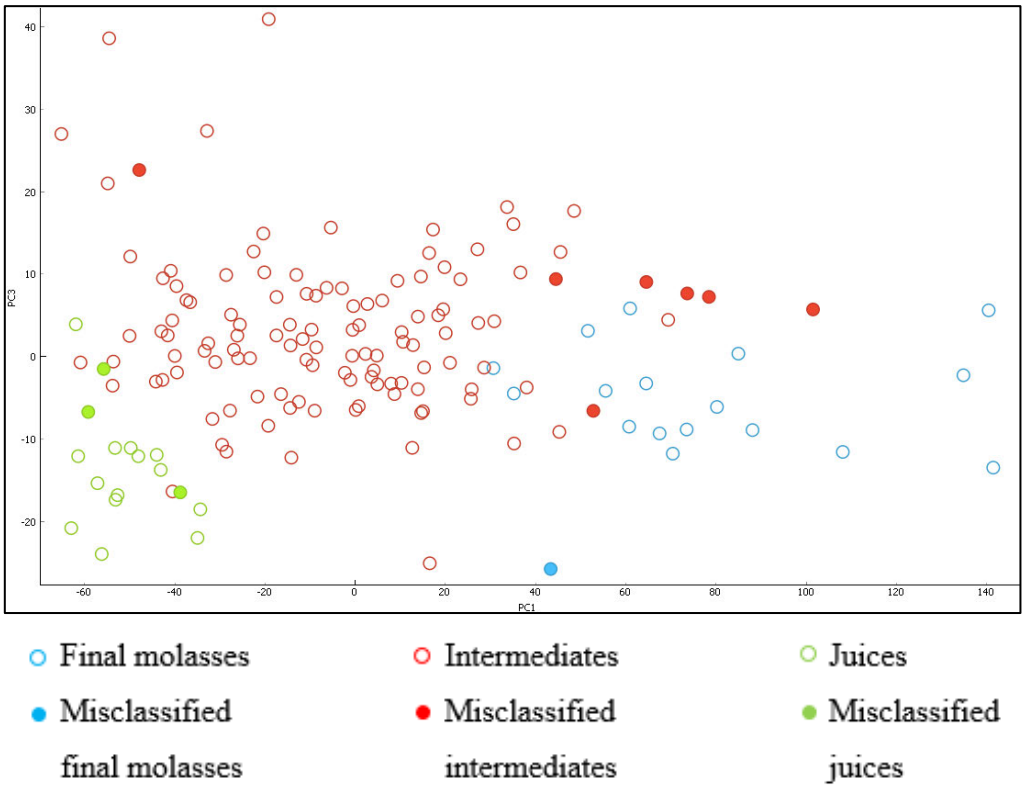

Figure 4.13: A scatterplot of misclassified samples accounting for 7.00 % of the total amount of samples analysed using the KNN (3) model

An evaluation of the performance measures of the model is tabulated in Table 4.2. These results imply that the model had a higher accuracy for juices, at 97.5 %; however, an F1 score of 87.5 % indicates that the model was more accurate for intermediates with a F1 score of 95.4 %. The higher precision and recall values for intermediates indicate the model's ability to predict positive cases for intermediates correctly. The model incorrectly predicted six intermediate samples as final molasses, which is reflected in the low precision score of 73.9 % for final molasses. This correlates to the lowest specificity with a score of 88.6 % for intermediates. Two of these intermediates were found to be C-massecuites which are similar to final molasses.

Although the model showed satisfactory performance regarding each product stream, the decreased validation accuracy indicates an over-fitted model. The model will thus have a low probability of correctly differentiating between products.

Table 4.2: An evaluation of performance measures for the KNN (3) model

K value	3		
Sugar stream product	**Final molasses**	**Intermediates**	**Juices**
Classification Accuracy (%)	95.5	93.0	97.5
Misclassification rate (%)	4.5	7.0	2.5
F1 score (%)	82.9	95.4	87.5
Precision (%)	73.9	96.6	93.3
Recall (%)	94.4	94.3	82.4
Specificity (%)	95.7	88.6	99.3

4.1.3.2 Classification tree

A classification tree model was applied to calibration and an independent validation set. The results were evaluated and tabulated in Table 4.3. An F1 score of 92.8 % was achieved for the classification model and a 90.2 % for the validation model. These results concurred with the classification accuracy of 93.0 % and a validation accuracy of 90.0 %. A decreased validation accuracy indicates a model that is overfitted. A scatterplot, shown in Figure 4.14, represents the misclassified samples of the calibration set, which accounts for 7.0 %. Five out of seventeen juice samples were misclassified as intermediates. These misclassified samples can be seen in the boundary between juices and intermediates of the PC1:3 plot. This indicates that the samples were too similar, making it difficult to differentiate.

Table 4.3: An evaluation of the calibration and independent validation sets using classification tree

Model type	Classification Accuracy (%)	Misclassification rate (%)	F1 score (%)
Calibration	93.0	7.0	92.8
Independent validation	90.0	10.0	90.2

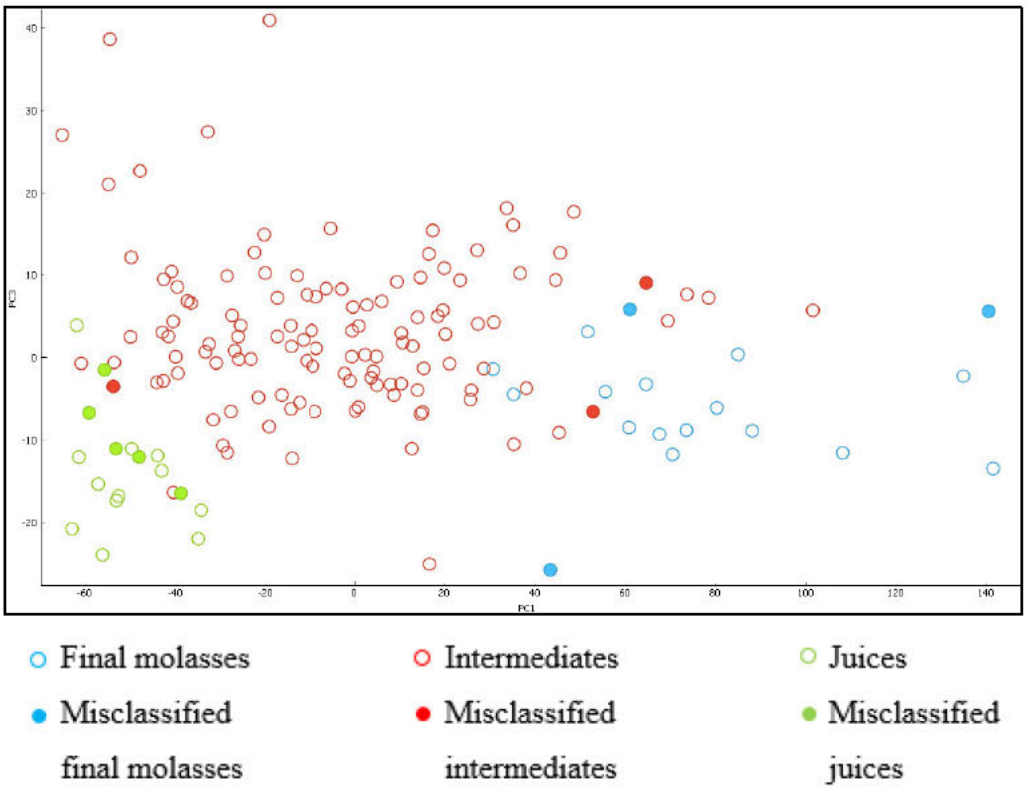

Figure 4.14: A scatterplot of misclassified samples accounting for 7.0 % of the total amount of samples analysed using the classification tree model

Upon further inspection, it was found that the juice samples had been misclassified as A-massecuites. A-massecuites have similar characteristics to juices. Table 4.4 shows a typical example of A-massecuite and juice NIRS predicted results indicating the similar values between the two different products.

Table 4.4: Example of A-massecuite and juice NIRS predicted results

Sample	Ash (%)	Brix (° Bx)	Fructose (%)	Glucose (%)	Pol (° Z)	Sucrose (%)
A Massecuite	0.49	13.61	0.2	0.19	12.07	12.13
Juice	0.36	13.64	0.21	0.15	12.34	12.27

An evaluation of the performance measures of the model is tabulated in Table 4.5. A low recall of 70.6 % was achieved for juices. This indicates that the model could not successfully predict positive cases for juices in particular. Although the intermediate samples had high scores relating to positive cases (F1 score = 95.6 %, precision = 93.7 %, Recall = 97.5 %), a low specificity value of 77.1 % was obtained, indicating negative samples that were not correctly classified as negative. These results indicate a higher probability that juices and intermediates

may not be identified correctly. The classification accuracy of final molasses was the highest implying that the model is best suited for this particular stream. However, the F1 score of 85.7 %, which is a better performance measure of imbalanced data, suggests that the model is not as accurate for final molasses.

The over-fitted model combined with the individual performance measures of the sugar streams indicates a poor performing model.

Table 4.5: An evaluation of performance measures for the classification tree model

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	96.8	93.0	96.2
Misclassification rate (%)	3.2	7.0	3.8
F1 score (%)	85.7	95.6	80.0
Precision (%)	88.2	93.7	92.3
Recall (%)	83.3	97.5	70.6
Specificity (%)	98.6	77.1	99.3

4.1.3.3 Support vector machine (SVM)

An SVM model was applied to a calibration and an independent validation set using various kernel types. Based on the results in Table 4.6, a linear kernel type was chosen as this gave the highest F1 score of 95.5 % with a corresponding highest classification accuracy of 95.5 %. This presents as an over-fitted model based on the lower validation F1 score of 83.0 % and the validation accuracy of 82.5 %. The calibration model had a misclassified rate of 4.5 %. Figure 4.15 shows a scatterplot of misclassified samples. These samples appear to be on the borders of the intermediates group, thus indicating samples that were spectrally similar to each other.

Table 4.6: An evaluation of the calibration and independent validation sets using various kernel types on the SVM model

Kernel type	Model	Classification Accuracy (%)	Misclassification rate (%)	F1 Score (%)
Linear	Calibration	95.5	4.5	95.5
	Independent validation	82.5	17.5	83.0
Polynomial	Calibration	77.7	22.3	68.0
	Independent validation	32.5	67.5	15.9
Radial Basis Function (RBF)	Calibration	91.1	8.9	90.1
	Independent validation	57.5	42.5	56.1
Sigmoid	Calibration	84.7	15.3	81.4
	Independent validation	55.0	45.0	53.0

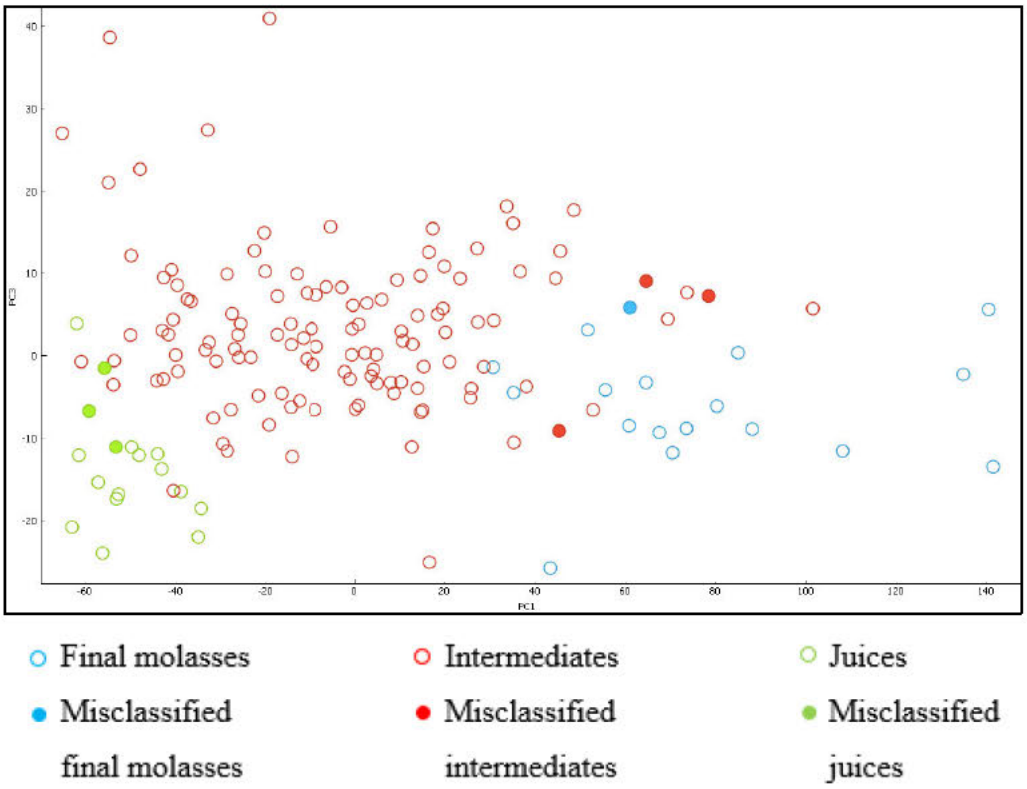

Figure 4.15: A scatterplot of misclassified samples accounting for 4.50 % of the total amount of samples analysed using the SVM model (Linear)

An evaluation of the performance measures of the model is tabulated in Table 4.7. All sugar streams products showed adequate separation. The final molasses had a low F1 score of 89.5 %

compared to the others. This contrasted with the high classification accuracy score of 97.5 %. The F1 score refers to the balance between precision and recall (Hand 2021). Final molasses scored lower because of its lower precision result of 85.0 %, indicating the model's poorer ability to classify final molasses samples correctly. Although juices exhibited precision and specificity of 100.0%, they also had a recall of 82.4 %. This lower value indicates that not all positive samples were correctly classified as positive. This is reflected in Figure 4.15. The model incorrectly predicted more intermediate samples resulting in a low specificity score of 88.6 %.

These results suggest that although the over-fitted model has good predictive powers for the calibration set, the model will have a low probability of predicting the classes accurately for independent validation or an unknown set.

Table 4.7: An evaluation of performance measures for the SVM model (Linear)

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	97.5	95.5	98.1
Misclassification rate (%)	2.5	4.5	1.9
F1 score (%)	89.5	97.1	90.3
Precision (%)	85.0	96.7	100.0
Recall (%)	94.4	97.5	82.4
Specificity (%)	97.8	88.6	100.0

4.1.3.4 Logistic regression

A logistic regression model was applied to calibration and an independent validation set. Two regularisation types were evaluated. Regularisation is when information is added to a model to reduce generalisation error and prevent overfitting (Shaer 2017). Table 4.8 shows that the Ridge model provides a higher F1 score of 95.6 %. This closely ties in with the classification accuracy of 95.5 %, which serves as an added measure of performance. However, the Ridge model has a lower F1 score of 90.1 % for the independent validation set, which corresponds to the lower validation accuracy of 90.0 %. These lower values indicate an over-fitted model.

On the other hand, the Lasso model shows a much lower F1 score and accuracy for the independent validation set. This suggests a model that is over-fitted to a higher degree. The results of Table 4.8 thus prove that the Ridge-type model is better suited to classifying sugar

stream products. A calibration misclassified rate of 4.5 % is represented in Figure 4.16. Three intermediate samples were misclassified as final molasses, and one intermediate sample was taken as a juice. In the scatterplot of PC1:3 in Figure 4.16, it is observed that juices and final molasses are grouped opposite each other. This indicates that the misclassified samples' features closely resemble the incorrect group, final molasses or mixed juice.

Table 4.8: An evaluation of the calibration and independent validation sets using two regularisation types on the logistic regression method

Regularization type	Model	Classification Accuracy (%)	Misclassification rate (%)	F1 score (%)
Lasso	Calibration	93.0	7.0	92.9
	Independent validation	75.0	25.0	75.7
Ridge	Calibration	95.5	4.5	95.6
	Independent validation	90.0	10.0	90.1

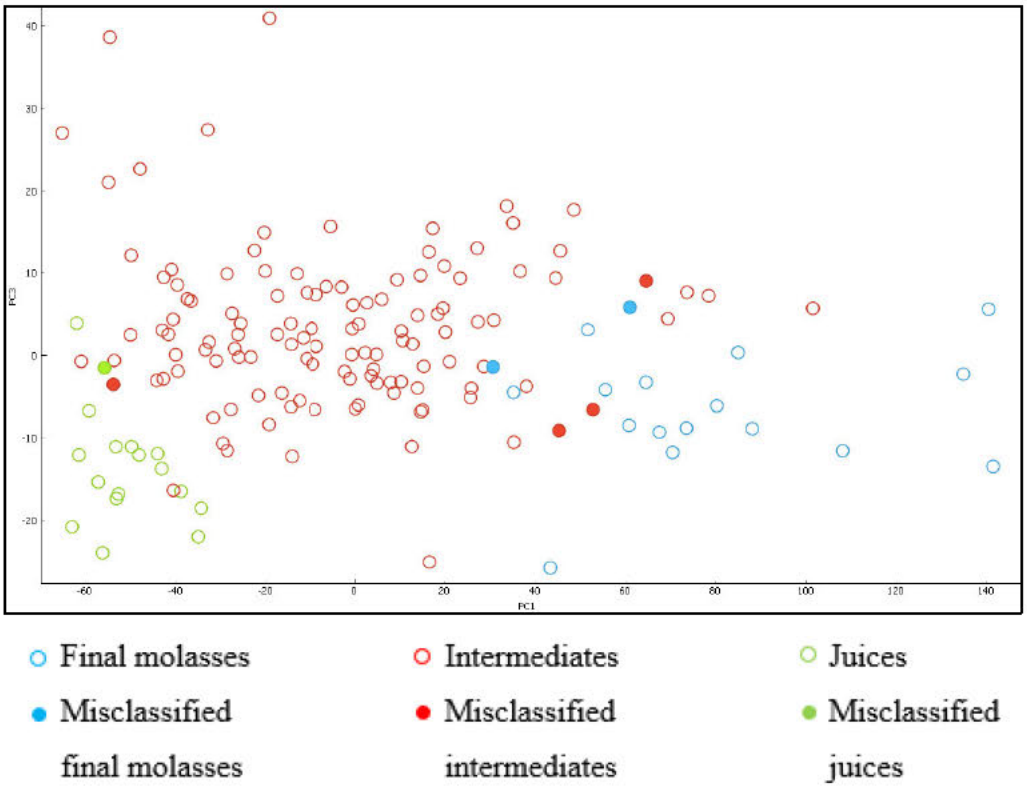

Figure 4.16: A scatterplot of misclassified samples accounting for 4.50 % of the total amount of samples analysed using the logistic regression (Ridge) model

Results of Table 4.9 show that intermediates had a higher F1 score of 97.1 %. Although final molasses had a high classification accuracy of 96.8 %, its F1 score revealed a lower value of 86.5 %. This suggests that the model is not as accurate as the classification accuracy suggests. Final molasses has a lower precision score of 84.2 %. This implies that a higher percentage of samples were incorrectly predicted as false positives for final molasses. However, a high specificity score of 97.8 % indicated the model was useful for correctly identifying negative samples. Intermediate and juice samples results suggest the model is better suited to correctly identifying those streams.

Based on the nature of the over-fitted model and the results of Table 4.9, this model has a low probability of positively identifying final molasses samples.

Table 4.9: An evaluation of performance measures for the logistic regression (Ridge) model

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	96.8	95.5	98.7
Misclassification rate (%)	3.2	4.5	1.3
F1 score (%)	86.5	97.1	94.1
Precision (%)	84.2	97.5	94.1
Recall (%)	88.9	96.7	94.1
Specificity (%)	97.8	91.4	99.3

4.1.4 Optimum model selection

All best calibration models from each model type were compared to determine which accuracy is best for differentiating between final molasses, intermediates and juices. Each model was pre-processed with Gaussian smoothing and used five PCs. The results are found in Table 4.10. Based on the results of Table 4.10, it is recommended that the KNN (3) model be put forward as a possible classification model for samples using absorbances. The logistic regression model had the highest calibration F1 score of 95.6 %; however, the decreased validation result suggested a further over-fit model than the KNN (3). The KNN (3) model had a calibration F1 score of 93.1 % and the independent validation set was a slightly lower 92.7 %. This indicates that the KNN (3) model would be better suited to separate sugar stream products.

Table 4.10: A comparison of accuracies of all models

		Classification accuracy (%)	F1 score (%)
KNN (3)	Calibration	93.0	93.1
	Independent validation	92.5	92.7
Classification tree	Calibration	93.0	92.8
	Independent validation	90.0	90.2
SVM (Linear)	Calibration	95.5	95.5
	Independent validation	82.5	83.0
Logistic regression (Ridge)	Calibration	95.5	95.6
	Independent validation	90.0	90.1

4.1.5 Development of optimum model

The KNN (3) model was then subjected to various PCs and pre-processing methods to determine a KNN (3) model with the optimum number of PCs. Results of this evaluation using varying PCs and pre-processing methods are shown in Table 4.11.

Table 4.11: Results of KNN (3) model using various pre-processing techniques in conjunction with a varying number of PCs

		Calibration		Independent validation	
Pre-processor	**Number of PCs**	**Classification accuracy (%)**	**F1 score (%)**	**Classification accuracy (%)**	**F1 score (%)**
Gaussian smoothing	2	91.7	92.1	90.0	90.3
	3	94.3	94.5	95.0	95.1
	4	93.0	93.1	90.0	90.2
	5	93.0	93.1	92.5	92.7
	6	93.0	93.1	90.0	90.2
Savitzky-Golay filter	2	91.1	91.4	92.5	92.5
	3	94.9	95.0	95.0	95.1
	4	93.0	93.1	87.5	87.7
	5	92.4	92.6	92.5	92.7
	6	92.4	92.6	90.0	90.2
Baseline	2	77.1	76.3	57.5	57.8
	3	90.4	90.5	72.5	72.3
	4	90.4	90.7	82.5	82.4
	5	93.0	93.1	87.5	87.3
	6	93.6	93.8	85.0	84.9
Min-max normalisation	2	94.9	95.0	85.0	85.3
	3	94.9	95.0	92.5	92.6
	4	94.9	95.0	90.0	90.1
	5	94.9	95.0	90.0	90.1
	6	94.9	95.0	92.5	92.6

Results of Table 4.11 indicate that the KNN (3) model using a Savitzky-Golay filter, of window size 5, polynomial order of 2 and set at a zero derivative, with three PCs gave optimum results. The F1 score of the calibration set was 95.0 %. This implies a model of high accuracy. The F1 score of the independent validation set was 95.1 %. The slightly higher results suggest the model is appropriate for classifying the different sugar streams based on their wavenumber and respective absorbances. The calibration set was trained with samples representing a season of

sugar processing from all 14 sugarcane processing factories across South Africa. This ensures a well-trained model that is then capable of effectively predicting unseen data. The baseline pre-processing showed the lowest probability for classifying the products, particularly for two PCs. The F1 scores of the calibration set showed significantly higher accuracies than those of the independent validation set. These implied over-fitted models, indicating that the models will not be able to predict unknown samples accurately.

After pre-processing with Savitzky-Golay, three PCs were used to explain 95% of the variability. PC1, PC2 and PC3 account for 84.3 %, 6.3 % and 4.6 % of the variation respectively. An optimum score plot of PC1:3 is given in Figure 4.17. The shape of this score plot follows the shape of the score plot created when the SMRI-NIRS equations were developed (Walford 2018) thus further confirming the potential of this model. The score plot shows separation was possible; however, there is a small degree of overlap between juices and intermediates and final molasses and intermediates. Ample separation occurred between juices and final molasses.

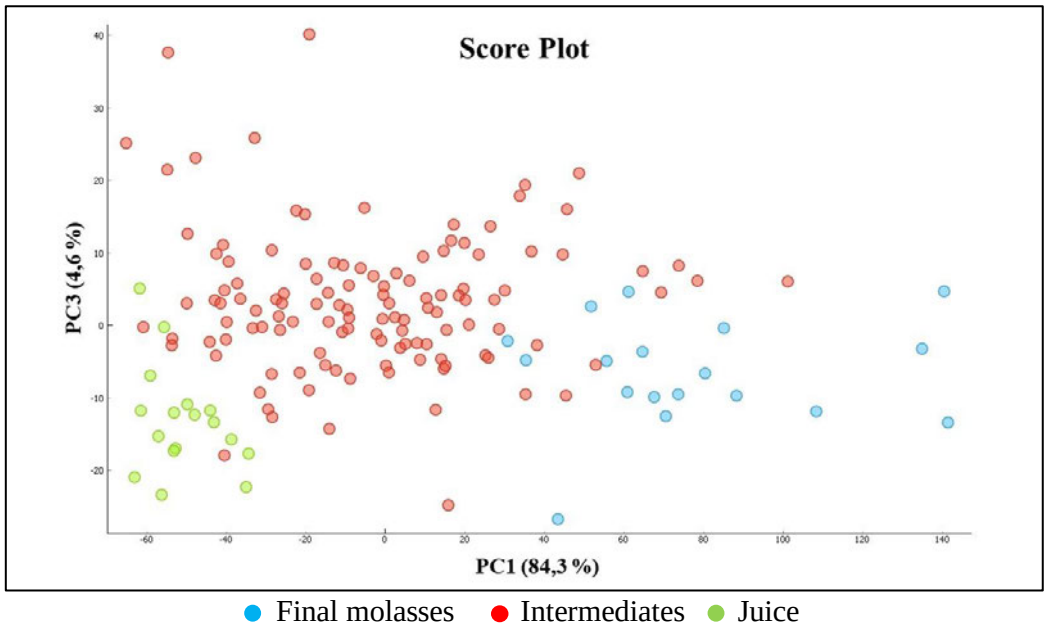

Figure 4.17: PCA score plot of PC1:3 for the optimum model of KNN (3) using the Savitzky-Golay filter

4.1.6 Optimum model performance measures

An evaluation of the performance measures for the optimum model are given in Table 4.12. The F1 score of the calibration set suggests an accurate model. This is confirmed by the higher F1 score of the independent validation set. A misclassification rate of 5.1 % was obtained. This

is depicted in Figure 4.18. One final molasses had misclassified as an intermediate; five intermediates were misclassified as final molasses and two juice samples were misclassified as intermediates. The misclassifications were mainly due to overlaps between the different sugar stream products. The distance between groups on the PC plot was miniscule, indicating that the spectra were very similar. The two juice samples were misclassified as A-massecuites. This suggests that although the model gives high accuracy results, further work needs to be undertaken to improve the Euclidean distance between the samples. More samples of a varying nature need to be added to the calibration set. This may result in a positive or negative change in the overlap.

Table 4.12: An evaluation of the calibration and independent validation set using the KNN (3) model with the Savitzky-Golay filter and three PCs

Model type	Classification Accuracy (%)	Misclassification rate (%)	F1 score (%)
Calibration	94.9	5.1	95.0
Independent validation	95.0	5.0	95.1

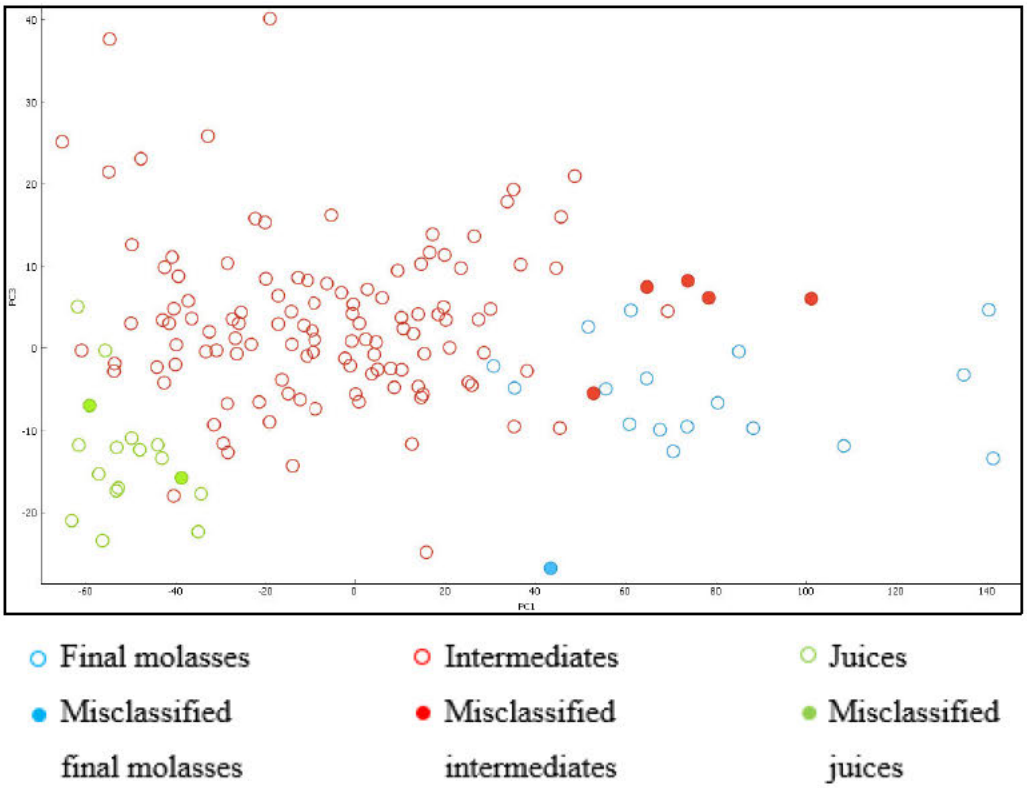

Figure 4.18: A scatterplot of misclassified samples accounting for 5.10 % of the total amount of samples analysed using the optimum KNN (3) model

Table 4.13 shows that intermediates had the highest F1 score of 96.7 %, indicating the model has a higher accuracy for intermediates samples. The classification accuracy for final molasses was 96.2 % which suggests that the model has a high accuracy for final molasses; however, the much lower F1 score of 85.0 % shows that this is not true. This is due to the imbalanced data for final molasses. Final molasses had a low precision score because it had a high rate of false positives. Juices achieved a 100.0 % precision score as the model had zero number of false positives. Although juices had a 100.0 % precision score, the lower recall score of 88.2 % was due to two false negatives.

The overall model results suggest an appropriate model for separation between sugar stream products.

Table 4.13: An evaluation of performance measures of the KNN (3) model with the Savitzky-Golay filter and three PCs

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	96.2	94.9	98.7
Misclassification rate (%)	3.8	5.1	1.3
F1 score (%)	85.0	96.7	93.8
Precision (%)	77.3	97.5	100.0
Recall (%)	94.4	95.9	88.2
Specificity (%)	96.4	91.4	100.0

4.1.7 Conclusion

The results of the KNN (3) model with the Savitzky-Golay filter and three PCs imply a satisfactory model of good predictive powers for unknown samples with known absorbances. The PC plot (Figure 4.3) and scatterplot (Figure 4.18) showed overlap between the three groups which account for the misclassification rates. More samples need to be added to the sample calibration group. This may positively affect the overlap by increasing the distance between the groups, thereby improving the robustness of this particular model.

4.2 Sugar stream product classification based on analyte concentrations

The second model focused on the analyte concentrations for a range of samples reflective of varying geographical and environmental conditions. The analyte concentration ranges are shown in Table 4.14.

Table 4.14: Range of analyte concentrations for each product

	Final Molasses	Juice	Intermediate
Pol (° Z)	27.9 - 36.0	9.1 - 13.0	36.1 - 87.5
Brix (° Bx)	74.0 - 85.7	10.9 - 15.6	57.2 - 96.6
Sucrose (%)	30.5 - 37.6	9.2 - 13.2	35.9 - 89.5
Glucose (%)	2.9 - 8.1	0.1 - 0.4	1.0 - 8.4
Fructose (%)	5.7 - 9.2	0.2 - 0.5	0.5 - 8.7
Ash (%)	9.1 - 16.1	0.5 - 0.67	1.1 - 13.2

4.2.1 Rank

Several scoring methods were investigated. These included information gain, information gain ratio and relief. Using the rank scoring system, glucose was found to be a negligible variable and was thus excluded from the development of the model. A distribution plot shown in Figure 4.19 shows a poor ranking score for glucose with final molasses, intermediates and juices. Alternatively, Figure 4.20 shows a distribution plot of pol, which scored the highest, making it the most useful variable for classifying sugar stream products.

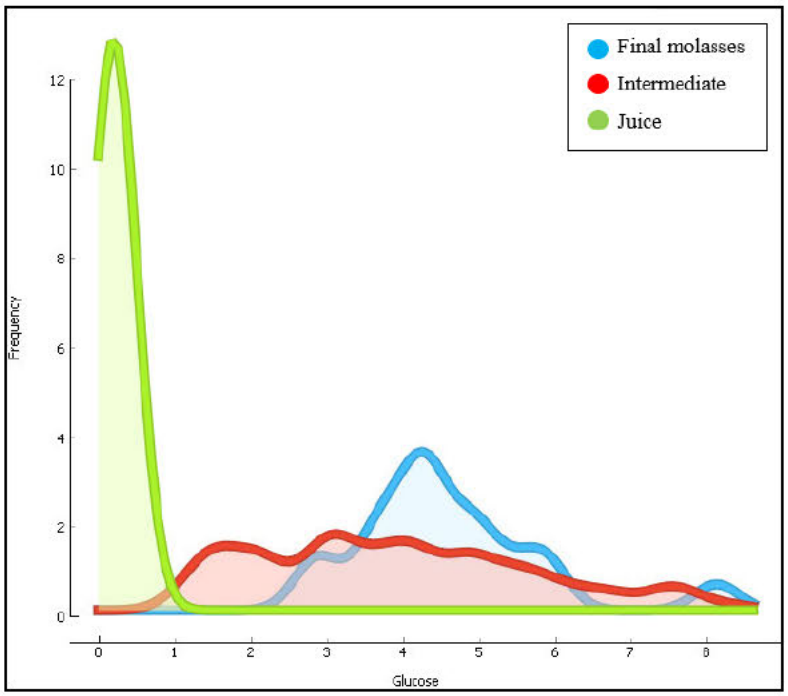

Figure 4.19: Distribution plot of Frequency vs Glucose

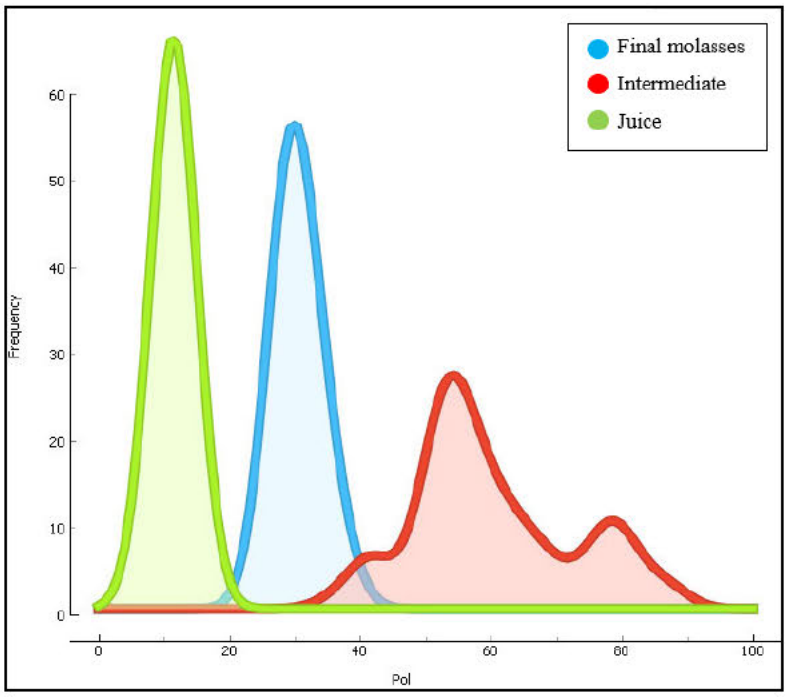

Figure 4.20: Distribution plot of Frequency vs Pol

4.2.2 Principal component analysis (PCA)

PCA was applied to the remainder of the variables after glucose was excluded from the data table of analyte concentrations. These were pol, sucrose, brix, ash and fructose. The first two PCs were used to explain 98 % of the variability. PC1 and PC2 accounted for 65.8 % and

32.4 %, respectively. PC1:2 gave good separation of sugar stream products shown in Figure 4.21. Coloured regions were used to depict a good separation.

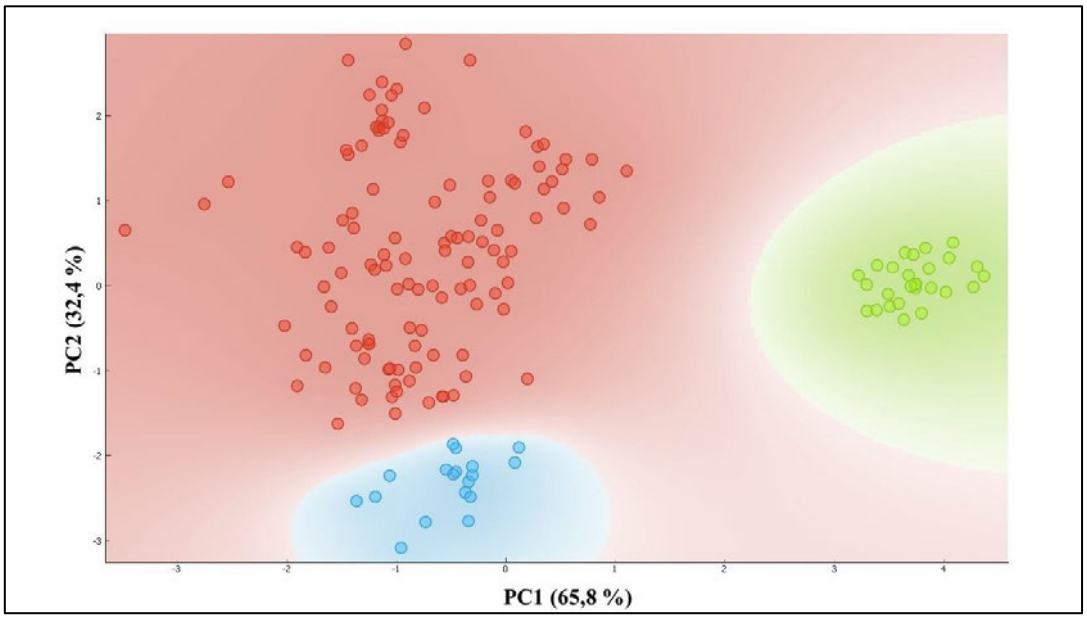

Figure 4.21: PCA score plot of PC1:2

4.2.3 Development of a classification model

All models were pre-processed with two PCs. A score plot of PC1:2 was then used for the development of all subsequent models. A cross-validation sampling method was also applied to all models for consistency. F1 score is used as a measure of accuracy across all models as imbalanced data was used to develop the model.

4.2.3.1 K-nearest neighbour (KNN)

KNN was applied to the PC scores, the optimum K value was determined, and the evaluation results of the calibration and independent validation sets are tabulated below in Table 4.15. The F1 score results of Table 4.15 show that a KNN model with two nearest neighbours gave the highest score of 99.4 % for the calibration set. The independent validation set achieved a lower score of 97.5 %, indicating a slightly overfitted model. Upon inspection, it was found that one intermediate sample was misidentified as final molasses. This confirms the misclassification rate of 0.6 % for the calibration set. This is visualised as a scatterplot in Figure 4.22. An overfitted model occurs when the model does not have a good representation of all sample results and is thus more specific to the calibration set (van der Schaaf 2012). The independent validation set thus has data that is not fully represented in the calibration set. This makes the

model sensitive to any deviating samples. An increase in the number of random samples will therefore improve the model.

Table 4.15: Evaluation of KNN model with various KNN values

K value	KNN Model	Classification Accuracy (%)	Misclassification rate (%)	F1 Score (%)
2	Calibration	99.4	0.6	99.4
	Independent validation	97.5	2.5	97.5
5	Calibration	98.8	1.2	98.7
	Independent validation	97.5	2.5	97.5

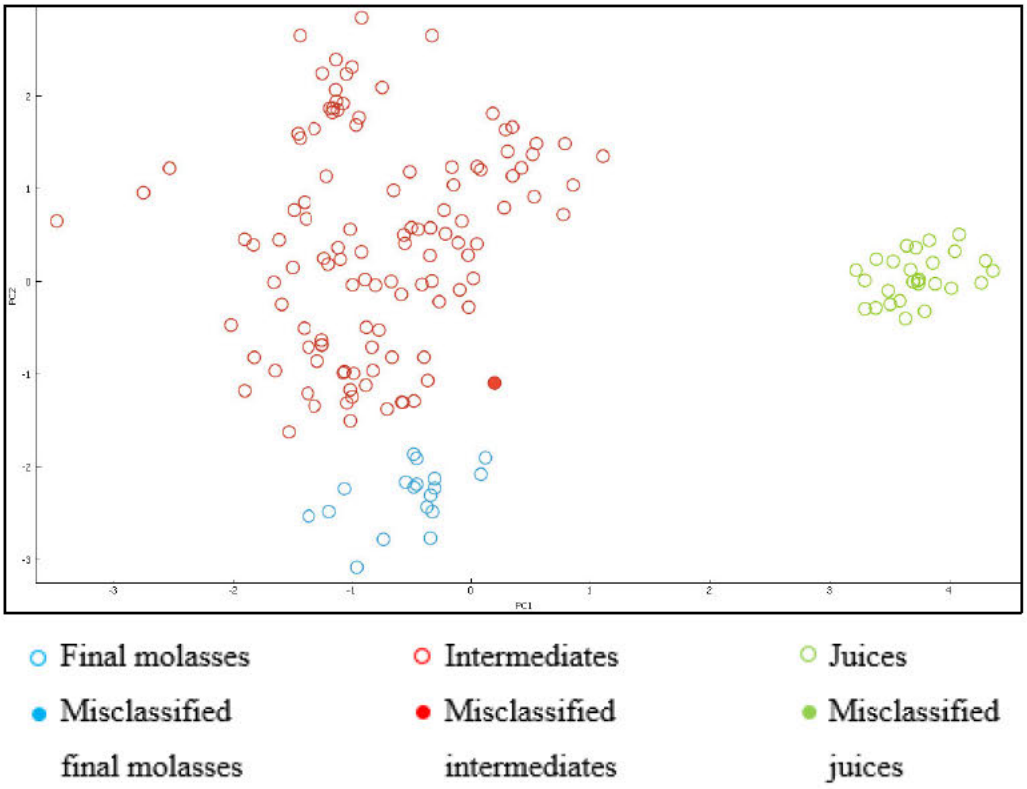

Figure 4.22: A scatterplot of misclassified samples accounting for 0.60 % of the total amount of samples analysed using the KNN (2) model

An evaluation of the performance measures for the KNN (2) model is tabulated in Table 4.16. These results show excellent model capabilities with high scores for all sugar stream products indicating a good separation. In addition, juices received a 100.0 % score for all parameters,

demonstrating the model's ability to correctly identify all positive and negative juice samples. This shows that the model works best for juice samples.

Table 4.16: An evaluation of performance measures for the KNN (2) model

K value	2		
Sugar stream product	**Final molasses**	**Intermediates**	**Juices**
Classification Accuracy (%)	99.4	99.4	100.0
Misclassification rate (%)	0.6	0.6	0.0
F1 score (%)	97.3	99.6	100.0
Precision (%)	94.7	100.0	100.0
Recall (%)	100.0	99.1	100.0
Specificity (%)	99.3	100.0	100.0

4.2.3.2 Classification tree

A classification tree model was applied to calibration and an independent validation set. The results were evaluated and tabulated in Table 4.17. The calibration set had an F1 score of 98.1 %. The independent validation set increased to 100.0 %, implying that the model has a high probability of correctly predicting unknown sugar stream products. A scatterplot of the misclassified samples is shown in Figure 4.23. Two intermediate samples were misclassified as juices. One of these intermediates was identified as a syrup similar to juice, hence the misclassification. One final molasses was incorrectly identified as an intermediate. Figure 4.23 shows that the misclassified final molasses was based at the border of the intermediates group. The two misclassified intermediates were also at the border of the intermediates group but in the direction of the juices. This implies that these misclassified samples were of similar analyte values to the group it had incorrectly predicted as positive.

Table 4.17: An evaluation of the calibration and independent validation sets using classification tree

Model type	**Classification Accuracy (%)**	**Misclassification rate (%)**	**F1 Score (%)**
Calibration	98.1	1.9	98.1
Independent validation	100.0	0.0	100.0

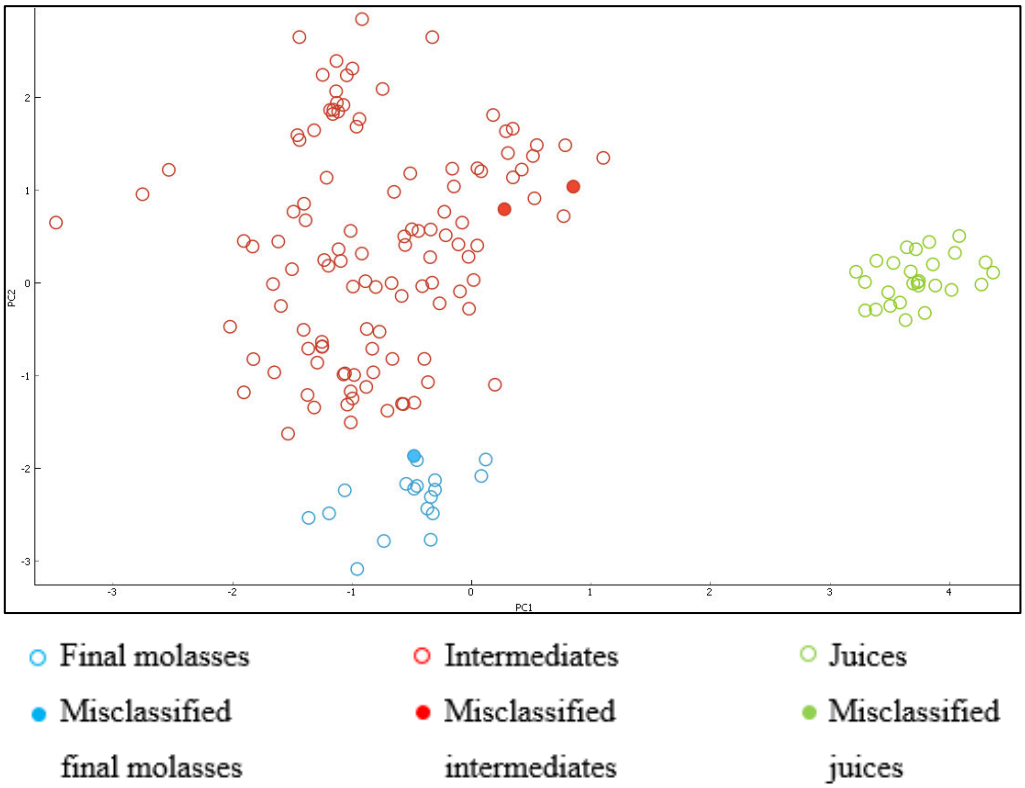

Figure 4.23: A scatterplot of misclassified samples accounting for 7.0 % of the total amount of samples analysed using the classification tree model

Table 4.18 shows that all sugar stream products gave similar high-scoring accuracies. Final molasses, intermediates, and juices had F1 scores of 97.1 %, 98.7% and 96.4 %, respectively. This shows good separation between the products. Final molasses gave 100.0 % scores for precision and specificity, indicating that there were no false positives and that all negative classes were correctly rejected. However, a lower recall score shows that the model could not predict positive final molasses samples as well as it did for intermediates and juices. In addition, juices had a lower precision score since two intermediate samples were falsely identified as juice.

Overall, the scores are relatively high, showing an acceptable model for the classification of the three sugar stream products.

Table 4.18: An evaluation of performance measures for the classification tree model

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	99.4	98.1	98.8
Misclassification rate (%)	0.6	1.9	1.2
F1 score (%)	97.1	98.7	96.4
Precision (%)	100.0	99.1	93.1
Recall (%)	94.4	98.3	100.0
Specificity (%)	100.0	97.8	98.5

4.2.3.3 Support vector machine (SVM)

An SVM method, using various kernel types, was applied to the calibration and independent validation sets. The results are shown in Table 4.19. Linear, polynomial and RBF kernels all gave the same F1 score of 98.7 % for the calibration set and 97.5 % for the independent validation set. The sigmoid kernel type gave unsatisfactory results with an F1 score of 81.9 % and a much lower validation score of 63.5 % indicating a highly over-fit model. Based on literature (Nisbet 2018), the RBF kernel is the preferred type for non-linear data. This is because it is localised and has a finite response along the complete x-axis. The SVM RBF model presented as an overfit model due to the F1 score for the independent validation set being lower than the F1 score of the calibration set. A misclassification rate of 1.2 % was observed in a scatterplot in Figure 4.24. Two final molasses samples were misclassified as intermediates. This implies that the misclassified samples were similar to the intermediates group in analyte concentrations.

Table 4.19: An evaluation of the calibration and independent validation sets using different kernel types on the SVM model

Kernel type	Model	Classification Accuracy (%)	Misclassification rate (%)	F1 score (%)
Linear	Calibration	98.8	1.2	98.7
	Independent validation	97.5	2.5	97.5
Polynomial	Calibration	98.8	1.2	98.7
	Independent validation	97.5	2.5	97.5
RBF	Calibration	98.8	1.2	98.7
	Independent validation	97.5	2.5	97.5
Sigmoid	Calibration	85.6	14.4	81.9
	Independent validation	70.0	30.0	63.5

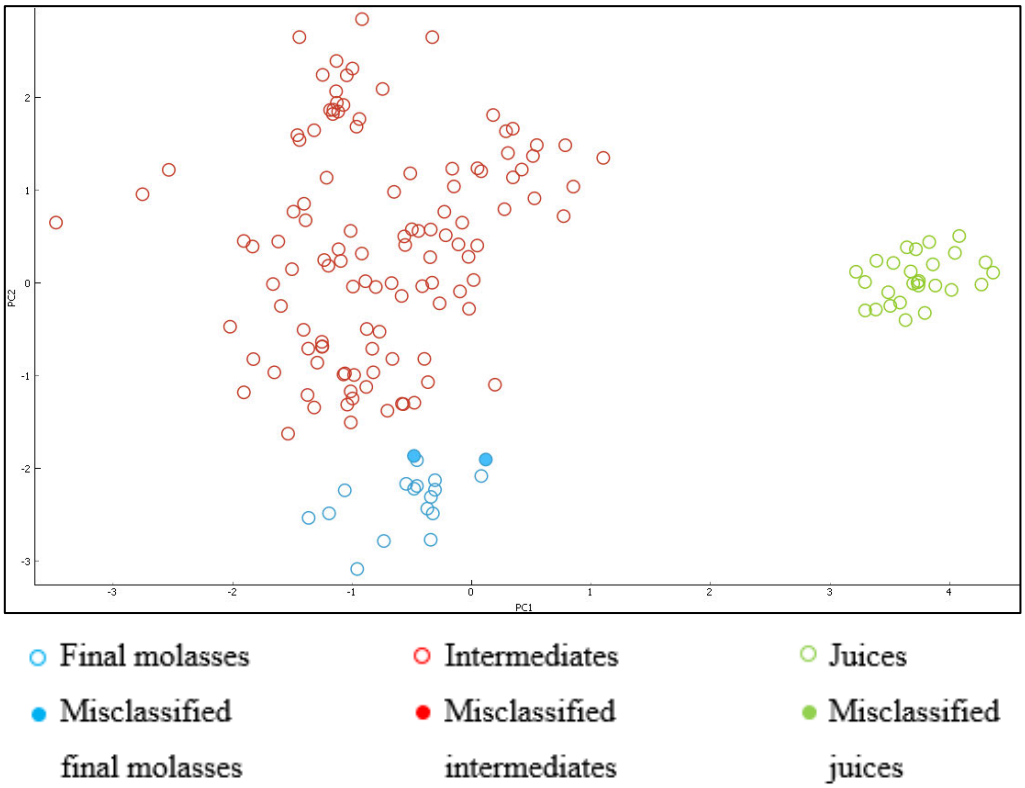

Figure 4.24: A scatterplot of misclassified samples accounting for 1.20 % of the total amount of samples analysed using the SVM RBF model

Table 4.20 gives the performance measures for individual products. F1 scores for the individual sugar stream products suggest that the model is accurate for the intermediate (99.1 %) and juice products (100.0 %). The F1 score for final molasses was slightly lower at 94.1 %. The 100.0 % scores for all juice product parameters imply that the model is highly suitable for classifying juice samples. Final molasses had a lower recall score since the model incorrectly predicted the two final molasses samples as intermediates. Concurrently, intermediates had a lower precision score resulting from the two false positives. Similarly, specificity was lower for intermediates since the model had not rejected the two negative samples.

Overall, the model exhibited high accuracy results and showed a high probability of correctly classifying sugar stream products.

Table 4.20: An evaluation of performance measures for the SVM RBF model

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	98.8	98.8	100.0
Misclassification rate (%)	1.2	1.2	0.0
F1 score (%)	94.1	99.1	100.0
Precision (%)	100.0	98.3	100.0
Recall (%)	88.9	100.0	100.0
Specificity (%)	100.0	95.6	100.0

4.2.3.4 Logistic regression

A Logistic regression method was applied to a calibration and an independent validation set. Both regularisation types, Lasso and Ridge, were applied to the sets and evaluated to determine the optimum model. The results are tabulated in Table 4.21. Both Lasso and Ridge gave the same F1 score of 99.4 %. The Ridge model was over-fit due to the lower validation F1 score. Lasso was chosen as the best type since it gave a validation F1 score of 100.0 %, suggesting the best fit model. This implies that the Lasso-type model has good predictive powers for unknown samples. A low misclassification rate of 0.6 % is depicted in Figure 4.25. Only one sample was misclassified, that is, a final molasses incorrectly identified as an intermediate. Figure 4.25 shows that the misclassified sample was on the outskirts of the final molasses group moving towards the direction of the intermediates group, indicating that the sample may have been of a similar nature to intermediates.

Table 4.21: An evaluation of the calibration and independent validation sets using different regularization types on the logistic regression model

Regularization type	Model	Classification Accuracy (%)	Misclassification rate (%)	F1 Score (%)
Lasso	Calibration	99.4	0.6	99.4
	Independent validation	100.0	0.0	100.0
Ridge	Calibration	99.4	0.6	99.4
	Independent validation	97.5	2.5	97.5

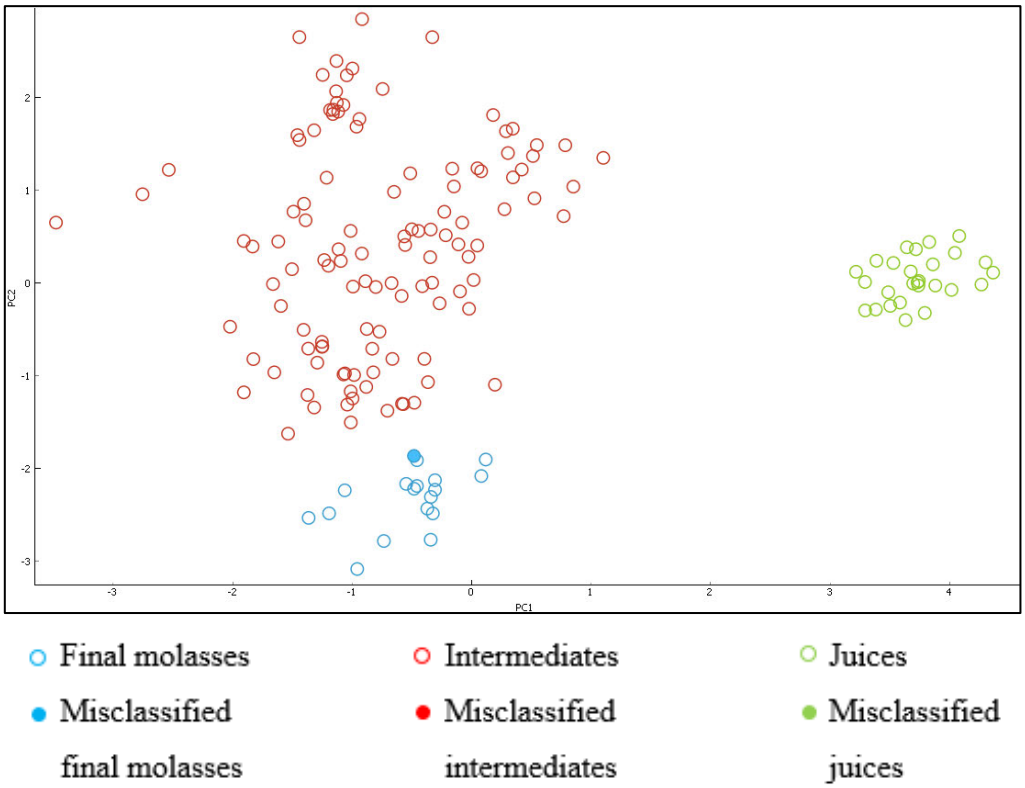

Figure 4.25: A scatterplot of misclassified samples accounting for 0.60 % of the total amount of samples analysed using the logistic regression (Lasso) method

Table 4.22 gives the performance measures of the individual products. The F1 scores of the sugar stream products were all relatively high with the highest score at 100.0 % for juices. The lower F1 accuracy of 97.1 % for final molasses correlates with the lowest recall value of 94.4 %. This recall value was obtained because the classifier model incorrectly predicted a final molasses sample as an intermediate. This was reflected in the slightly lower precision score of 99.1 % for intermediates, in which the one misclassified final molasses sample presented as a false positive for intermediates. This inaccuracy was echoed in the lower specificity of 97.8 % for intermediates. Thus, the intermediates group had failed to reject all the negative classes correctly. This model holds a high affinity for juices as it had 100.0 % scores of all parameters.

Overall, this model performed very well and showed a high probability of correctly predicting and classifying the sugar stream products.

Table 4.22: An evaluation of performance measures for the logistic regression (Lasso) method

Sugar stream product	Final molasses	Intermediates	Juices
Classification Accuracy (%)	99.4	99.4	100.0
Misclassification rate (%)	0.6	0.6	0.0
F1 score (%)	97.1	99.6	100.0
Precision (%)	100.0	99.1	100.0
Recall (%)	94.4	100.0	100.0
Specificity (%)	100.0	97.8	100.0

4.2.4 Optimum model selection

All optimum calibration models from each model type were compared to determine the best accuracy for differentiating between final molasses, intermediates and juices based on their analyte concentrations. Each model had used two PCs. The results are found in Table 4.23 below. Based on the results of Table 4.23, it is recommended that the logistic regression (Lasso) model be put forward as a possible classification model for samples using analyte concentrations. The KNN (2) and logistic regression (Lasso) models had the highest F1 scores of 99.4 % pertaining to the calibration set. The logistic regression was then determined as the optimum model since the independent validation set yielded a higher F1 score at 100.0 %. This indicates the best fit model capable of accurately predicting unknown sugar stream samples based on a well-represented calibration set.

Table 4.23: A comparison of accuracies of all models

		Classification Accuracy (%)	F1 Score (%)
KNN (2)	Calibration	99.4	99.4
	Independent validation	97.5	97.5
Classification tree	Calibration	98.1	98.1
	Independent validation	100.0	100.0
SVM (RBF)	Calibration	98.8	98.7
	Independent validation	97.5	97.5
Logistic regression (Lasso)	Calibration	99.4	99.4
	Independent validation	100.0	100.0

4.2.5 Development of optimum model

The logistic regression (lasso) model was then subjected to varying PCs to determine a logistic regression (lasso) model with the optimum number of PCs. Results of this evaluation using varying PCs are shown in Table 4.24.

Table 4.24: Logistic regression (Lasso) model with varying number of PCs

	Calibration		Independent validation	
Number of PCs	**Classification accuracy (%)**	**F1 score (%)**	**Classification accuracy (%)**	**F1 score (%)**
2	99.4	99.4	100.0	100.0
3	99.4	99.4	100.0	100.0

The F1 scores of both the calibration and independent validation sets did not change when the number of PCs was increased from two to three. The logistic regression (lasso) model with two PCs was then kept as the optimum model. As described in **4.2.3.4**, this was the best fit model, with the independent validation set scoring higher than the calibration set. This suggests that the calibration set was trained with well-represented samples that could effectively predict unseen data. Table 4.22 describes the accuracy results of the individual products. The high F1 scores of the three sugar stream products indicate the model's ability to correctly predict unknown sugar stream samples. The overall model results suggest a model with good predictive powers for separation between sugar stream products.

4.2.6 Conclusion

The results of the logistic regression (lasso) model with two PCs imply a satisfactory model of good predictive powers for unknown samples with known analyte concentrations. Furthermore, the PC plot (Figure 4.21) and scatterplot (Figure 4.25) show that there is a good separation between the three groups.

4.3 Comparison between the optimum model for absorbances and the optimum model for analyte concentrations of sugar stream products

The KNN (3) model with the Savitzky-Golay filter and three PCs was compared with the logistic regression (lasso) model with two PCs. The KNN (3) model was developed using the absorbances of the samples from a wide range of sugar stream products. The logistic regression

model was designed using the analyte concentrations of those same samples. The objective of this thesis was to determine which model development is more accurate for sugar stream products.

4.3.1 Evaluation of the F1 scores and classification accuracies for the two models

Table 4.25 shows results of the two models based on the different types of data from the same samples. The logistic regression (lasso) model suggests a higher accuracy with an F1 score of 99.4 % for the calibration set and a higher validation F1 score of 100.0 %. This implies the best fit model with excellent predictive powers for unknown sugar stream product samples. Tables 4.21, 4.22 and 4.24 and Figure 4.25, all show results for a model with excellent capabilities. Figure 4.25 depicts a model with classification groups that are separated. Table 4.22 shows tabulated results for the individual sugar stream product groups. These results showed the model's high accuracy values for all three groups. The minor misclassification rate was due to only one sample that was misclassified. This is acceptable for the overall 99.4 % accuracy of the model.

Table 4.25: An evaluation of F1 scores and classification accuracies for both models

Model		Classification Accuracy (%)	F1 Score (%)
KNN (3)	Calibration	94.9	95.0
	Independent validation	95.0	95.1
Logistic regression (lasso)	Calibration	99.4	99.4
	Independent validation	100.0	100.0

4.4 General comments, Recommendations and Conclusion

The project aimed to use chemometric methods to develop an optimum classification model. In order to obtain information to develop the models, a product had first to be manually chosen. This approach may appear contradictory as by manually choosing the product, a discrimination analysis is essentially conducted. However, when the model is applied to intermediate products, it can then be used to differentiate between the sugar stream products within the intermediate group. The intermediates consist of A-, B-molasses, A-, B-, C-massecuite, and syrup, which, whilst all of which use the same equations, is important in the industry that they are analysed as individual products for process control purposes. It is thus necessary for the model to be able to assign the correct equation to the sugar stream product within the intermediate group.

Furthermore, mixed juice and molasses products have different analyte concentrations even when analysed using the intermediate equations. Thus, the results will still be able to indicate a product type. Table 4.26 gives an example of a mixed juice and final molasses sample analysed with intermediate equations. The results show the differences between the two groups based on their analyte concentrations.

Table 4.26: Results of mixed juice and final molasses using the intermediate equations

Product	Ash	Brix	Fructose	Glucose	Pol	Sucrose
Mixed juice	3.1	92.3	1.7	1.5	79.7	80.9
Final Molasses	6.5	73.2	6.1	4.9	35.5	37.6

The two classification models that were explored (Classification models using sample absorbances or sample analyte concentrations) can be improved upon in the following ways:

1. The classification model-based absorbances should be improved on by excluding the water peaks before pre-processing. This will eliminate common peaks due to water. The PCA may improve, thus providing a better model,
2. Adding more sample data. This can assist in having more significant groups of defined sugar stream products,
3. All outlier values should be taken into consideration, and these should be analysed together with other variables. Although this may result in inaccurate predictions, once more data is included, the predictive power will increase and
4. By tuning the model. Various parameters can be changed to improve the robustness of the model, such as other pre-processing methods and models that were not explored in this study.

In conclusion, analyte concentration values are best used to develop an accurate and reliable model for classifying sugar stream products, namely final molasses, intermediates, and juices. The concentrations of the various analytes can display the differences between the different sugar stream products, whereas this cannot be as easily done with the spectra. This study has the potential to be used in the South African sugar industry to rapidly and accurately classify a sugar stream product so that the appropriate quantification methods can be automatically chosen. This will create more efficiency in the factory, enabling management to make decisions that will positively contribute to the effective running of a sugar mill.

4.5 References

Hand, D.J.; Christen, P. and Kirielle, N. F. (2021). F*: an interpretable transformation of the F-measure. Machine Learning. 110 (1), p451–456.

Nisbet, R.; Miner, G. and Yale, K. (2018). Handbook of Statistical Analysis and Data Mining Applications. In: Nisbet, R; Miner, G and Yale, K Basic Algorithms for Data Mining: A Brief Overview. Handbook of Statistical Analysis and Data Mining Applications. 2nd ed. USA: Academic Press. p121–147.

Shaer, L.; Kanj, R.; Joshi, R.; Malik, M. and Chehab, A. (2017). Regularized Logistic Regression for Fast Importance Sampling Based SRAM Yield Analysis. 18th International Symposium on Quality Electronic Design. p119-124.

Shahraki, H.R.; Pourahmad, S. and Zare, N. (2017). *K* Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data. BioMed Research International. 2017, p1- 9.

van der Schaaf, A.; Xu, C.; van Luijk, P.; van't Veld, A.; Langendijk, J. and Schilstra, C. (2012). Multivariate modeling of complications with data driven variable selection: Guarding against overfitting and effects of data set size. Radiotherapy and Oncology. 105 (1), p115-121.

Walford, S. (2018). The SMRI-NIRS Technology: Development, validation and application. [Poster]. South African Sugar Technologists' Association, 14 – 16 August, International Convention Centre Durban.