THESIS

PREDICTING THE COMPOSITION OF SUCROSE–ETHANOL–WATER MIXTURES AND ALCOHOLIC BEVERAGES USING DENSITY, REFRACTIVE INDEX, AND TEMPERATURE DATA

Submitted by

Rishi Vojjala

Department of Chemical and Biological Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2025

Master's Committee:

Advisor: David Dandy

David Morett
Edwin Chong
Kenneth Reardon

ABSTRACT

PREDICTING THE COMPOSITION OF SUCROSE–ETHANOL–WATER MIXTURES AND ALCOHOLIC BEVERAGES USING DENSITY, REFRACTIVE INDEX, AND TEMPERATURE DATA

Accurate alcohol content measurements are a critical regulatory requirement in the alcoholic beverage industry, where bottles labeled with the percentage of alcohol by volume must be within ± 0.3 percentage points of the actual value. While large-scale producers use expensive analysis equipment, smaller breweries and distilleries often rely on manual hydrometer readings, which are prone to error and inefficiency. This project explores a cost-effective alternative to estimate beverage compositions, by using widely available in-line instruments such as Coriolis meters and refractometers. The primary aim is to predict the composition of an unknown solution in terms of Brix for sugar content and percentage alcohol by weight (ABW%) for alcohol content, using the physical properties of temperature, density, and refractive index of the solution. Solutions with known compositions were prepared in the lab, and their corresponding physical measurements were recorded across a wide range of mixture compositions and temperatures. These data were used to train regression models that learned the relationship between physical properties and solution composition. Given the temperature, density, and refractive index of an unknown solution, the models could then predict the mixture composition. The Brix model achieved high accuracy, with most predictions within ±1 Brix. Alcohol predictions were less precise, typically within ± 2 ABV percentage points. The models were validated using both lab-made mixtures and commercially produced beverages. The models performed comparably on both, showing strong generalization. The results also suggest that prediction accuracy would be improved by collecting more measurements. With additional parameters, this approach could be extended beyond three-

component mixtures to more complex formulations, and beyond alcoholic beverages to applications in industries such as chemical processing, cosmetics, and food quality monitoring.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. David Dandy, and Mr. David Morett from Emerson, for their unwavering guidance and support throughout this project. Their consistent involvement with weekly meetings and invaluable mentorship played a pivotal role throughout my research endeavor. I am also thankful to my committee members, Dr. Edwin Chong and Dr. Kenneth Reardon, for their valuable input and support during the project. Additionally, I am grateful to Emerson for initiating and sponsoring this study. I would like to thank Mr. Anand and Mr. Joel from Emerson for their valuable insights and contributions to the project.

I am deeply grateful to my mother for her unwavering encouragement and support, no matter what path I choose to pursue. I am thankful to my friends, whose support extended far beyond encouragement, from introducing me to helpful tools to sharing advice on managing deadlines, throughout my time in grad school. Their insights, emotional guidance, and even ensuring I was properly fed and cared for played a crucial role in helping me navigate my academic journey. I genuinely appreciate every conversation, meal, and check-in during the most stressful times. While I had every intention of naming each friend and highlighting their individual contributions, such acknowledgments would exceed the space constraints of this document. In the final stretch of thesis writing, I found myself pet-sitting six dogs and ten cats. Their companionship and occasional chaos brought me comfort during some of my most stressful moments.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In the food and beverage industry, the Alcohol and Tobacco Tax and Trade Bureau (TTB) requires that the alcohol content labeled as “Alcohol by Volume” on beverages must be within ± 0.3 percentage points of the actual value (Alcohol and Tobacco Tax and Trade Bureau, 2022). Hence, breweries and distilleries have a crucial step in measuring alcohol before bottling their products. The current inline measurement options are limited, with Anton Paar’s Cobrix considered the standard for accuracy. However, these systems are expensive, with the Cobrix starting at $95,000, which is not economical for smaller breweries or distilleries (Anton Paar, n.d.). Samples of the beverages also have to be sent to the instrument manufacturer for analysis when a component is changed, which leads to inefficiencies and production downtime. These instruments do not measure flow properties such as volumetric flow rates but measure a specified property such as alcohol and sugar content, or carbonation. Smaller industries use an alternative to these inline instruments by measuring the alcohol content using offline analog techniques such as a hydrometer. This requires making batches of products to be measured before bottling instead of a continuous bottling process. Analog instruments also rely on manual reading of a scale, which can be prone to errors by human observation.

In partnership with the industry sponsor Emerson, this project was initiated, to measure the composition of unknown alcoholic beverages using the already prevalent flow measurement devices in industries such as Coriolis meters and refractometers. Coriolis meters are used for the measurement of density and flow rate, while refractometers are used for the measurement of refractive index. The primary components of alcoholic beverages are sucrose, water, and ethanol. Hence the ternary mixture of sucrose, ethanol, and water was chosen for comparison and measurements. The alcohol content is generally indicated in terms of alcohol by volume as a percentage (ABV), and the sugar content is expressed in the form of Brix degrees. 1 Brix denotes

the presence of 1 g of sucrose in 100 g of a solution. If we consider the ternary solution, density, and refractive index are both functions of composition and temperature. Hence, the primary goal was to find the ABV and Brix of the solution by measuring the density, refractive index, and temperature of the mixture. This can be looked at as a system of two unknowns and three equations from the density, refractive index, and temperature measurements. This would be a competitive alternative to the expensive Cobrix setups, offering a more accessible solution to the smaller industries using Emerson's Coriolis meters and refractometers (Emerson, 2025a). The project involved collecting density and refractive index measurements across a range of temperatures and compositions for the ternary mixture. This data was used to train a machine learning model to predict ABV and Brix of an unknown composition. The model would be later validated using beverage samples from distilleries.

Accurate density measurements were required for the model training, which could be done by analog instruments such as a hydrometer and pycnometer, or by digital devices such as a Coriolis flow meter. A hydrometer consists of a glass tube with a weighted bottom made of lead or steel, with a long narrow stem showing a scale. It is immersed in a liquid till it floats and the reading on the stem's specific gravity scale. A pycnometer, on the other hand, is a flask of known volume, and the density is determined by weighing the filled flask. While a hydrometer can achieve an accuracy of ± 0.01 g/cm^3, a pycnometer's accuracy can reach ± 0.0001 g/cm^3 (Mettler Toledo, n.d.-a). However, these instruments are not very reliable due to errors in the reading of scales by a human eye and potential issues of manually increasing the fluid's temperature. Hence, with the partnership from Emerson, it was initially proposed to use their Coriolis meter for better accuracy.

The Coriolis meter is an in-line density measurement instrument, consisting of two parallel tubes through which the flowing fluid is evenly split. A driver causes the tubes to oscillate in opposite to each other at the natural resonant frequency of the tubes. Magnet and coil assemblies

called pickoffs are placed at the two measurement locations along the tube, as illustrated in **Figure 1**. Each pickoff produces voltage signals that create sine waves, which indicate the motion of one tube relative to the other. When there is no fluid flow, the sine waves of the inlet and outlet pickoff are synchronized i.e., in-phase. When a fluid flows through the tubes, the force arising through the motion of fluid inside the tubes called Coriolis force, causes the tubes to twist in opposite directions. This creates a phase shift between the inlet and outlet sine waves. The time delay (Δt) between the sine waves is directly proportional to the mass flow rate of the fluid ($\dot{m}$), as a higher mass flow would induce greater Coriolis force effects (Emerson, 2025b). The oscillation frequency of the tubes depends on the density of the fluid. A denser fluid will result in the tubes being heavier, resulting in a lower natural frequency. The fluid density (ρ) is inversely proportional to the square of the frequency (f). After calibrations with high and low-density fluids, the density of an unknown fluid can be found by measuring the frequency of oscillation (Weinstein, 2010). The stiffness of the tubes is constant, implying that the mass flow rate and density are the only factors affecting the frequency, allowing the calculation of the fluid volumetric flow rate (V). Hence, the Coriolis meter can be used to measure the density, mass, and volumetric flow rate of a fluid passing through it, noting that

$$\dot{m} \propto \Delta t$$

$$\rho \propto \frac{1}{f^2}$$

$$V \propto \frac{\dot{m}}{f^2}$$

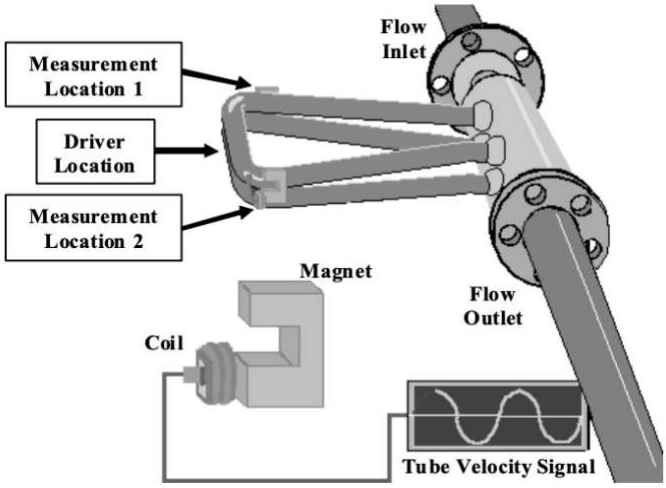

Figure 1: Anatomy of a Coriolis Meter (Weinstein, 2010)

The other property of the mixtures to be measured is refractive index, which is a dimensionless property, defined as the ratio of the speed of light in vacuum to the speed of light in a given fluid (Reis et al., 2010). Digital refractometers measure the refractive index using Snell's law, which states that the product of the refractive index of the fluid and sine of the angle between the incident ray of light and the normal to the surface of the fluid, is always constant: $n \cdot sin\,\alpha = constant$, where n is the refractive index of the medium the light is passing through and α is the angle of reflection or incidence of ray of light.

As shown in **Figure 2**, a digital refractometer consists of a light-emitting diode, whose emitted light passes through a polarization filter, an interference filter, and a focal lens to reach the sapphire prism where the sample is located (Mettler Toledo, n.d.-b). The reflected light is directed through a lens to an optical sensor, which determines the critical angle to measure the refractive index. The refractive index is dependent on the temperature of the sample and the wavelength of the light. A light source of wavelength of 589 nm, also known as the sodium D-line, is highly stable and reliable. Hence, it is widely used in digital refractometers and the refractive index is expressed as nD.

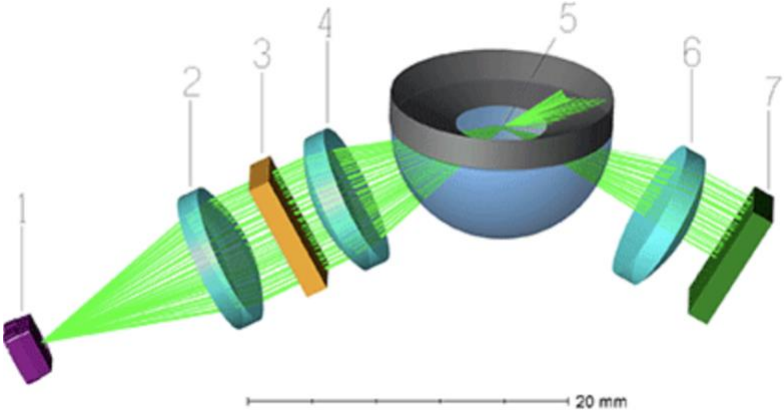

Figure 2: Anatomy of a digital refractometer with its components: (1) light-emitting diode, (2) polarization filter, (3) interference filter, (4) focal lens, (5) sapphire prism, (6) focal lens, (7) optical sensor (Mettler Toledo, n.d.-b)

The initial setup proposed was to use a Coriolis meter, combined with an inline refractometer, to measure the density and refractive index of the sampling flowing through it. The fluid volume inside the Coriolis meter was 250 mL. The challenges of preparing large volumes of mixtures in liters for the flow setup would not be practical for a large range of compositions to be measured. Hence alternatives were considered for a flow setup. With the support of Emerson, a static benchtop densitometer and refractometer were obtained. This allowed for multiple measurements using smaller sample preparations of less than 50 mL. The benchtop densitometer used the same principles of vibration of a small tube cell as that of a Coriolis meter to measure density. A sample mixture would be injected into it and the measurements can be taken without the need of a flow setup.

Machine learning was employed to predict the composition of solutions. The alcohol and sugar content are treated as dependent variables, and the inputs used to determine them are density, refractive index, and temperature, which are the independent variables. Machine learning methods are mainly categorized into supervised, unsupervised, and reinforcement learning.

Reinforcement learning involves an agent that interacts with the dataset without any input or output factors and it learns to make decisions by taking feedback in the form of rewards or penalties. Unsupervised learning deals with datasets where outputs are not well defined by attempting to infer patterns in the data. Our project primarily deals with supervised learning, in which the goal is to train a model to map a function from the inputs to the correct outputs using a training dataset. This project focused on supervised learning models such as linear regression, random forests, support vector regression, and neural networks.

Linear regression is the most basic of the regression models. It models the relationship between the dependent and independent variables by fitting a straight line through the data. The straight-line is used in the case of a two-dimensional dataset, it can be considered a hyperplane for higher-dimensional data (Géron, 2019). To find the best fit, it tries to find the hyperplane with the least sum of squared between actual and predicted values. The basic model of linear regression can be denoted by:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

where $\hat{y}$ is the predicted value, n is the number of features, x_i denotes the feature values and θ_i denotes the model parameters or weights. Linear regression's predictive accuracy tends to be lower than other complex models, but it serves as a useful benchmark for basic analysis and comparison with other models.

Unlike linear regression, which aims to minimize the error between predicted and actual values by fitting a line through all data points, support vector regression focuses on maintaining the predicted values within a specified tolerance level known as epsilon (ε). While linear regression can be thought of as a straight line passing through the data, support vector regression can be thought of as fitting a tube instead of a line through the data with a defined width of epsilon.

It seeks to find a function that fits the majority of the data through the tube and the value of the tolerance can be varied to include or exclude outliers. Data points that fall outside the ε-insensitive tube are referred to as support vectors, and they contribute to the construction of the final model (Theobald, 2024). Different kernels can be used for the model to also work with nonlinear relationships, such as a polynomial or a sigmoid kernel. The kernels transform input space to allow complex fitting and curved prediction boundaries.

Random Forest is an ensemble method based on decision trees. Decision trees learn from a hierarchy of if-else questions, leading to a decision. To build a tree, the algorithm searches over possible tests and finds the one that is most informative about the target variable. This recursive process yields a binary tree of decisions, with each node containing a test. This partitioning is repeated until each region in the partition or leaf only contains a single target value or a single regression value (Müller & Guido, 2016). A prediction on a new data point is made by checking which region of the partition of the feature space the point lies in, and then predicting the majority target. For regression, to make a prediction, the tree is traversed based on the tests in each node to find the leaf the new data point falls into. Decision trees are accurate at decoding patterns in training data, but because there is a fixed sequence of decision baths, any variance and outliers can result in poor predictions and also overfitting. Random forests are one way to address these issues, which is a collection of decision trees, where each tree is slightly different. The idea is to use many trees, each of which works well and overfits the data in different ways, the overall overfitting is reduced by averaging the results. To randomize the trees in a forest, the variables selected for dividing the data are randomized and the input data selection for each tree is also randomized. In this project, random forest models have had a relatively high prediction accuracy, as discussed in the results section.

To capture complex relationships, neural network models were also explored. A neural network is a model inspired by the structure of the human neuron. They can be viewed as a generalization of linear regression that performs multiple processing stages to come to a decision (Müller & Guido, 2016). While in linear regression, there is only one step of calculation of the output using weighted coefficients or model parameters, here the process is repeated multiple times. It first computes hidden units that represent the intermediate processing, then applies a nonlinear function called rectifying linear unit, before combining it again using weighted sums to yield the final result. This is also called multilayer perceptron (MLP). In **Figure 3**, each left node represents an input feature, the connecting line represents learned coefficients and the right node represents output which is the weighted sum.

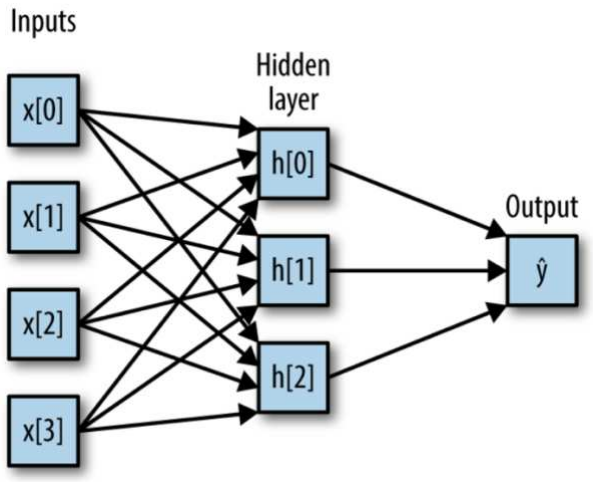

Figure 3: A multilayer perceptron with a single hidden layer (Müller & Guido, 2016)

The weights are set randomly for a neural network before learning starts. It then uses backpropagation, i.e., the model trains itself by updating weights to minimize prediction error. Even when the same parameters are used, a very different final model can be obtained using different seeds (Géron, 2019). It is also sensitive to the hyperparameters, which are settings not

learned by the model but the properties such as learning rate or the number of hidden layers which are set before the training begins.

These models were systematically explored, and the most promising models were further analyzed and tuned to improve predictive accuracy. Their performance was assessed using real beverage samples by industry partners, to assess how well the models generalize beyond the controlled ternary mixtures used in the dataset.

CHAPTER 2: MATERIALS AND METHODS

Instruments and Materials Used

Molecular biology grade sucrose and reagent grade ethanol (200 proof), along with distilled water, were used to prepare the solutions (Fisher Scientific, 2025). The densitometer, refractometer, and analytical balance employed in this study were obtained from Mettler Toledo, with their specifications detailed in **Table 1**.

Table 1: Technical specifications of the experimental equipment (Mettler-Toledo, n.d.-c; Mettler-Toledo, n.d.-d; Mettler-Toledo, n.d.-e)

Instrument	Accuracy	Repeatability	Resolution
Mettler Toledo Densitometer D5 (g/cm^3)	± 0.00005 g/cm^3	± 0.000005 g/cm^3	± 0.00001 g/cm^3
Mettler Toledo Refractometer RX5 (nD)	± 0.00002	± 0.00001	± 0.00001
Mettler Toledo Analytical Balance MS204Sx (g)	± 0.0002 g	± 0.00005 g	± 0.0001 g
Temperature adjustment of Densitometer and Refractometer (°C)	± 0.02 °C	± 0.01 °C	± 0.01 °C

Mass-based vs volume-based alcohol measurement

Alcohol by Volume (ABV %) refers to the percentage volume of ethanol in the total volume of the mixture. However, in ethanol-water solutions, the total volume of the solution is always lower than the sum of the individual component volumes. This is due to intermolecular interactions such as hydrogen bonding and also because the system has a negative excess molar volume (Grolier & Wilhelm, 1981; Parke & Birch, 1999). Due to non-ideal behavior and the difficulty associated with preparing and reading volume-based measurements in flasks, mass-based compositions were used in the project. The composition of the ternary mixtures prepared was indicated using the Percentage of Alcohol by Weight (ABW %) and the mass fraction of sugar denoted by Brix. The percentage of alcohol by weight can be converted to the percentage of alcohol by volume. At a specified temperature T (°C),

$$\begin{aligned} \%\, Alcohol\ by\ Volume &= 100 \times \frac{V_{ethanol}}{V_{solution}} \\ &= 100 \times \left(\frac{\frac{m_{ethanol}}{\rho_{ethanol}}}{\frac{m_{solution}}{\rho_{solution}}} \right) \\ &= (\%\, Alcohol\ by\ Weight) \times \frac{\rho_{solution}}{\rho_{ethanol}} \end{aligned}$$

where V_i denotes the volume, m_i denotes the mass and ρ_i denotes the density of the i^{th} component. The density of ethanol at the specified temperature was obtained from the NIST Chemistry webbook (NIST, n.d.). The density of the solution was determined from measurements taken during the experimental procedure. As the volume of a solution varies with temperature, institutions such as the European Union have set the standard temperature for the ABV % labelling of a beverage to be measured at 20°C (European Commission, n.d.).

Estimation of component masses of a mixture before preparation

To prepare a ternary mixture solution of a specific composition, the weight of each component was measured. The Mettler Toledo equipment requires an injection of 10 mL of sample to fill both the densitometer cell and the refractometer cell completely. Hence the total solution volume to be prepared was chosen to be 30 mL to allow for repeat measurements in case of potential issues such as the presence of air bubbles in the densitometer cell. Given that the density of water is 0.9977 g/mL at 22°C, a ternary solution consisting of 30 grams of water will have a total volume greater than 30 mL. A 30 mL sample mixture would be sufficient for three separate measurements.

For example, in a mixture of composition of 10% ABW and 10 Brix, we can consider the weight of ethanol and sucrose to be x and y (in grams) respectively. Using 30 grams of water, to make the specified composition:

$$\frac{x}{x+y+30} = 0.1$$

$$\frac{y}{x+y+30} = 0.1$$

Solving the equations yielded the mixture to be made of 3.75 g of Ethanol, 3.75 g of Sucrose, and 30 g of Water.

Preparation of solution of specific composition

To describe the preparation, the mixture composition of 10% ethanol by weight and 10 Brix, as calculated above, was selected. Before weighing the components, the weighing station in the analytical balance was cleaned with dry wipes. The tare function on the analytical balance

was used to reset it to 0.0000 g. A clean 100 mL beaker was placed on the scale. The weight of the beaker was noted and then the balance was again tared to 0.0000 g. Using a clean spatula, sucrose was gradually added until the weight was as close to the targeted 3.75 g. The weight of sucrose was noted down. The balance was then tared again. Using a 5 mL pipette and subsequently, a 0.5 mL pipette, water was gradually added until the target of 30 g was reached. The weight of the water was noted down and the balance was tared again. Using the 0.5 mL pipette, ethanol was gradually added to get the target 3.75 g weight. After mixing the solution, it was covered with parafilm to rest for a minimum of 3 h before measurement. This helped in the dissipation of air bubbles which were created from the mixing process. This reduced the possible formation of air bubbles during injection into the densitometer.

This mixture prepared consisted of 3.758 g of sucrose, 3.724 g of ethanol, and 30.381 g of water, with a final total volume of approximately 35 mL. This was a solution of 9.84% ethanol by weight and 9.92 Brix. While the target was 10% ethanol by weight and 10 Brix, it was not practically possible to make 10.00 % ethanol by weight and a 10.00 Brix solution. The objective was to get density and refractive index data for a wide range of compositions.

Experimental Setup

The setup consisted of the densitometer connected in series with the refractometer. This allowed a single sample injection to the densitometer to flow into the refractometer for simultaneous measurement of both density and refractive index (**Figure 4**). Downstream of the refractometer was a three-way valve. The three-way valve directed the excess liquid from the refractometer to the collection container. The valve was also connected to the drying pump DryPro instrument. The drying pump helped in cleaning the instrument by suctioning in the cleaning agents. It was also used in the reverse mode to purge the instrument of liquids and then helped clean it by passing dry air continuously throughout the system. All the tubing between instruments

was provided by the manufacturer, which ensured no leakage. The final outlet tube in the collection beaker was positioned well above the collected liquid level to avoid any backflow.

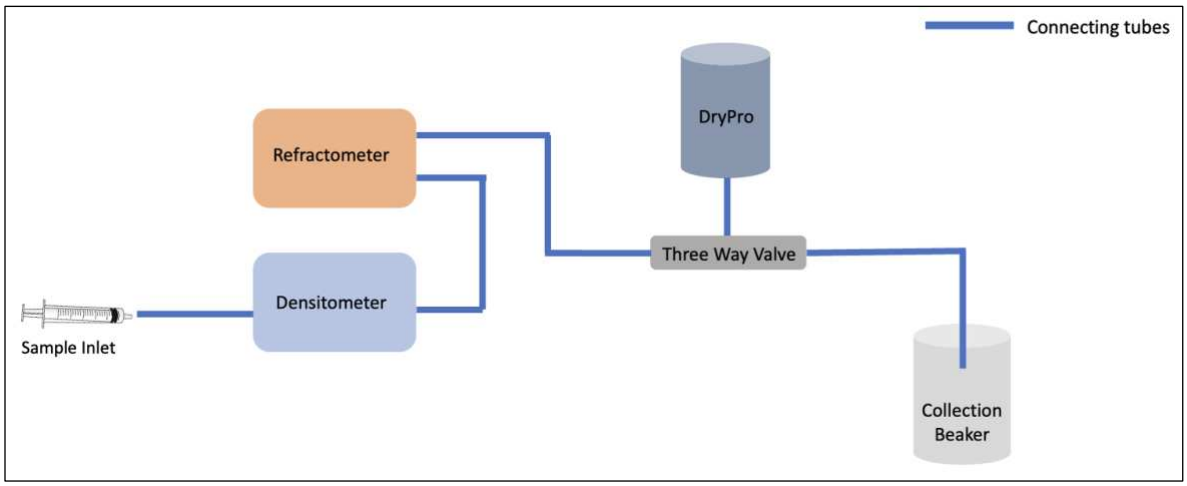

Figure 4: Illustration of the experimental setup.

The densitometer has a display, which allowed for real time inspection of the measuring cell. A USB drive was connected to the densitometer to always back up the data and for the import of the measurement data to a computer. The refractometer used the display of the densitometer for error messages and prompts. The densitometer and refractometer stop the measuring process when they detect the presence of a large number of air bubbles.

Measurement of a solution

The aim was to obtain density and refractive index measurements of a given mixture composition across a temperature range of 10-40°C, in 2°C increments. Before measuring a sample, the equipment undergoes a cleaning cycle to ensure the removal of residual components from the previous sample. A 10 mL syringe was used to draw the sample from the prepared solution beaker. It was then injected into the densitometer inlet. The sample was injected very slowly until the densitometer and refractometer cells were full, and excess flowed into the

collection beaker. It was made sure the injection was slow to ensure laminar flow to avoid the formation of air bubbles. The syringe was not fully depressed to avoid rebounding it, which would introduce air into the system. The densitometer display was checked for the presence of air bubbles. If small air bubbles were observed, additional sample was injected in to push them out. In case of a significant number of air bubbles, the liquid was flushed out and the sample injection was repeated.

Once it was confirmed that there were no air bubbles present, the measurement sequence at the desired temperature range was started. The system starts measuring the density and refractive index at 10°C. Once completed, the system was programmed to automatically increase the temperature by 2°C and start the measurements again. It stops the measurements after the final measurement at 40°C. For a few solutions to investigate the accuracy of the measurements, presence of air bubbles, and potential hysteresis effects, after the final measurement of 40°C, the temperature was decreased in decrements of 2°C, and measurements were taken until it reached 10°C. In the absence of air bubbles in these measurements, there were no hysteresis effects noted and the repeatability of measurements was also excellent, always remaining within each instrument's accuracy of ± 0.00005 g/cm^3 for density and ± 0.00002 nD. After the completion of measurements of a sample, the USB drive was ejected from the densitometer, and the data were backed up to a computer. The USB drive was plugged back into the densitometer and a cleaning cycle was performed on the setup.

System Cleaning and Maintenance

A cleaning cycle was performed before and after each sample measurement. Any remaining solution in the system was first flushed out using the flush function of the drying pump which pushed dry air into the system. A small tube was attached to the densitometer inlet. The tube was then submerged in a beaker of cleaning agent which was water or acetone. First, distilled

water was pumped throughout the system continuously for one minute and the excess flowed into the collection beaker. This was followed by pumping acetone throughout the system for thirty seconds in a similar manner. Dry air was then passed through the system for three minutes to evaporate the acetone and ensure complete drying of the system. The inlet tube was removed at the end of the process.

Instrument maintenance included twice-weekly cleaning, inspection, and calibration of the instruments. The calibration of the densitometer and refractometer was done by using distilled water as the reference standard. If the density and refractive index values were not within the tolerance of ±0.00002 g/cm^3 and ±0.00001 nD, respectively, the instruments had inbuilt adjustment functions to adjust the reading to be within the accuracy range. Similarly, the analytical balance was checked at the start of each day using a certified 0.0500 g standard weight. On certain occasions, the measured calibration weight was not within the ± 0.00002 g tolerance, which was resolved by fixing the level of the balance for proper alignment.

Benchmarking instrument performance from literature comparisons

To verify the accuracy of the instrument and validate the measurement data obtained, measurements were compared to literature values. In **Table 2**, the density and refractive index of pure water were recorded and compared with standard reference values from the National Institute of Standards and Technology (NIST, n.d.-a; NIST, 2024). In **Table 3**, density values for the ternary mixture of ethanol, water, and sucrose at 30°C were compared with data from Resa et al. (2005). This validated the accuracy of the instrument and the experimental process.

Table 2: Comparison of measurements of pure water. Water standard values were obtained from the National Institute of Standards and Technology.

Temperature (°C)	Density standard (g/cm^3)	Density measured (g/cm^3)	Deviation (g/cm^3)	Refractive index standard	Refractive index measured	Deviation
10	0.9997	0.9997	0.00000	1.33407	1.33407	0.00000
15	0.9991	0.99908	-0.00002	1.33376	1.33376	0.00000
20	0.99821	0.9982	-0.00001	1.33336	1.33334	-0.00002
25	0.99705	0.99704	-0.00001	1.33287	1.33287	0.00000
30	0.99565	0.99566	+0.00001	1.33230	1.33231	+0.00001

Table 3: Comparison of ternary mixture density. At 30°C, density measurements of ternary mixtures from Resa et al. (2005) were compared with similar composition prepared in the lab.

Mixture of Composition from Resa et al. (2005)	Density from Resa et al. (2005) (g/cm^3)	Mixture of Composition prepared for comparison	Density measured (g/cm^3)	Absolute difference (g/cm^3)	Percentage Difference
5% ABW, 10.07 Brix	1.025	4.98% ABW, 9.92 Brix	1.02844	0.00344	0.33561
5% ABW, 4.95 Brix	1.004	4.85% ABW, 4.98 Brix	1.00578	0.00178	0.17729
10% ABW, 5.66 Brix	1.001	10.65% ABW, 4.99 Brix	1.0093	0.00831	0.82333
10% ABW, 10 Brix	1.016	9.83% ABW, 9.94 Brix	1.02295	0.00695	0.68406

In **Table 2**, the measurement deviations of pure water between the standard and measured data were within the accuracy range of ± 0.00005 g/cm^3 and ± 0.00002 nD of the equipment. In **Table 3**, the percentage difference between the density values of the compared

ternary compositions was at most 0.82333 %. This can be attributed to the slightly different composition of the mixture prepared for comparison.

Additional validation checks on measured data

Based on the known physical properties of this ternary system, basic consistency checks were employed to verify the accuracy of measured data. These checks help identify potential issues during experimentation such as the presence of air bubbles that could introduce measurement errors. The density of ethanol is lower than the density of water for this study's temperature range of 10-40°C (NIST, 2024; NIST, n.d.). Hence, for a fixed Brix and constant temperature, the density of a mixture would decrease with an increase of the ABW %, and the density of the ternary mixture would be lower than that of a binary solution of sugar and water. Similarly, for a fixed ABW % and constant temperature, the density of a solution increases as the Brix of the composition increases. Comparing two component data of ethanol-water and sucrose-water systems were also used as base checks for measurements of the ternary solution.

Hysteresis checks were also performed to assess measurement repeatability. Density and refractive index measurements were noted as the temperature increased from 10°C to 40°C, and then again as it decreased back to 10°C. Solutions that showed graphical anomalies or repeated measurements that deviated heavily from the accuracy of ± 0.00002 g/cm^3 or ± 0.00001 nD were flagged for potential issues. These flagged solutions were reviewed using recorded videos in the densitometer to check for potential air bubbles and measurements were redone whenever necessary. **Figure 5**, **Figure 6** and **Figure 7** denote a few examples of remeasurements.

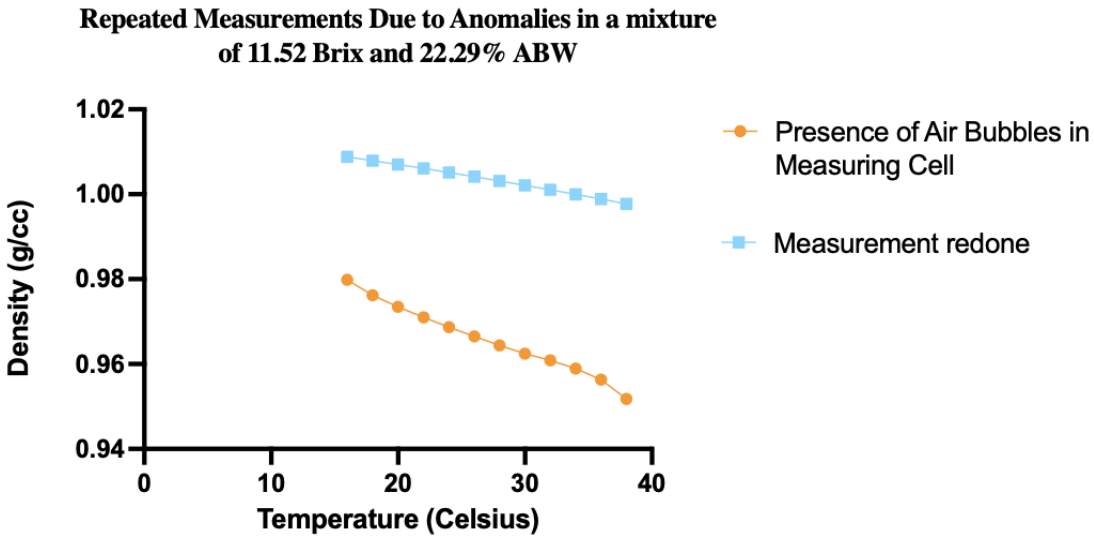

Figure 5: Remeasurement of 11.52 Brix and 22.29% ABW composition solution due to air bubbles. The original measurements showed graphical anomalies indicative of air bubble interference. Upon remeasurement, the correct data was found with a similar graphical trend with other composition measurements.

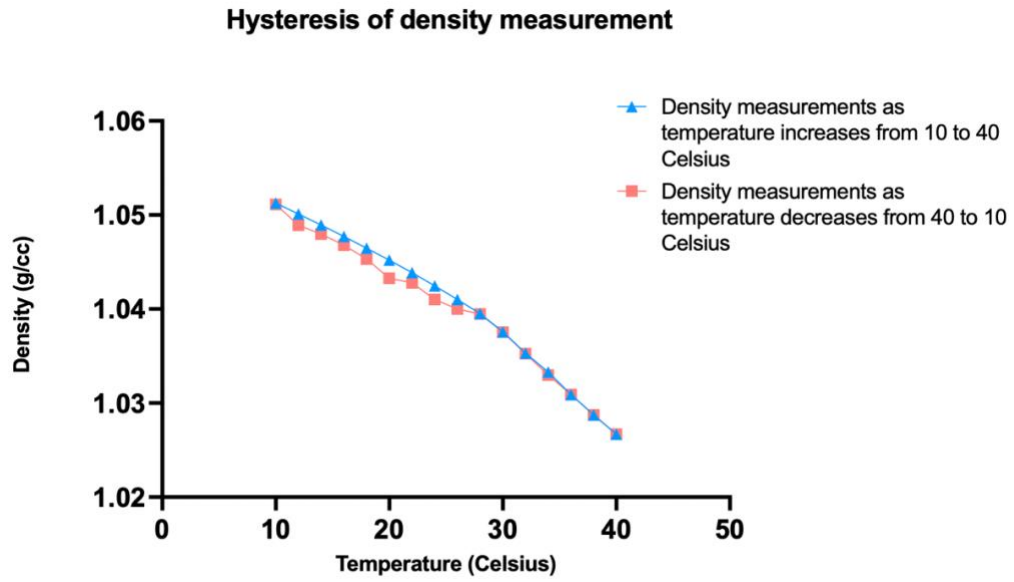

Figure 6: Hysteresis of a solution of 15.04 Brix and 33.09% ABW resulted in discrepancies between measurements taken during temperature increase and decrease. This indicated potential measurement issues, which were found to be changes in settings of the instrument which were subsequently fixed.

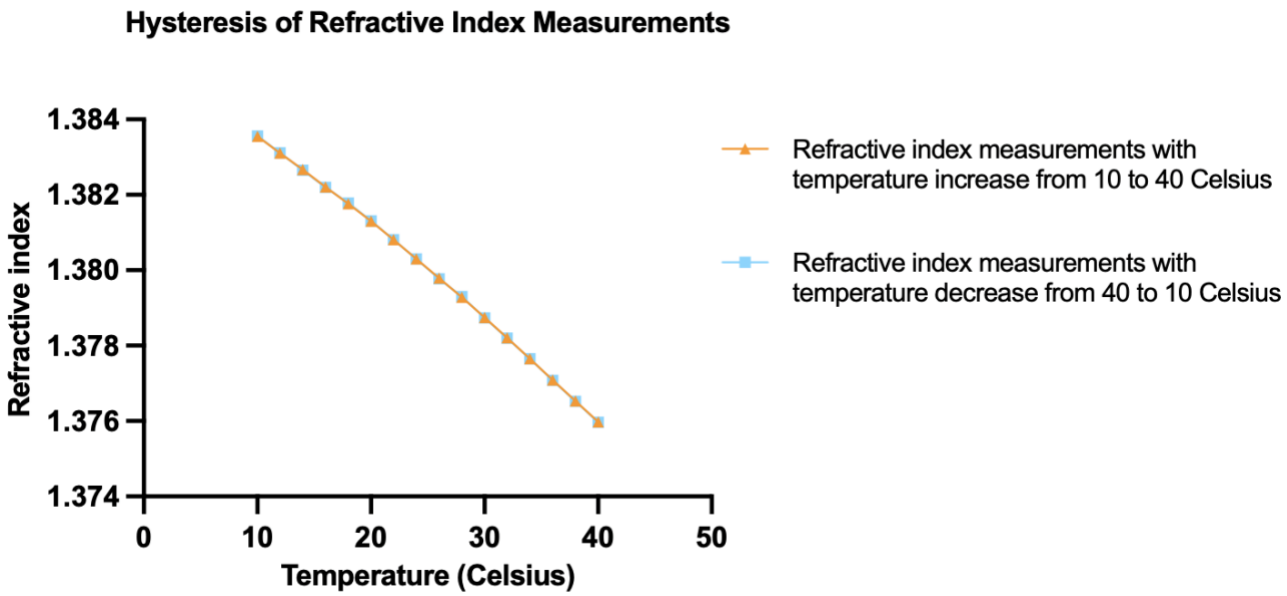

Figure 7: Refractive index hysteresis for the mixture of 15.04 Brix and 33.09% ABW. Identical results were observed during heating and cooling phases, demonstrating good repeatability and consistency of measurements.

Compositions Measured

Out of all the mixtures of ethanol, water, and sucrose that were measured, 98 compositions (**Figure 8**) were identified as highly accurate, showing no interferences of air bubbles in measurements and passing the validation checks. Each of these composition's density and refractive index data was measured over the temperature range of 10°C to 40°C, in 2°C increments. Due to refractometer malfunction and downtime caused by waiting for a replacement, 10 of the compositions had missing refractive index data points. They were excluded from the training dataset of the machine learning model. The exclusion was considered to have minimal impact, as data from closely related compositions were measured and included. Binary systems of sucrose-water and ethanol-water are well-established systems, their density and refractive index measurements are taken from literature (Anton Paar, 2025; Charles, 1965; Mettler Toledo, n.d.-f; Mettler Toledo, n.d.-g; Scott, 1946; Smithsonian Institution, n.d.).

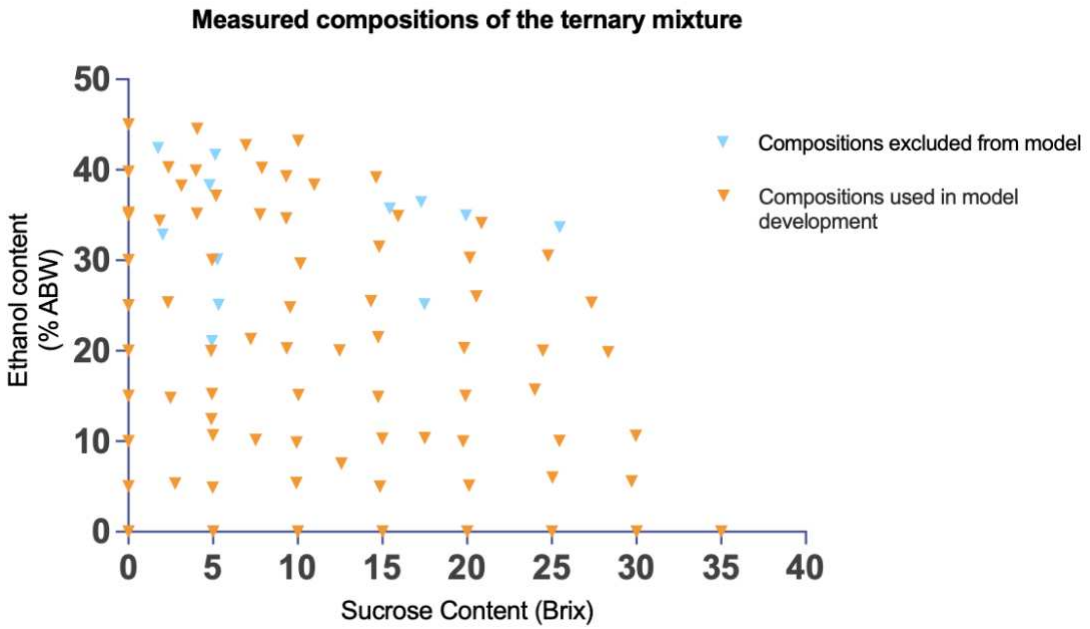

Figure 8: Compositions that were measured for density and refractive index at 10°C to 40°C.

Model development

A .csv file was compiled, containing all the measured values of temperature, density, refractive index, and composition in terms of the ABW% and Brix of the solution. The temperature, density, and refractive index are features, and the ABW% and the Brix of the solution are considered targets. Exploratory data analysis was conducted on the dataset. This included generating summary statistics, consisting of the mean, standard deviation, and range, to understand the distribution of each variable. Histograms were generated to visualize distributions and identify potential outliers (Sarkis, 2023). A correlation matrix was generated to assess the relationship between pairs of variables. This helped confirm known dependencies such as an increase in Brix leads to an increase in density and refractive index at constant temperature. Data integrity checks were also performed to find missing values, duplicates, and outliers. All the computation and modeling were done using Google Collab and the scripts attached in the Appendix section.

Data scaling refers to the transformation of numerical features to a consistent scale. While simpler models such as linear regression would not benefit from scaling the data, methods such as support vector regression and neural networks perform significantly better on scaled input data. Two common scaling techniques are min-max scaling and standardization (Géron, 2019). Min-max scaling is the simplest where the values are shifted and rescaled so they end up in the range of 0 to 1, but is highly sensitive to outliers. In standardization, each feature is transformed to have a mean of 0 and a standard deviation of 1. Standardization was applied using the StandardScaler transformer. To assess the effect of scaling, models were evaluated using both standardized and unscaled datasets.

In the initial model analysis, the Mean Squared Error, Coefficient of Determination, and Mean Absolute Error were calculated for predicted ABW% and predicted Brix in each model. These metrics were then compared to assess each model's performance. Mean squared error (MSE) measures the average of the squared differences between predicted and actual values. A lower MSE indicates better predictive performance. The coefficient of determination (R^2) indicates how well the model explains the variance in the target variable. The R^2 value of 1.0 indicates perfect prediction, while 0 means the model explains none of the variability. Mean absolute error (MAE) represents the average of the absolute differences between predicted and actual values. It gives a straightforward interpretation of error in the same units as the target variables. The next step involved training and evaluating the predictive models.

Linear regression models were applied using the scikit-learn library. The performance was assessed using 5-fold cross-validation, where the dataset is randomly partitioned into five equal-sized folds. For each fold, a linear regression model is trained on four folds and evaluated on the remaining fold. This process is repeated five times, with each fold serving as the test set once.

The results are then averaged to mitigate the effects of variation due to data split. For reproducibility, a fixed random seed was used. Polynomial regression is the transformation of features based on standard linear regression. Models of polynomial regression of degree 2 to 7 were also generated.

A Random Forest regression model consisting of 1000 decision trees was also implemented to predict ABW% and Brix values. Five-fold cross-validation was conducted to evaluate model performance across different data splits. Additionally, feature importance scores were computed to identify the most influential input variables contributing to the predictions of ABW% and Brix.

The next model implemented was support vector regression (SVR) with a radial basis function kernel. This kernel enables the model to capture non-linear relationships by mapping the input data into a higher-dimensional feature space where linear separation becomes feasible. The regularization parameter was initially set to 1, and the tolerance for stopping criteria was left at the default value of 0.1. These hyperparameters were subsequently varied to identify the values that minimized prediction errors.

K-nearest neighbors (KNN) regression was also explored as a predictive model. A range of k values, from 1 to 20, was evaluated to identify the optimal number of neighbors that yielded the best predictions. The model was implemented using scikit-learn's MultiOutputRegressor, enabling simultaneous prediction of both ABW% and Brix. To understand which features had the most influence on the KNN model, the variance of each input variable was examined. Features with higher variance tend to contribute more to the model's predictions, similar to how feature importance is calculated in Random Forest models.

Neural network models were also implemented using scikit-learn's Multi-Layer Perceptron regressor. Six different configurations with different number of hidden layers were tested: a small network with a single hidden layer of 10 neurons, a medium network with two hidden layers of 50 and 25 neurons, a large network with two hidden layers of 100 and 50 neurons, a deep network with three hidden layers of 100, 50, and 25 neurons, an extended configuration with two hidden layers of 50 and 100 neurons, and a minimal configuration with two hidden layers of just 2 and 4 neurons. The default activation function of the Rectified Linear Unit was used across all models.

Based on the performance of the various models, as detailed in the Results and Discussion section, the most effective models were identified and further refined. Improvements were guided by how the accuracy of predictions changed with the model parameters.

Model predictions of distillery samples

In collaboration with Emerson and the local distillery Copper Muse, an inline Coriolis meter and refractometer were installed before the bottling stage of a select few alcoholic beverages. This allowed for real-time data collection directly before bottling. The bottled beverages were also passed through the Mettler Toledo setup to obtain density and refractive index data over the range 10-40°C, similar to the prepared ternary mixture samples. These measurements were inputs to the model to evaluate the accuracy of predictions. Since real alcoholic beverages contain additional additives beyond ethanol, sucrose, and water, this served as a useful test to evaluate whether the three-component model could be generalized to more complex, real-world mixtures.

CHAPTER 3: RESULTS AND DISCUSSION

Compositions Selected for Modeling and Measurement

The composition ranges selected for measurements and modeling was based on the typical compositions of the commonly available alcoholic beverages. High-alcohol, low-sugar beverages such as vodka and gin were considered, which commonly have an alcohol content of around 40% ABV (approximately 35% ABW) and negligible sugar content (Alcohol Help, 2025). The other end of the spectrum included low-alcohol, high-sugar beverages like canned fruit ciders with sugar levels ranging from 10 to 26 Brix and alcohol contents around 5% ABV (approximately 3% ABW). Certain liqueurs such as amaretto and crème liqueurs represent examples of beverages with both high sugar and high alcohol levels, typically around 25 Brix and 30% ABV (approximately 25% ABW). Using these examples, the composition range for preparing the ternary ethanol, sucrose, and water solutions was set to a range of 0 to 45% ABW and 0 to 30 Brix.

After defining the composition range, samples were selected to cover the full range with approximately uniform increments of 5% ABW and 5 Brix. Initially, emphasis was placed on compositions near 40% ABV as it is the most common alcohol content found in popular distilled beverages like vodka, gin, tequila, and whisky. The finalized set of 90 compositions for regression modeling, having passed all consistency checks, is presented in **Figure 9.**

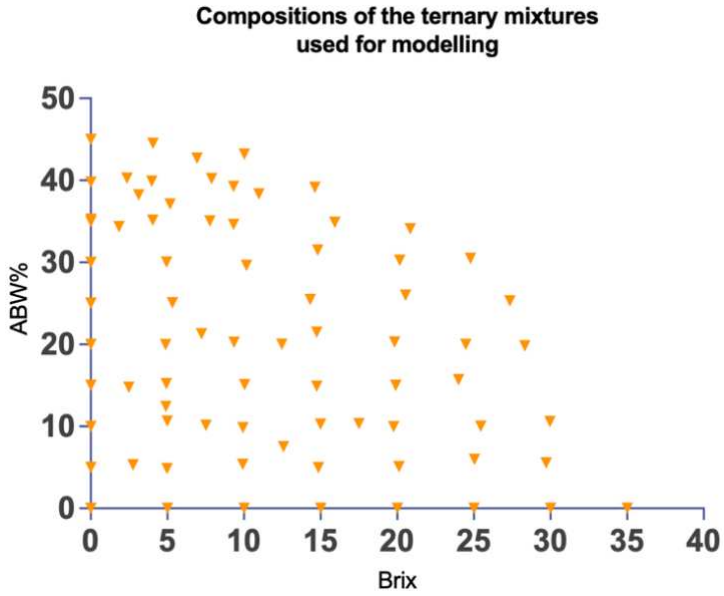

Figure 9: Compositions of ternary mixture used for the regression models.

For each composition, density and refractive index measurements were conducted over a temperature range of 10°C to 40°C in 2°C increments. The subsequent analysis will examine how the physical properties of density and refractive index vary with compositional factors such as ABW% and Brix across temperatures. The industry sponsor for the project requested that the amount of data presented in the thesis be limited. Therefore, the following sections focus on specific representative cases that effectively illustrate the key situations of interest.

Behavior of density and refractive index with varying temperature at constant composition

For a fixed composition, density was observed to decrease with increasing temperature, consistent with established thermodynamic principles. Depending on the composition, the decrease in density exhibited either a linear or non-linear pattern. It was initially hypothesized that a solution with a high Brix and ABW% would show a more pronounced non-linear behavior (Bouchard et al., 2007). However, this was not consistently observed. As illustrated in **Figure 10**,

a solution of 20.84 Brix and 34.11% ABW, which can be considered as a relatively high Brix and ABW% solution, exhibits linear behavior.

Nonlinear behavior in the density–temperature relationship was primarily observed in compositions with either low Brix and high ABW% or low ABW% and high Brix. An illustrative case is shown in **Figure 11**, where a solution containing 5.13 Brix and 41.64 % ABW exhibits clear nonlinearity with increasing temperature. This deviation from linear trends may be attributed to non-ideal interactions between ethanol, water, and sucrose at the compositional boundaries.

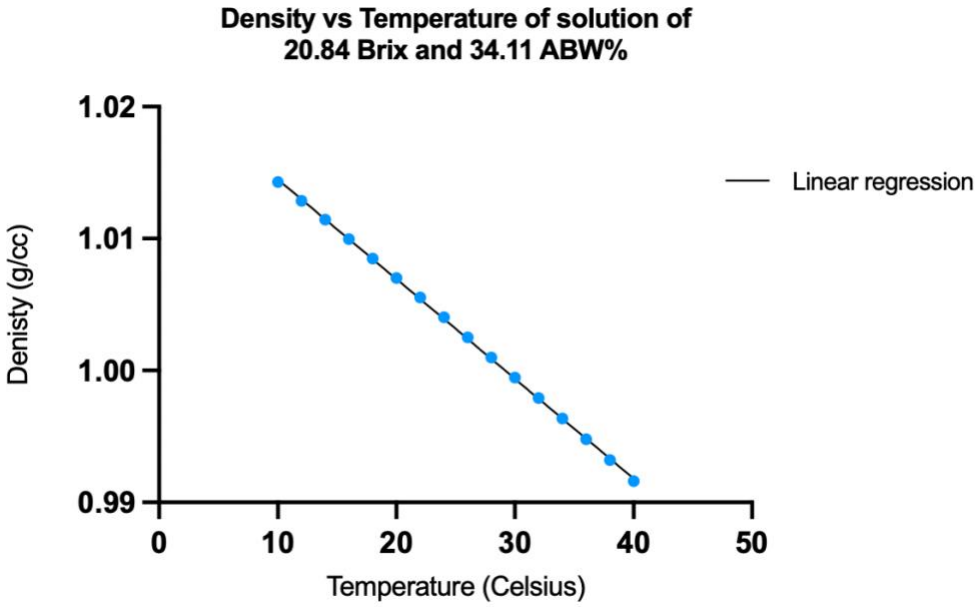

Figure 10: Density versus temperature for a solution with 20.84 Brix and 34.11% ABW exhibiting linear behavior. A simple linear regression model has been used to show the trend.

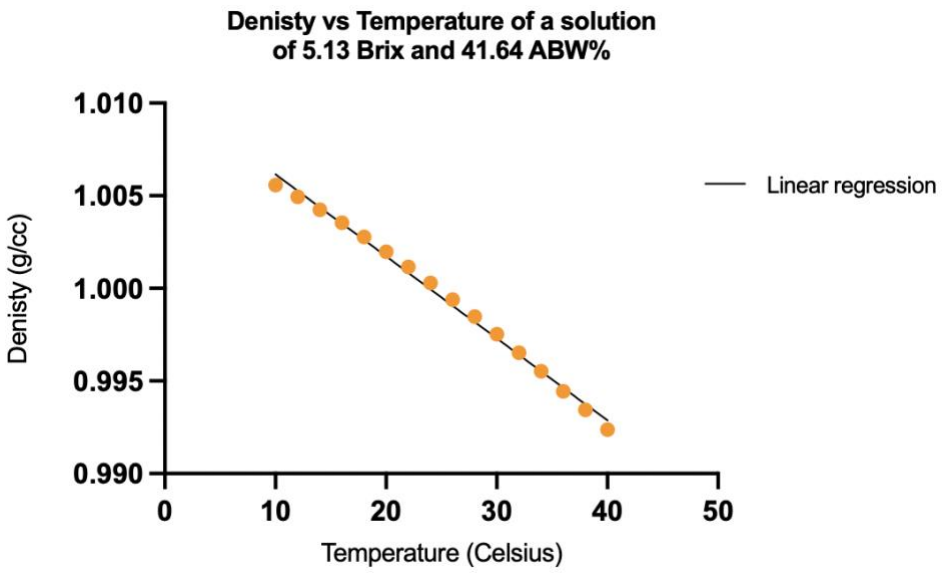

Figure 11: Density versus temperature for a solution with 5.13 Brix and 41.64% ABW exhibiting non-linear behavior. A simple linear regression model has been used to show the trend.

The refractive index decreased with increasing temperature, which is consistent with known behavior of aqueous solutions. Similar to the behavior of density, the refractive index did not show any consistent trend of linearity or non-linearity with changes in temperature across different compositions. In some cases, the behavior followed a highly linear trend, as shown in **Figure 12**. In other cases, it was a non-linear trend, as shown in **Figure 13**. The observed non-linearity can be attributed to compositional boundary effects, as previously discussed in relation of the density–temperature behavior.

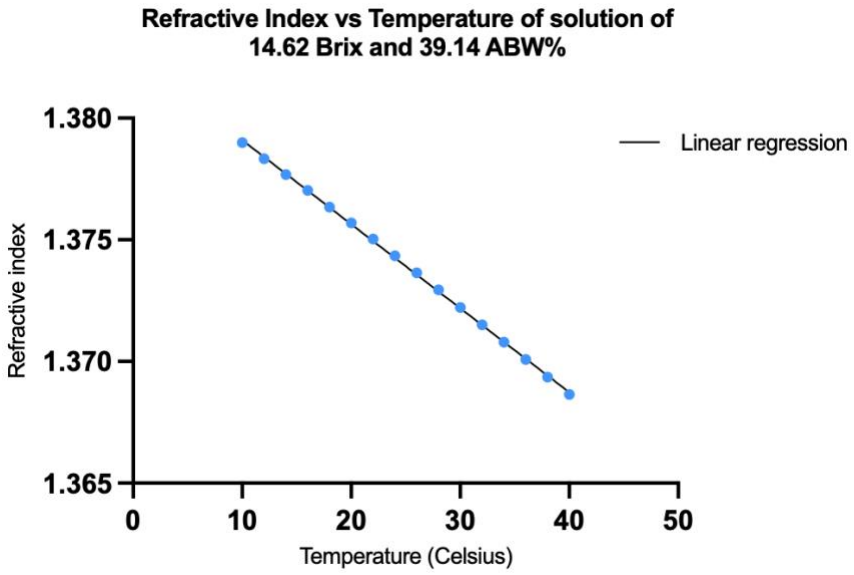

Figure 12: Refractive index versus temperature for a solution with 14.62 Brix and 39.14% ABW exhibiting linear behavior. A simple linear regression model has been used to show the trend.

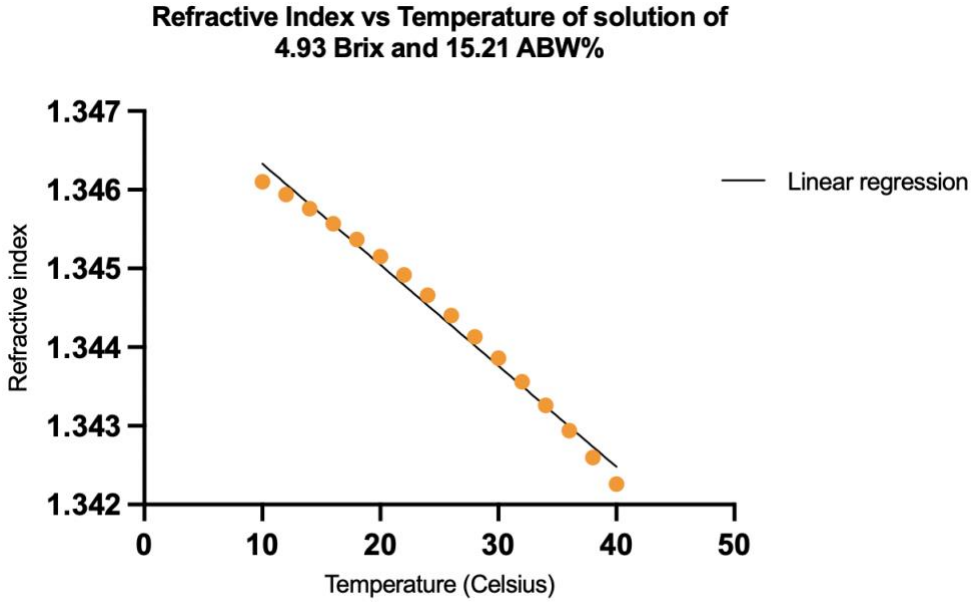

Figure 13: Refractive index versus temperature for a solution with 4.93 Brix and 15.21% ABW exhibiting non-linear behavior. A simple linear regression model has been used to show the trend.

Behavior of density and refractive index with varying Brix and temperature at fixed ABW%

Both density and refractive index exhibited close to a linear increase with rising Brix of the solution when the ABW% and temperature were maintained constant. As previously discussed, it is not practically feasible to prepare solutions with an exact and perfectly constant % ABW or Brix. In **Figure 14**, the behavior of density with an increase in Brix is shown for solutions with approximately 5% ABW at 20°C. While the target was 5% ABW, slight variations occurred during preparation, with values ranging from approximately 4.98% ABW to 5.54% ABW. These deviations may have introduced irregularities in the plotted curve. **Figure 15** illustrates the shift in the trend of density as temperature increases for the same composition.

In **Figure 16**, a similar linear increase was observed in the behavior of refractive index versus Brix at a fixed temperature of 20°C and fixed 5% ABW. **Figure 17** illustrates the shift in the trend of refractive as temperature increases for the same composition.

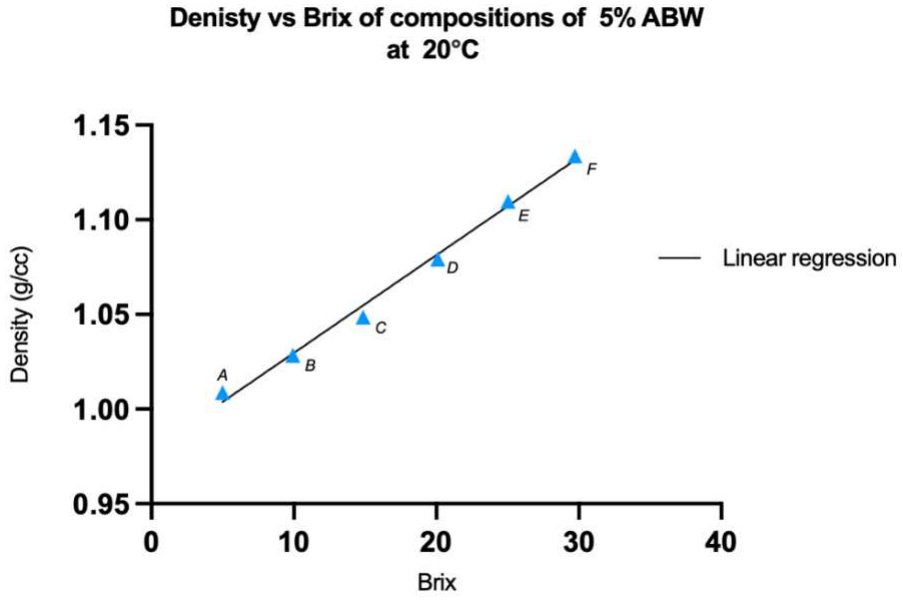

Figure 14: Density versus brix of different compositions of 5% ABW at 20°C. Each letter denotes a different composition. A) 4.98 Brix and 4.85% ABW, B) 9.92 Brix and 5.37% ABW, C) 14.86 Brix and 4.97% ABW, D) 20.11 Brix and 5.08% ABW, E) 25.03 Brix and 5.97% ABW, F) 29.72 Brix and 5.54% ABW.

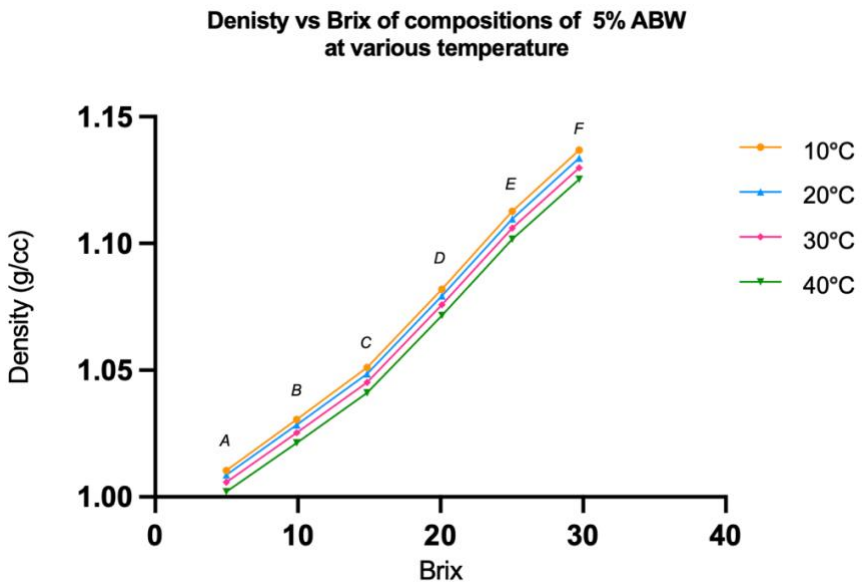

Figure 15: Density versus brix of different compositions of 5% ABW at various temperatures. Each letter denotes a different composition. A) 4.98 Brix and 4.85% ABW, B) 9.92 Brix and 5.37% ABW, C) 14.86 Brix and 4.97% ABW, D) 20.11 Brix and 5.08% ABW, E) 25.03 Brix and 5.97% ABW, F) 29.72 Brix and 5.54% ABW.

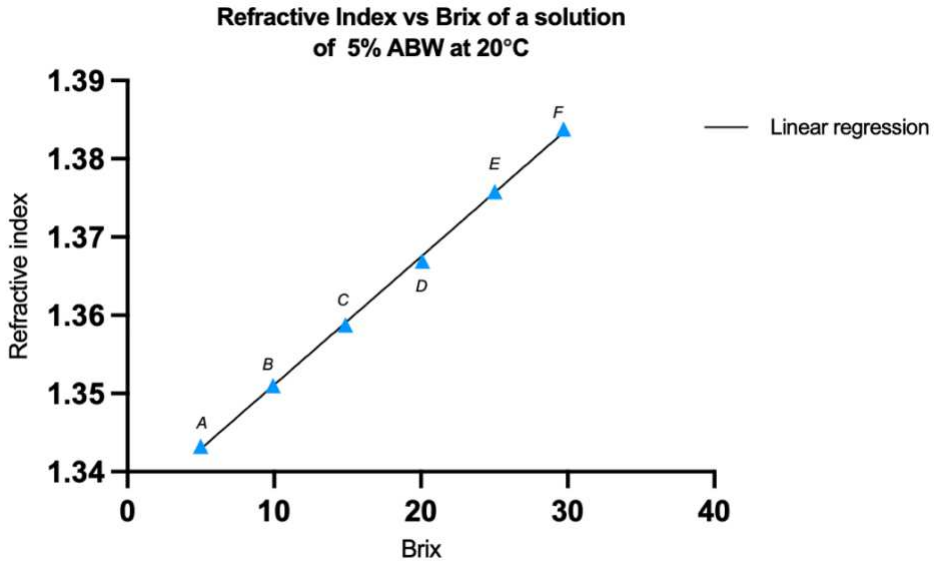

Figure 16: Refractive index versus brix of different compositions of 5% ABW at 20°C. Each letter denotes a different composition. A) 4.98 Brix and 4.85% ABW, B) 9.92 Brix and 5.37% ABW, C) 14.86 Brix and 4.97% ABW, D) 20.11 Brix and 5.08% ABW, E) 25.03 Brix and 5.97% ABW, F) 29.72 Brix and 5.54% ABW.

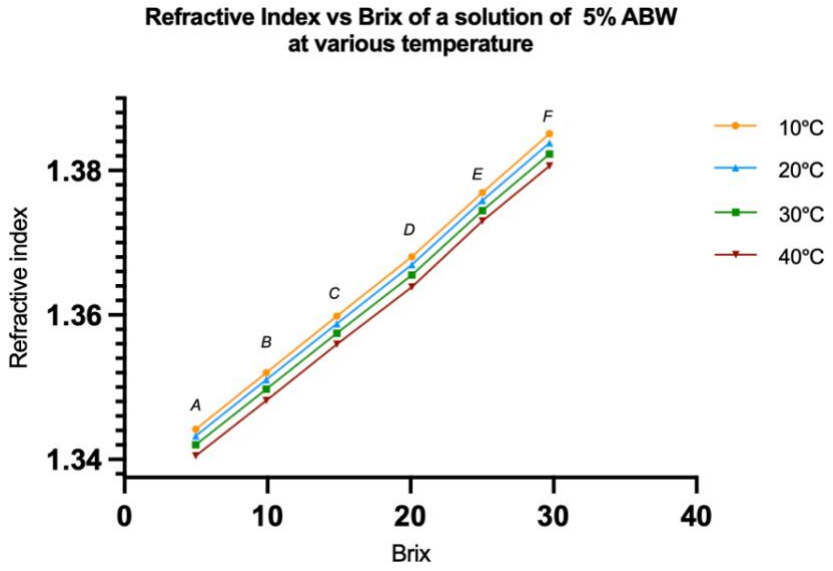

Figure 17: Refractive index versus brix of different compositions of 5% ABW at various temperatures. Each letter denotes a different composition. A) 4.98 Brix and 4.85% ABW, B) 9.92 Brix and 5.37% ABW, C) 14.86 Brix and 4.97% ABW, D) 20.11 Brix and 5.08% ABW, E) 25.03 Brix and 5.97% ABW, F) 29.72 Brix and 5.54% ABW.

As shown earlier, the nearly linear increase in density with an increase in Brix at a fixed ABW% and temperature can be attributed to the behavior of sucrose in aqueous solutions. **Figure 18** illustrates this relationship in a binary mixture of sucrose and water at a constant temperature, where density increases linearly with sucrose concentration (Mettler Toledo, n.d.-f). This linear trend occurs because sucrose dissolves readily in water, resulting in minimal volume contraction effects. Similarly, the same explanation could be used for the ternary system investigated here. When ABW% and temperature are held constant, increasing Brix corresponds to adding sucrose to the solution without any change in the amount of ethanol. Since sucrose is soluble in water and does not significantly alter the ternary solution volume in a nonlinear fashion, the resulting density increase remains close to linear.

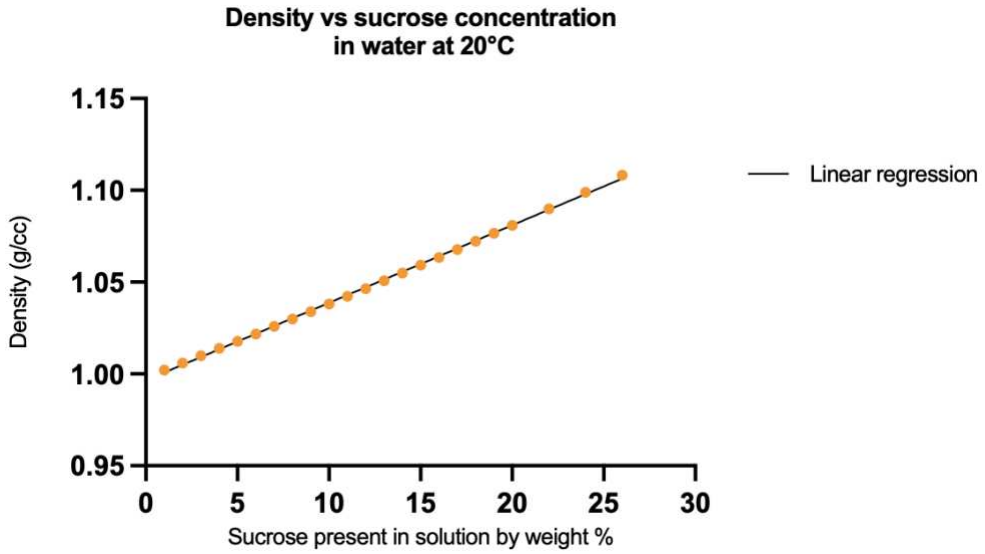

Figure 18: Density versus sucrose concentration in a binary solution of water and sucrose at 20°C. Data points were obtained from Mettler Toledo, n.d.-f.

Behavior of density and refractive index with varying ABW% and temperature at fixed Brix

Both density and refractive index exhibited a nonlinear decrease with increasing ABW% when Brix and temperature were held constant, which is consistent with established thermodynamic principles. This behavior is illustrated in **Figure 19**, which shows the trend at a fixed 5 Brix and 30°C. The data for this figure were sourced from Resa et al. (2005).

A similar trend is observed in **Figure 20**, where the effect of varying ABW% on density across different temperatures is shown at a fixed composition of 10 Brix. Although 10 Brix was the intended value for all solutions, slight deviations occurred during preparation, with measured Brix values ranging from approximately 9.31 to 10.25 Brix. In **Figure 19** and **Figure 20**, a third-degree polynomial regression line was fitted to the data points to highlight the nonlinearity of the observed trends.

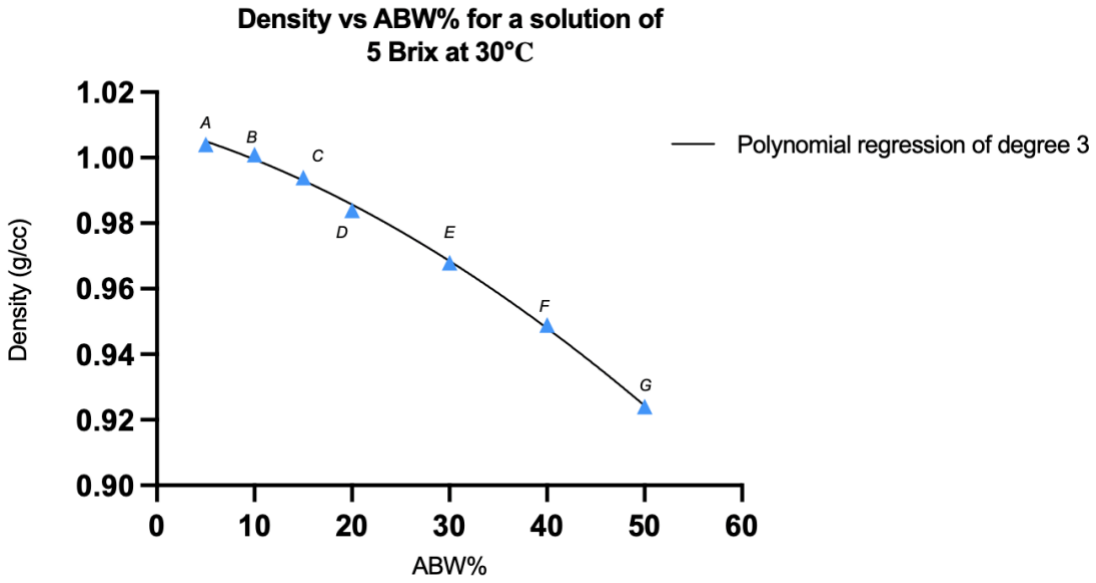

Figure 19: Density versus ABW% for a solution of 5 Brix at 30°C. Each letter denotes a different composition. A) 5% ABW and 4.95 Brix, B) 10% ABW and 5.66 Brix, C) 15% ABW and 5.66 Brix, D) 20% ABW and 5.66 Brix, E) 30% ABW and 5 Brix, F) 40% ABW and 5.21 Brix, G) 50% ABW and 5 Brix. (Resa et al., 2005)

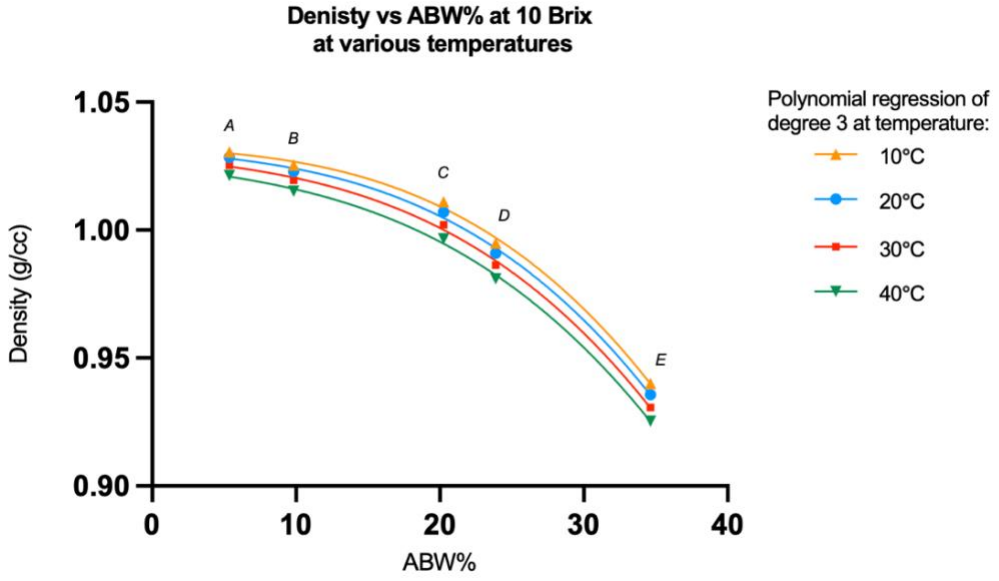

Figure 20: Density vs ABW% of solution of 10 Brix at various temperatures. Each letter denotes a different composition. A) 9.92 Brix and 5.37% ABW, B) 9.92 Brix and 9.835% ABW, C) 9.35 Brix and 20.26% ABW, D) 10.25 Brix and 23.88% ABW, E) 9.31 Brix and 34.63% ABW

The non-linear decrease in density with increasing ABW% at constant temperature can be attributed to the non-ideal mixing behavior of ethanol and water. **Figure 21** illustrates this relationship in a binary mixture of ethanol and water at 20°C, where density decreases in a non-linear trend as ABW% increases (Mettler Toledo, n.d.-g). This behavior arises from volume contraction that occurs when ethanol is mixed with water because of hydrogen bonding between molecules. For example, when 50 mL of ethanol is mixed with 50 mL of water, the resulting total volume is approximately 97 mL, instead of 100 mL. This deviation from ideal behavior significantly affects the physical properties of the solution. Similarly, In the context of the ternary system, when Brix is constant and ABW% is increased at a fixed temperature, effectively there is ethanol being added to the system while keeping the amount of sucrose the same. Hence the observed non-linear trends can be attributed to the non-ideal mixing effects.

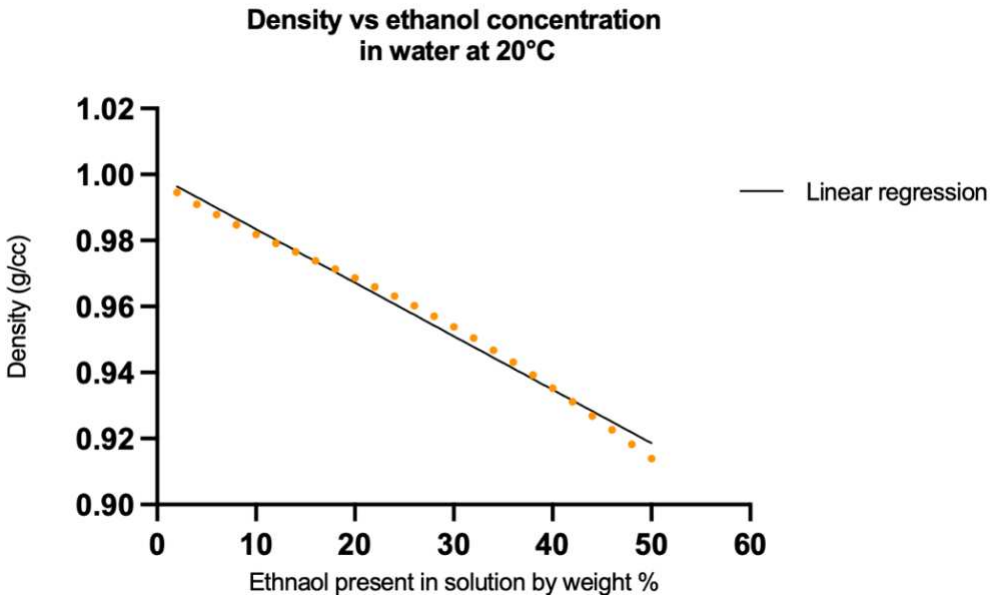

Figure 21: Density versus ethanol concentration in binary mixture of ethanol and water at 20°C. Data points were obtained from Mettler Toledo, n.d.-g.

Modelling for prediction of composition of an unknown solution

For the predictive modeling, a comprehensive dataset was required. As described previously, measurements of density and refractive index were collected for each composition across a temperature range from 10°C to 40°C, in 2°C increments. This experimental dataset comprised 90 unique compositions. Each data entry can be represented as a row in the .csv file in the form of [temperature, density, refractive index, ABW%, Brix]. For each unique composition defined by a fixed ABW% and Brix, there should ideally be 16 data points, corresponding to measurements at each 2°C increment between 10°C and 40°C. Therefore, for 90 compositions, a complete dataset would contain 1,440 rows. However, the compositions also included binary mixtures obtained from literature sources. These literature datasets typically reported measurements at less uniform temperature intervals such as 5°C or 10°C and often contained missing values for either density or refractive index. As a result, after combining the experimental and literature data, the final dataset comprised 1210 valid rows in the specified format. Hence the model can be considered to have 1210 data points.

Given the inputs as temperature, density, and refractive index of an unknown solution, the goal was to predict the two outputs or targets which were ABW% and Brix. Rather than employing a single regression model to predict both variables, it was observed that different regression types performed better for each target. For example, as will be shown in subsequent sections, polynomial regression demonstrated strong performance in predicting Brix but was less effective for ABW% predictions. Conversely, random forest regression yielded accurate ABW% predictions but was not optimal for Brix predictions. Hence, the prediction task was divided into two separate regression models, one for Brix and another for ABW%. For each target, a range of regression techniques was evaluated, including linear regression, polynomial regression, random forest regression, k-nearest neighbors, SVM regression, and neural network.

The predictive performance of each of these regression types was then compared, and the best-performing model type for each target variable was selected and further optimized to develop the final predictive models. As will be shown subsequently, it was found that a polynomial regression model of degree 8 provided the best predictions for Brix without signs of overfitting. For ABW%, the random forest regression model emerged as the most accurate predictor. These two final models each take the temperature, density, and refractive index of the unknown solution as inputs. The polynomial regression model outputs the predicted Brix, while the random forest model predicts the ABW%.

For all modeling tasks, the dataset was initially split into training and testing sets in an 80:20 ratio. To ensure a robust evaluation of the model and to minimize the influence of a potentially unrepresentative split of the data, k-fold cross-validation was applied. Since the dataset was not very large, k-fold cross-validation was considered an appropriate choice. For example, in 5-fold cross-validation, the training data is divided into five equal parts or folds. The model is trained on four of these folds and validated on the remaining one. This process is repeated five times, with each fold serving as the validation set once. The prediction evaluation metrics such as mean squared error (MSE) are evaluated for each fold, and the metric's variation across folds is analyzed. A low variation in the metrics across folds suggests that the model's predictive performance is consistent and the data split is reliable. Based on this, the best-performing fold or the average across folds can be used to select optimal model settings. Additionally, a fixed random seed was assigned to ensure that the data-splitting process and model outputs were reproducible.

The following sections detail the selection and optimization of regression models for the separate prediction of Brix and ABW%.

Model for prediction of Brix

Performance of exploratory predictive models for Brix

The different regression models were first assessed on their ability to predict Brix based on the input variables of temperature, density, and refractive index. In **Table 4**, the performance of models was evaluated using the following metrics: Mean Squared Error (MSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE). Lower values of the error metrics and higher values of R^2 indicate better predictive performance of the model.

Table 4: Evaluation metrics for accuracy of Brix prediction of a model.

Model	Unscaled Data			Scaled Data		
	MSE	**R^2**	**MAE**	**MSE**	**R^2**	**MAE**
Linear regression	3.5039	0.9514	1.5230	3.5039	0.9514	1.5230
Polynomial Regression Degree 2	1.9323	0.9732	1.0432	1.9323	0.9732	1.0432
Polynomial Regression Degree 3	1.1906	0.9835	0.8322	1.1906	0.9835	0.8322
Polynomial Regression Degree 4	0.9573	0.9867	0.7081	0.8887	0.9877	0.7007

Polynomial Regression Degree 5	1.0327	0.9857	0.7424	0.6047	0.9916	0.5814
Polynomial Regression Degree 6	1.1427	0.9841	0.7962	0.3372	0.9953	0.4328
Random forest	1.6876	0.9806	1.1475	1.6876	0.9806	1.1475
SVM regression	78.4892	0.7748	7.9653	4.2800	0.9253	1.7019
K Nearest Neighbors	36.4530	0.1241	6.4781	13.7163	0.7470	3.0524
Neural network (10)	121.8450	0.0495	9.2722	15.3000	0.8672	3.0845
Neural network (50, 25)	115.8957	0.0016	9.1757	8.994152	0.9194	2.4376
Neural network (100, 50)	150.9866	0.3921	10.0558	7.1159	0.9491	2.1630
Neural Network (100, 50, 25)	115.0271	0.0006	9.1157	7.4633	0.9464	2.2083

Neural Network (50, 100)	116.2734	0.0187	9.1436	8.6956	0.9209	2.3567
Neural Network (2, 4)	145.1868	0.3441	9.9203	13.6117	0.8983	2.9541

Regression models, such as neural networks and support vector machines (SVMs), have demonstrated improved performance when trained on scaled data, which aligns with expectations from existing literature (Géron, 2019). Among the models tested, polynomial regression of degree 6 with scaled data showed the best overall performance in predicting Brix by having the lowest MSE and MAE, closely followed by the random forest model. Given its strong predictive capability, polynomial regression was selected as the primary model for Brix estimation. It was also observed that increasing the polynomial degree led to improved accuracy. Finding the most optimal degree for the polynomial was the next step.

Polynomial regression for Brix prediction

Polynomial regression models of degree 2 to 15 were evaluated on both scaled and unscaled data. For each model, the metrics of mean squared error (MSE), coefficient of determination (R^2), and mean absolute error (MAE) were compared. In addition to evaluating performance on the training and test sets, 5-fold cross-validation was performed, where the training set was split into five folds: the model was trained on four folds and validated on the fifth, repeating this process five times. The evaluation metrics were compared across the training, test, and cross-validation results to identify potential signs of overfitting. A significant discrepancy between training and test or cross-validation metrics, such as a high R^2 on the training set and a

noticeably lower R^2 on the test set suggests that the model may be overfitting by memorizing the data rather than generalization. Performance metrics should be consistent across all sets, indicating good generalization. The impact of data scaling on model performance was examined as shown in **Figure 22**, **Figure 23**, and **Figure 24**.

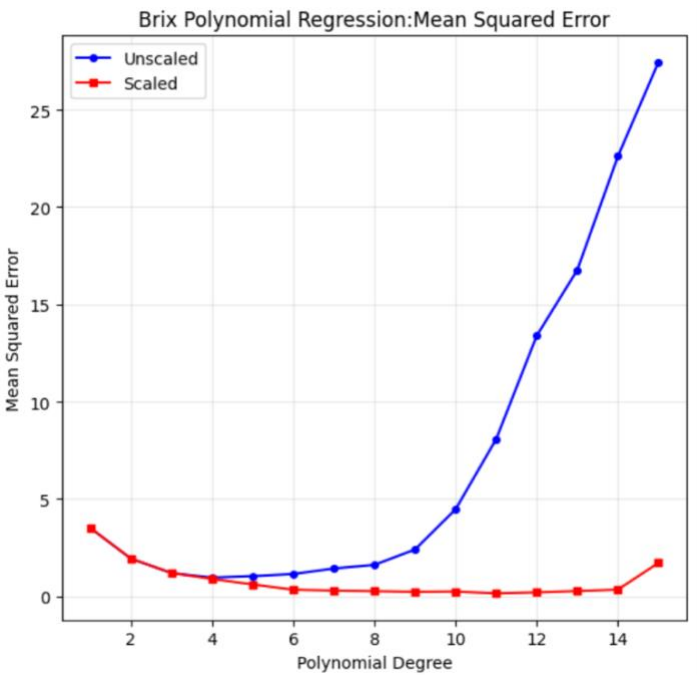

Figure 22: Mean Squared Error (MSE) of polynomial regression models for Brix prediction. It illustrates the impact of using scaled versus unscaled input data.

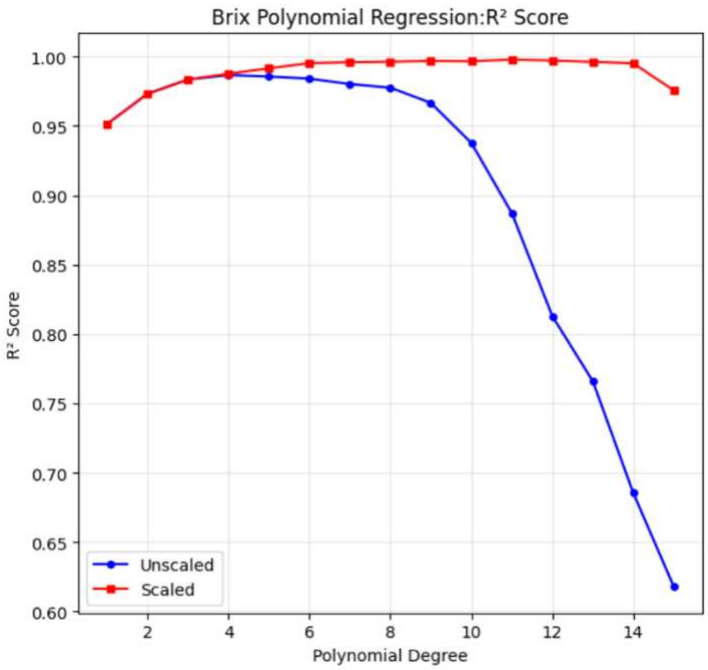

Figure 23: Coefficient of Determination (R^2) of polynomial regression models for Brix prediction. It illustrates the impact of using scaled versus unscaled input data.

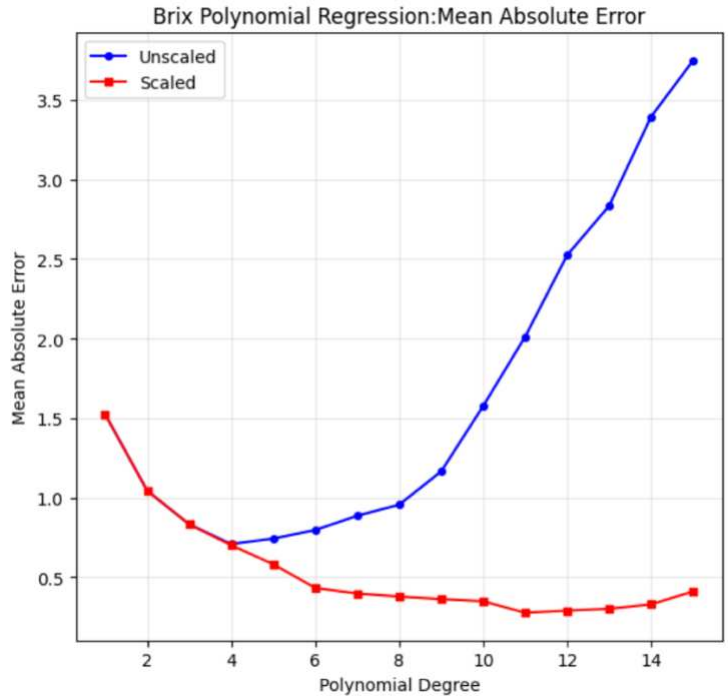

Figure 24: Mean Absolute Error (MAE) of polynomial regression models for Brix prediction. It illustrates the impact of using scaled versus unscaled input data.

Since scaled data yielded better performance, further analysis was conducted on the metrics of models trained with scaled data, comparing results across the training, test, and 5-fold cross-validation sets, as shown in **Figure 25** and **Figure 26**. Since 5-fold cross-validation produces five sets of metrics, one for each fold, the validation performance shown in the respective figures represents the mean of these five values.

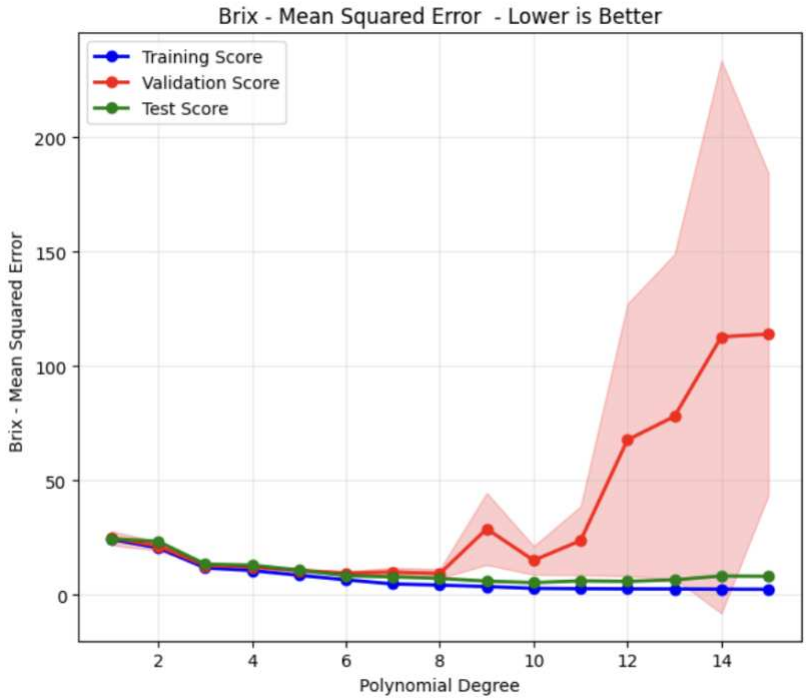

Figure 25: Mean Squared Error (MSE) across varying polynomial degrees for training, testing, and 5-fold cross-validation sets in polynomial regression using scaled data.

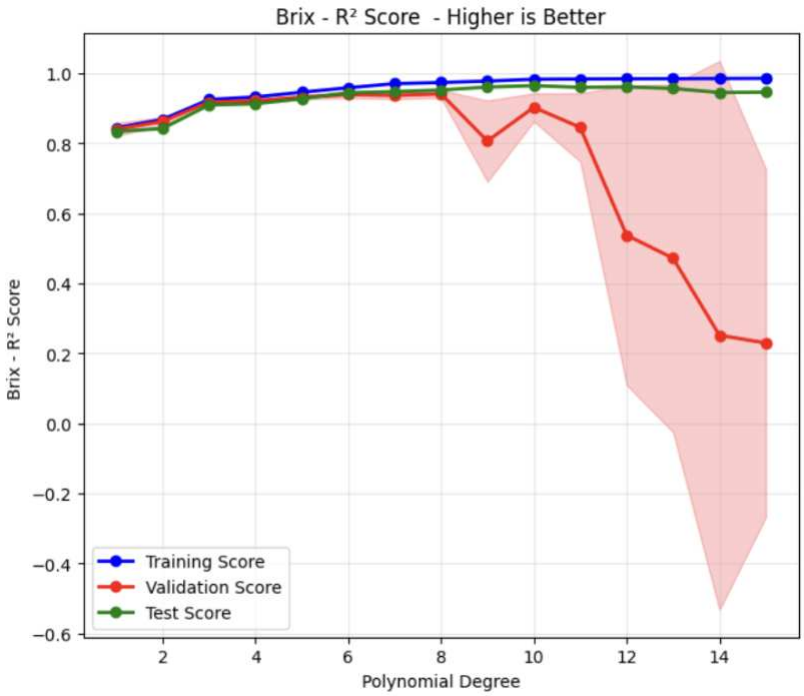

Figure 26: Coefficient of Determination (R^2) across varying polynomial degrees for training, testing, and 5-fold cross-validation sets in polynomial regression using scaled data.

As shown in **Figure 25** and **Figure 26**, increasing the polynomial degree beyond 8 leads to a noticeable divergence in metric values between the training and test set, and cross-validation set. Higher-degree polynomial models show very low MSE and R^2 values approaching 1, indicating overfitting. The model with degree 8 shows a good balance between training and validation performance, with consistent metrics across all sets, suggesting it may be the optimal degree for the model. Hence, further analysis was conducted on the polynomial of degree 8.

The relationship between actual and predicted values was examined for both the training and testing datasets of the polynomial of degree 8. In **Figure 27** and **Figure 28**, in the actual vs. predicted plots, the diagonal line (x = y) represents perfect predictions. Points deviating from the diagonal indicate prediction errors. Additionally, residual plots were generated to assess model performance (**Figure 29** and **Figure 30**). Residuals are the difference between the actual and predicted values.

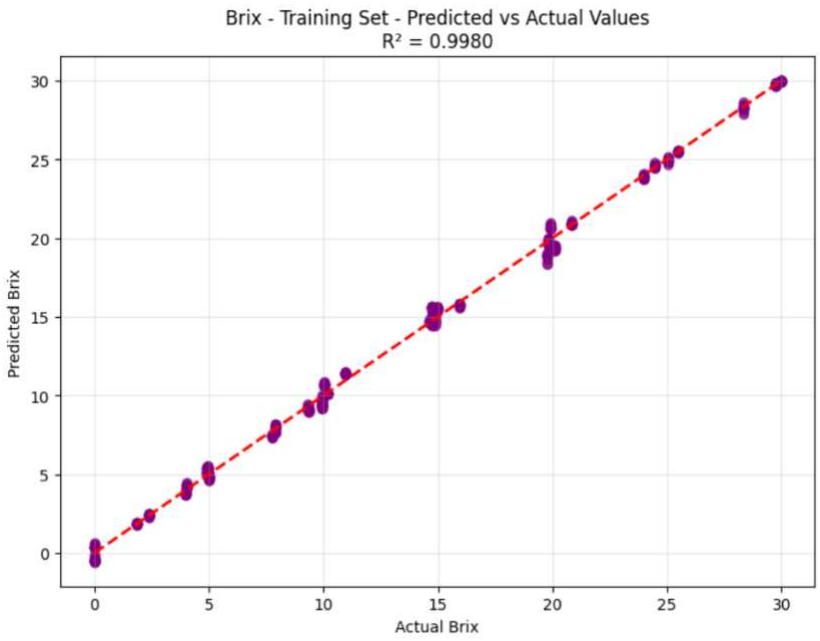

Figure 27: Actual vs. predicted Brix values for the training dataset using the degree 8 polynomial regression model.

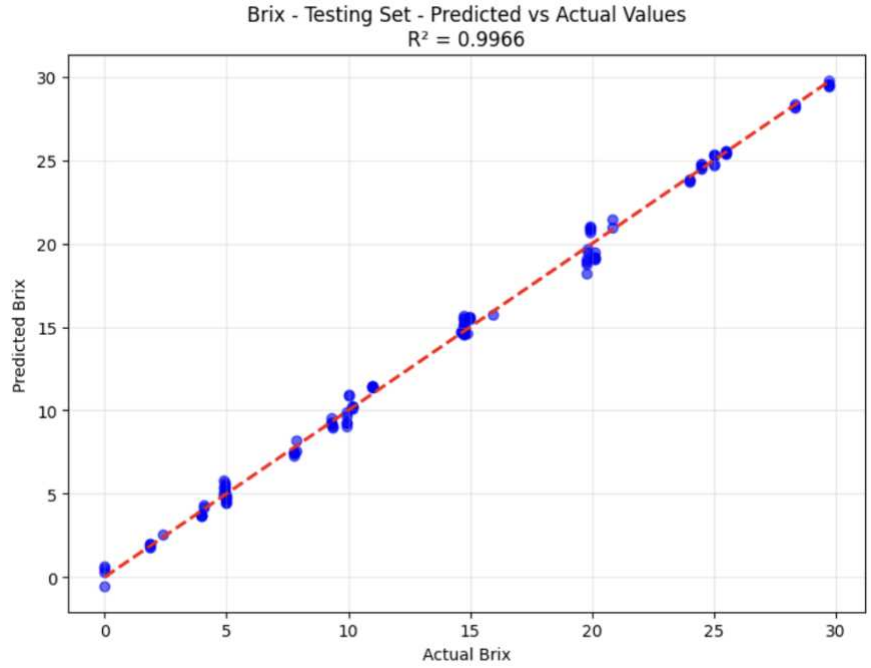

Figure 28: Actual vs. predicted Brix values for the testing dataset using the degree 8 polynomial regression model.

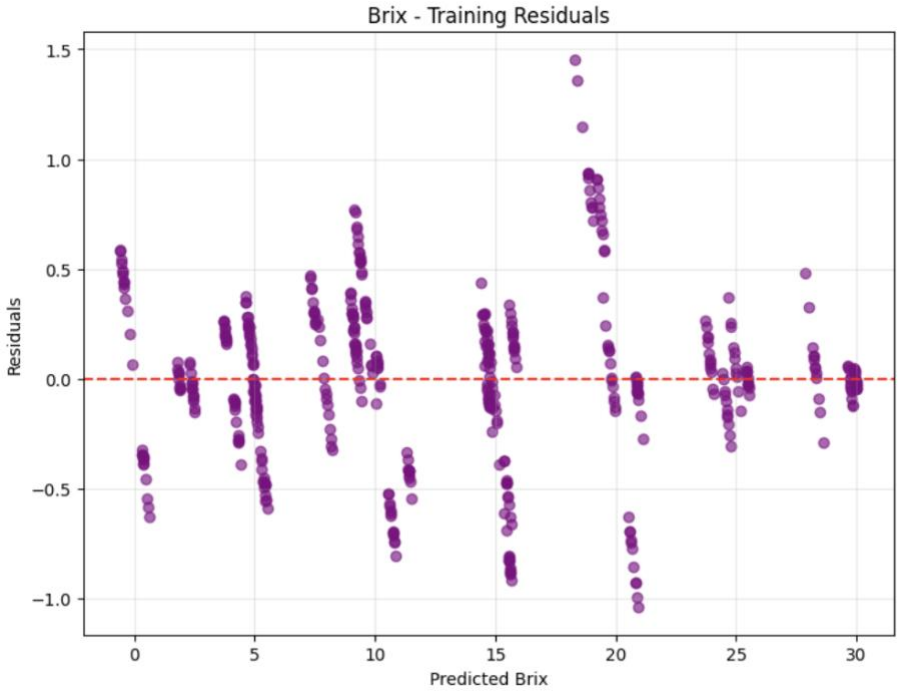

Figure 29: Residuals of Brix prediction on the training dataset using the degree 8 polynomial regression model.

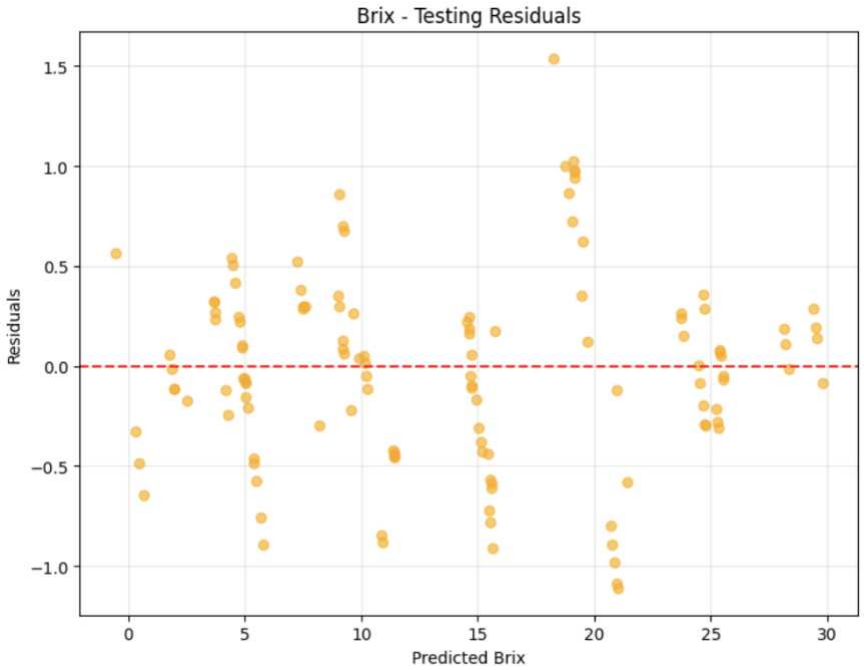

Figure 30: Residuals of Brix prediction on the testing dataset using the degree 8 polynomial regression model.

The residual analysis for Brix prediction showed that the predicted values deviate from the actual values with a maximum residual of approximately +1.5 Brix and a minimum of -1.1 Brix, indicating relatively low error. Furthermore, the actual vs. predicted plot shows that most points lie close to the diagonal line. The Root Mean Squared Error (RMSE) was 0.3813 Brix for the training set and 0.4832 Brix for the testing set, reflecting good generalization performance, given the limited size of the dataset. A 10-fold cross-validation process with R^2 scores ranging from 0.9253 to 0.9623, with a mean of 0.9377 and a standard deviation of 0.0161, indicated that the model performs consistently and is stable across different data splits. Therefore, the polynomial regression model of degree 8 with scaled input data was selected as the model for Brix prediction. Although the improvement in prediction was minimal, the degree 8 model was selected over a degree 6 model, given the negligible difference in computation time. Feature analysis also revealed that the most influential predictors in the model were engineered features such as (density) ^{5}, (density × refractive index) ^{2}, and (density × refractive index) ^{6}, among others. In contrast, temperature showed very low importance.

Prediction of ABW%

Performance of exploratory predictive models for ABW%

The models were first assessed on their ability to predict ABW% based on the input variables of temperature, density, and refractive index. In **Table 5,** the performance of models was evaluated using the following metrics: mean squared error (MSE), coefficient of determination (R^2), and mean absolute error (MAE). Lower values of the error metrics and higher values of R^2 indicate better predictive performance of the model.

Table 5: Evaluation metrics for accuracy of ABW% prediction of a model.

Model	Unscaled			Scaled		
	MSE	R^2	MAE	MSE	R^2	MAE
Linear Regression	24.3893	0.8400	4.2678	24.3893	0.8400	4.2678
Polynomial Degree 2	21.5665	0.8583	3.9372	21.5665	0.8583	3.9372
Polynomial Degree 3	12.773	0.9163	2.9348	12.773	0.9163	2.9348
Polynomial Degree 4	13.0384	0.9146	2.9061	12.1084	0.9207	2.7543
Polynomial Degree 5	13.4597	0.9118	2.9647	10.3915	0.9318	2.4923
Polynomial Degree 6	13.6923	0.9103	2.9996	9.5223	0.9441	2.2746
Random Forest	2.0401	0.9717	1.3653	2.0314	0.9719	1.3619
SVM Regression	78.4892	0.0723	7.2203	21.7361	0.8520	3.6748
K Nearest Neighbors	29.8266	0.1241	4.6410	21.6504	0.3642	3.8428
Neural network (10)	121.8450	0.0495	9.272277	15.3001	0.8853	3.0845

Neural network (50, 25)	115.8957	0.0016	9.1757	8.9941	0.9345	2.4376
Neural network (100,50)	150.9866	0.0021	10.0558	8.5338	0.9362	2.1630
Neural Network (100, 50, 25)	110.1879	0.0006	9.1157	9.0292	0.9318	2.2083
Neural Network (50, 100)	112.5129	0.0187	9.1436	10.6480	0.9209	2.3567
Neural Network (2, 4)	147.1248	0.3441	9.920368	13.6117	0.8983	2.9541

Similar to the Brix analysis, ABW% prediction models such as neural networks performed better when trained on scaled data across most algorithms. The Random Forest model demonstrated the best overall performance, with polynomial regression following behind, while the remaining models showed considerably lower predictive accuracy. Random Forest achieved the lowest values for MSE and MAE, along with an R^2 value closest to 1, indicating stronger predictive capability. The performance of the scaled and unscaled Random Forest models was nearly identical. This observation is consistent with existing literature, as Random Forest

predictions rely on tree-based decisions rather than feature scaling. Optimizing the model with unscaled data was the next step.

Random Forest model optimization for ABW% prediction

The base Random Forest model, configured with 1000 decision trees, achieved a baseline performance of MAE = 1.3619 and R^2 = 0.9719. Feature engineering was first employed in the input space. This created additional features, such as polynomial terms (e.g., temperature²), feature ratios (e.g., density/refractive index), and interaction terms (e.g., density × temperature), all of which aimed to capture nonlinear and combined effects among variables. Then hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. To identify an optimized configuration, 2160 combinations were tested across the parameters of "n_estimators", "max_depth", "min_samples_split", "min_samples_leaf", and "max_features". Finally, ensemble methods were applied, which included simple voting (np.mean), weighted voting (performance-based weight assignment), and stacking ensemble (StackingRegressor). This resulted in a final model performance with an MAE of 1.0575 and an R^2 of 0.9810. The improvement over the baseline is relatively modest, which indicates the model reaching the upper limits of predictive performance given the limited size of the dataset. The relationship between actual and predicted values was examined for both the training and testing datasets. In **Figure 31** and **Figure 32**, in the actual vs. predicted plots, the diagonal line (x = y) represents perfect predictions. Additionally, residual plots were generated to assess model performance (**Figure 33** and **Figure 34**).

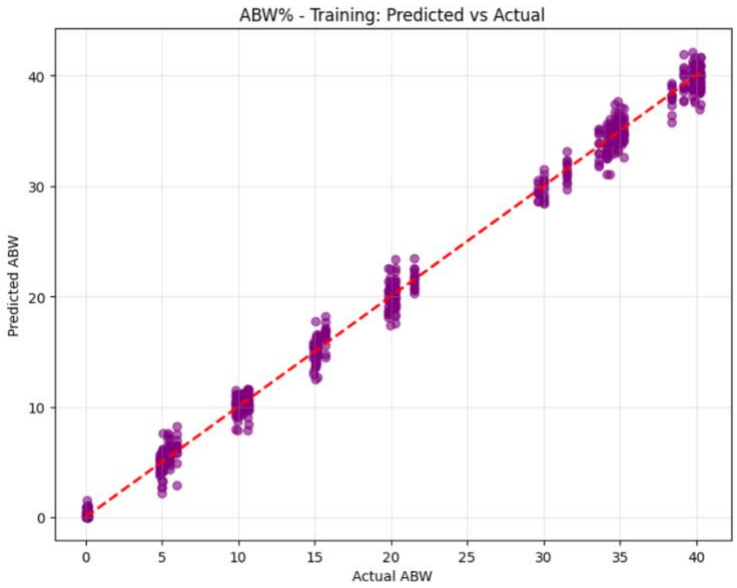

Figure 31: Actual vs. predicted ABW% values for the training dataset using the Random Forest regression model.

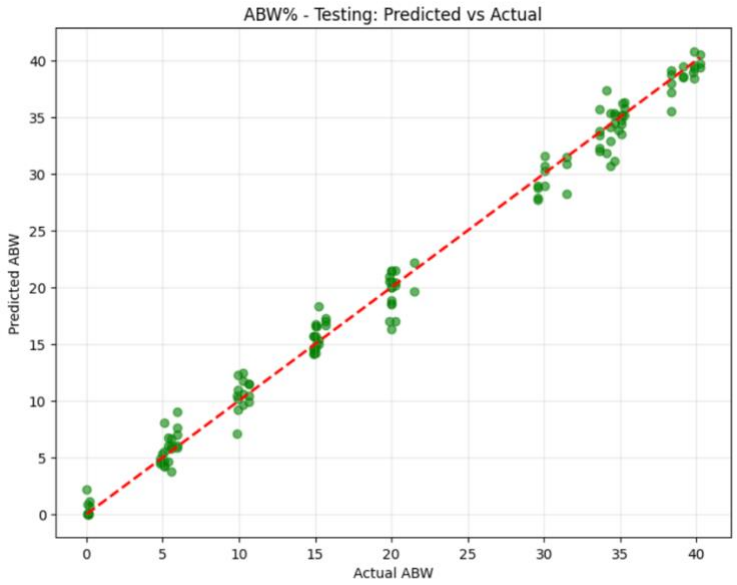

Figure 32: Actual vs. predicted ABW% values for the testing dataset using the Random Forest regression model

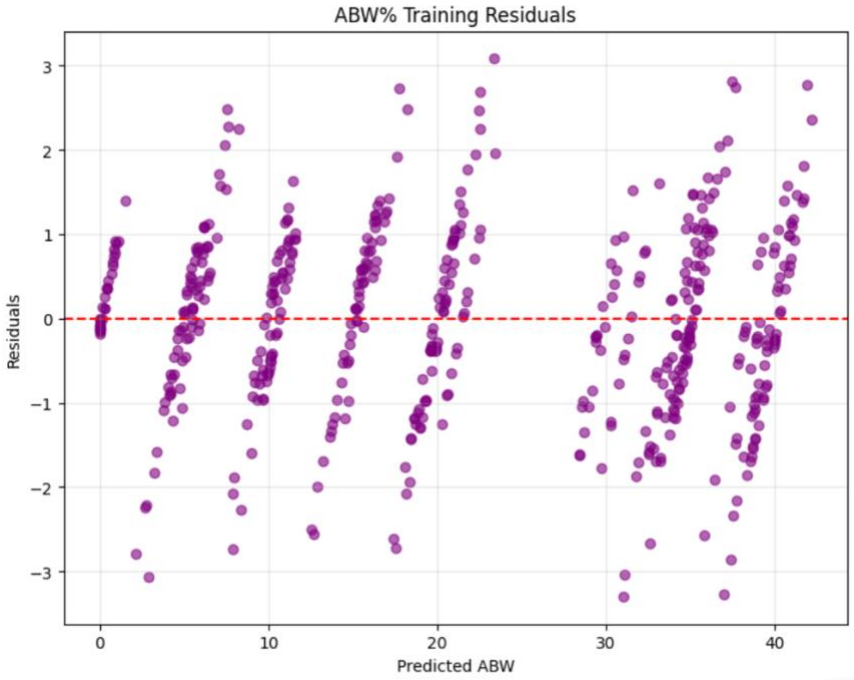

Figure 33: Residuals of ABW% prediction on the training dataset using the Random Forest regression model.

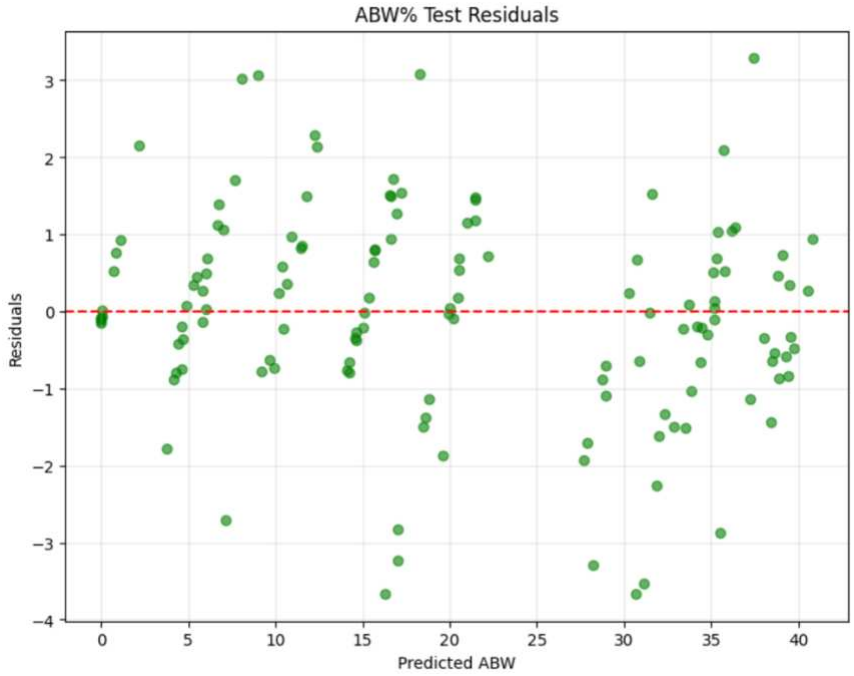

Figure 34: Residuals of ABW% prediction on the testing dataset using the Random Forest regression model.

The residual analysis for ABW% prediction indicated that approximately 63.7% of the predicted values fell within ± 1 ABW% of the actual values. Additionally, 87.3% of the predicted values were within ± 2 ABW% of the actual values, and 95.3% of the predicted values were within ±3 ABW% of the actual values. These results suggest that predicting ABW% is considerably more challenging than predicting Brix. Despite model optimization, the predictions are still not within the ± 0.3 ABW% prediction accuracy required for industrial applications. This limitation can be attributed to the nature of the dataset, which spans a broad range from 0 to 40 ABW%. Concentrating on a smaller ABW% range with denser data coverage would likely yield significantly better predictive performance. The model generally predicts ABW% within ± 2% percentage points, as will be further demonstrated in the subsequent analyses.

Similar to the Brix prediction model, feature analysis of the Random Forest model showed that temperature as a feature had minimal influence on the model's predictive performance. The most important features were the ratio between density and refractive index, and the square of refractive index.

Prediction of unknown ternary solutions prepared in the laboratory

Measured data from solutions not included in the training or testing sets were used to evaluate the predictive performance of the models (**Table 6**). The predictions were consistent with the previous model behavior and residual analysis. Notably, the models for both ABW% and Brix exhibited greater difficulty when the unknown solution contained 0% ABW or 0 Brix respectively.

Table 6: Model predictions for solutions not included in the training or testing datasets.

Model Inputs	Random forest regression model for ABW% prediction			Degree 8 polynomial regression model for Brix prediction		
	Actual ABW%	Predicted ABW%	Absolute Error in ABW%	Actual Brix	Predicted Brix	Absolute Error in Brix
20°C 0.99099 g/cm^3 1.35331 nD	23.88	21.70	2.18	5.02	5.61	0.59
10°C 0.99488 g/cm^3 1.35494 nD	23.88	23.98	0.1	5.02	5.56	0.54
25°C 1.00327 g/cm^3 1.38158 nD	34.12	33.67	0.45	20.84	20.87	0.03
27°C 1.0913 g/cm^3 1.37904 nD	20.14	18.27	1.87	22.60	21.74	0.86
20°C 1.09576 g/cm^3 1.38019 nD	20.14	18.19	1.95	22.60	22.90	0.3
20°C 1.0810 g/cm^3 1.3639 nD	0	4.32	4.32	20.13	19.73	0.4

20°C 0.985299 g/cm^3 1.33795 nD	7.61	8.95	1.34	0	2.04	2.04

Prediction of commercial beverage samples

The beverages were products of CopperMuse distillery, who were collaborating with the project's industry partner Emerson (CopperMuse Distillery, n.d.). As ABV% is defined at 20°C, the beverage samples were passed through the experimental setup at 20°C to obtain the density and refractive index measurements. The measurements were then inputs to the models to generate the predicted composition (**Table 7**).

Table 7: Model predictions for real-world beverage samples at 20°C.

Beverages at 20°C	**Random forest regression model for ABW% prediction**				**Degree 8 polynomial regression model for Brix prediction**		
	ABV% labelled on the bottle	**ABW% predicted**	**Converted to ABV% predicted**	**Absolute error in ABV%**	**Brix actual**	**Brix predicted**	**Absolute Error in Brix**
Vertueux Vodka	40	36.57	43.92	3.92	0	0.02	0.02
Amaretto Liqueur	25	19.979	26.73	1.73	23.3	21.38	1.92

Coffee Liqueur	25	21.535	28.03	3.03	14	13.55	0.45
Silver Rum	40	37.52	45.19	5.19	1	0.76	0.24
Jalapeño Vodka	40	35.84	43.12	3.12	0	0.29	0.29

These real-world beverages differ significantly from the simple three-component mixtures used in model development. They contain a variety of flavoring agents, and sugars not in the form of pure sucrose. Hence, it was encouraging to observe that the model's predictions closely matched the actual values, particularly for that of sugar content. The models demonstrate potential for further refinement with additional data. Future approaches could include focusing on gathering additional data centered around popular beverage compositions to improve prediction accuracy to reach the ideal ± 0.3 ABV percentage point range.

Future directions and work discussion

One of the key recommendations for future work is to collect more data in the range of 0 to 10 Brix and 35 to 45 % ABV range. This would significantly improve model performance around the most commercially relevant composition of 40% ABV spirits. Also, with by collecting more data, other complex machine learning models can be explored which have not performed very well with the current limited dataset. A larger and more diverse dataset would enable the development of region-specific models, tailored to particular composition ranges based on the intended beverage type to be predicted.

At the start of the project, minimal planning had been done for mapping of the composition space prior to making measurements. In hindsight, plotting the target compositions on ABW% vs Brix graphs to identify underrepresented regions, would have been far more efficient. This would have also reduced redundancy.

Another improvement relates to the temperature measurement protocol. Initially data were collected at 1°C intervals, which proved to be highly time-consuming. Switching to 2°C increments significantly reduced measurement time. Consulting with more distilleries to obtain the specific temperature ranges used during bottling or quality control could help narrow the experimental temperature range even further, instead of using the entire possible range of 10-40°C. Feature analysis from the current models has shown that temperature contributes less to prediction accuracy compared to density and refractive index. This suggests that future efforts could adopt even coarser temperature increments for measurements.

Initial sample preparation and measurement was another area that presented significant challenges. Issues such as uneven mixing for high ABW% and high Brix compositions, unforeseen instrument downtime, and the sensitivity of the measurement equipment to air bubbles often led to delays and required repeated trial and error adjustments. While much of the initial work involved learning through experimentation, future personnel can benefit from these insights. With a better idea of potential pitfalls, subsequent work can adopt the current preparation and measurement protocols to minimize variability and reduce the likelihood of measurement errors. Instrument cleaning also turned out to be more time-consuming than initially anticipated. While the manufacturer recommends short drying and cleaning times of 10 to 15 s, extending the drying period to 2 to 3 min significantly improved measurement consistency due to proper removal of residual water or acetone.

Instead of relying solely on the equipment's inbuilt screen, future projects would benefit from connecting the instrument to a dedicated computer with the Mettler Toledo software installed. While this would require a relatively expensive annual software license, continuous connection would reduce the risk of data loss due to undetected interruptions and would also enable sending real time notifications to personnel in the event of an error. Although the instrument's accuracy and repeatability have been reliable, there were several instances where settings unexpectedly reverted, causing substantial delays. It is likely that the newer dedicated computer software versions offer better safeguards against such issues and quicker troubleshooting.

CHAPTER 4: CONCLUSION

This study demonstrated the feasibility of predicting solution compositions of liquid mixtures using physical measurements, primarily density and refractive index. Building each model separately for sugar and alcohol content proved effective for handling their different behavior. The Brix prediction model performed well, with most predictions falling within ± 1 Brix, even for commercially produced beverages beyond lab-made mixtures. In contrast, the alcohol prediction proved more challenging. While most errors remained within ± 2 ABV percentage points, the actual result did not meet the initial target of ± 0.3 ABV percentage points. This inability appeared to be the limited coverage of measurements over all possible compositions. More targeted data collection around specific beverage profiles, such as those with high alcohol content and low sugar content, in products like gin or vodka, would likely improve performance for distillery specific use cases. Using relatively simple models such as polynomial regression and random forest outperformed more complex models like neural networks. This is due to the limited dataset size, as complex regression models typically require more data to achieve optimal performance. A key takeaway was good predictions, especially for Brix, with actual commercial beverages that include additional additives beyond the ethanol, sugar, and water system. While alcohol predictions struggled when sugar content was near zero (e.g., jalapeño vodka), the model performed better for liqueurs with moderate sugar content (e.g., Amaretto).

This approach can be extended beyond this specific ternary solution. Since the ethanol, sugar, and water mixture exhibits non-ideal behavior and is relatively difficult to model, the modeling method could be even more effective in systems with more ideal behavior. Industries such as petrochemicals (e.g., water-to-oil ratio analysis) or cosmetics (e.g., ingredient concentration measurements for emulsion stability in lotions and shampoos), could benefit from similar modeling frameworks. Using additional measurement parameters such as viscosity and

speed of sound, models can handle mixtures with more than three components, opening the door to broader industrial applications.

REFERENCES

Alcohol and Tobacco Tax and Trade Bureau. (2022). *Distilled spirits labeling: Alcohol content*. Alcohol and Tobacco Tax and Trade Bureau. https://www.ttb.gov/regulated-commodities/beverage-alcohol/distilled-spirits/ds-labeling-home/ds-alcohol-content

Alcohol Help. (2025, June 3). *Types of alcohol*. Recovery Worldwide, LLC. https://www.alcoholhelp.com/alcohol/types/

Anton Paar. (2025). *Sucrose density*. Anton Paar Wiki. https://wiki.anton-paar.com/us-en/density-and-density-measurement/sucrose-density/

Anton Paar. (n.d.). *Inline beverage analyzer: Cobrix*. Anton Paar. https://www.anton-paar.com/us-en/products/details/cobrix/?srsltid=AfmBOor3BHP6T3-l6jzoPFr2s5Yyp05o4MmHcXioMy1hw2GiNkWEVbyf

Bouchard, A., Hofland, G. W., & Witkamp, G. J. (2007). Properties of sugar, polyol, and polysaccharide water-ethanol solutions. *Journal of Chemical and Engineering Data*, *52*, 1838–1842. https://doi.org/10.1021/je700190m

Burkov, A. (2020). *Machine learning engineering*. True Positive Inc.

Charles, D. F. (1965). Refractive Indices of Sucrose-Water Solutions in the Range from 24 to 53% Sucrose. *Analytical Chemistry*, *37*, 405–406. https://doi.org/10.1021/ac60222a027

CopperMuse Distillery. (n.d.). *Spirits*. CopperMuse. https://www.coppermuse.com/

Emerson. (2025a). *Densitometers, Viscometers, Refractometers*. https://www.emerson.com/en-us/catalog/micro-motion-coriolis?fetchFacets=true#facet:&partsFacet:&modelsFacet:&facetLimit:&searchTerm:&partsSearchTerm:&modelsSearchTerm:&productBeginIndex:0&partsBeginIndex:0&modelsBeginIndex:0&orderBy:&partsOrderBy:&modelsOrderBy:&pageView:grid&minPrice:&maxPrice:&pageSize:&facetRange:&

Emerson. (2025b). *Flow Meters for Mass, Volume & Density Measurement*. https://www.emerson.com/en-us/automation/measurement-instrumentation/flow-measurement/coriolis-flow-meters

European Commission. (n.d.). *Alcoholic beverages*. European Union. https://knowledge4policy.ec.europa.eu/health-promotion-knowledge-gateway/alcoholic-beverages_en

Fisher Scientific. (2025). *Sucrose, OmniPur™, Calbiochem™*. Fisher Scientific. https://www.fishersci.com/shop/products/sucrose-omnipur-calbiochem-emd-millipore-2/8510500GM#?keyword=sucrose

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd edition.)*. O'Reilly Media.

Grolier, J. P. E., & Wilhelm, E. (1981). Excess volumes and excess heat capacities of water + ethanol at 298.15 K. *Fluid Phase Equilibria*, *6*, 283–287. https://doi.org/10.1016/0378-3812(81)85011-X

Mettler Toledo. (n.d.-a). *Density measurement*. Mettler Toledo. https://www.mt.com/ca/en/home/applications/Application_Browse_Laboratory_Analytics/Density/density-measurement.html

Mettler Toledo. (n.d.-b). *Refractive index*. Mettler Toledo. https://www.mt.com/us/en/home/applications/Application_Browse_Laboratory_Analytics/Refractive_index/definition_and_measurement.html

Mettler-Toledo. (n.d.-c). *Operating instructions: Excellence density meter D4/D5/D6*. Mettler Toledo. https://www.mt.com/us/en/home/library/operating-instructions/lab-analytical-instruments/excellence-density-meter-d4-d5-d6-manual.html

Mettler-Toledo. (n.d.-d). *Operating instructions: Excellence refractometers R4/R5*. Mettler Toledo. https://www.mt.com/us/en/home/library/operating-instructions/lab-analytical-instruments/excellence-refractometers-r4-r5-manual.html

Mettler-Toledo. (n.d.-e). *MS204 analytical balance (phased-out product)*. Mettler Toledo. https://www.mt.com/us/en/home/phased_out_products/Laboratory_Weighing_Solutions/analytical-balances/MS204.html

Mettler Toledo. (n.d.-f). *Sucrose*. Mettler Toledo. https://www.mt.com/au/en/home/supportive_content/concentration-tables-ana/Sucrose_de_e.html

Mettler Toledo. (n.d.-g). *Ethanol*. Mettler Toledo. https://www.mt.com/us/en/home/supportive_content/concentration-tables-ana/Ethanol_de_e.html

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media.

National Institute of Standards and Technology. (2024, May 1). *Index of refraction of liquid water.* NIST Publications. https://webbook.nist.gov/cgi/fluid.cgi?P=1&TLow=10&THigh=40&TInc=2&Digits=5&ID=C7732185&Action=Load&Type=IsoBar&TUnit=C&PUnit=atm&DUnit=kg%2Fm3&HUnit=kJ%2Fmol&WUnit=m%2Fs&VisUnit=uPa*s&STUnit=N%2Fm&RefState=DEF

National Institute of Standards and Technology. (n.d.-a). *Isobaric Properties for Water*. NIST Chemistry WebBook. https://webbook.nist.gov/cgi/fluid.cgi?P=1&TLow=10&THigh=40&TInc=2&Digits=5&ID=C7732185&Action=Load&Type=IsoBar&TUnit=C&PUnit=atm&DUnit=kg%2Fm3&HUnit=kJ%2Fmol&WUnit=m%2Fs&VisUnit=uPa*s&STUnit=N%2Fm&RefState=DEF

National Institute of Standards and Technology (NIST). (n.d.). *Ethanol, physical and thermodynamic properties.* NIST Chemistry WebBook. https://webbook.nist.gov/cgi/cbook.cgi?ID=C64175&Mask=4

Parke, S. A., & Birch, G. G. (1999). Solution properties of ethanol in water. *Food Chemistry*, *67*, 241–246. https://doi.org/10.1016/S0308-8146(99)00124-7

Reis, J. C. R., Lampreia, I. M. S., Santos, Â. F. S., Moita, M. L. C. J., & Douhéret, G. (2010). Refractive Index of Liquid Mixtures: Theory and Experiment. *ChemPhysChem, 11, 3722–3733*. https://doi.org/10.1002/cphc.201000566

Resa, P., Elvira, L., Montero De Espinosa, F., & Gómez-Ullate, Y. (2005). Ultrasonic velocity in water-ethanol-sucrose mixtures during alcoholic fermentation. *Ultrasonics, 43, 247–252*. https://doi.org/10.1016/j.ultras.2004.06.005

Sarkis, A. (2023). *Training data for machine learning*. O'Reilly Media

Scott, T. A. (1946). Refractive index of ethanol-water mixtures and density and refractive index of ethanol-water-ethyl ether mixtures. *Journal of Physical Chemistry*, *50, 406–412.* https://doi.org/10.1021/j150449a003

Smithsonian Institution. (n.d.). Ethanol–water density tables. Document provided by Emerson Micro Motion

Theobald, O. (2024). *Machine learning with Python.* Packt Publishing.

Weinstein, J. (2010). Multiphase Flow in Coriolis Mass Flow Meters – Error Sources and Best Practices. *28th International North Sea Flow Measurement Workshop October 2010.* https://nfogm.no/wp-content/uploads/2019/02/2010-17-Multiphase-Flow-in-Coriolis-Mass-Flow-Meters-%E2%80%93-Error-Sources-and-Best-Practices-Weinstein-Emerson.pdf