UNIVERSIDADE DE TRÁS-OS-MONTES E ALTO DOURO
SCHOOL OF SCIENCES AND TECHNOLOGY
DEPARTMENT OF ENGINEERING

Predicting Oenological Attributes Using Machine Learning Models

Master's Thesis in Computer Engineering

Rui Manuel Machado Silva

Advisor:
Pedro José de Melo Teixeira Pinto, Ph.D.

Vila Real, 2017

Universidade de Trás-os-Montes e Alto Douro

Predicting Oenological Attributes Using Machine Learning Models

Master's Thesis in Computer Engineering

Rui Manuel Machado Silva

Advisor: Pedro José de Melo Teixeira Pinto, Ph.D.

Composition of the Jury:

Luís Filipe Leite Barbosa, Ph.D.

José Luís Guimarães Oliveira, Ph.D.

Pedro José de Melo Teixeira Pinto, Ph.D.

Vila Real, 2017

Thesis presented to the Universidade de Trás-os-Montes e Alto Douro, as a requirement to obtain a Master’s Degree in Computer Engineering under the guidance of Professor Pedro José de Melo Teixeira Pinto.

Author’s statement:

I declare that this written work has been carried out in accordance with the regulations of the Universidade de Trás-os-Montes e Alto Douro. The work is original, except where indicated by special reference in the text. Any views expressed are those of the author and do not represent in any way the views of the Universidade de Trás-os-Montes e Alto Douro. This work, in whole or in part, was not submitted for evaluation in other Portuguese or foreign higher education institutions.

Signature: Rui Manuel Machado Silva Date: 27/11/2017

ACKNOWLEDGEMENTS

To the magnificent dean of the University of Trás-os-Montes e Alto Douro for providing the opportunity to attend the Master's Degree in Computer Engineering.

To the European Investment Funds by FEDER/COMPETE/POCI – Operational Competitiveness and Internacionalization Programme, under Project POCI-01-0145-FEDER-006958 and National Funds by FCT – Portuguese Foundation for Science and Technology, under the project UID/AGR/04033/2013, and through Project I&D Interact – Integrative Research in Environment, Agro-Chains and Technology, under NORTE-01-0145-FEDER-000017, in their research entitled "Fostering viticulture sustainability for Douro Valley: multidisciplinary efforts from field to wine (VitalyWINE)", co-financed by "Fundo Europeu de Desenvolvimento Regional (FEDER)" under NORTE 2020, for their support in this investigation.

To my advisor, Professor Pedro José de Melo Teixeira Pinto, for the extreme dedication, total availability and support he provided me in the preparation of this thesis.

To my family for the support and advice that allowed me to keep the focus in ending this extensive work.

ABSTRACT

The potential of hyperspectral images combined with machine learning algorithms to predict anthocyanin concentration, pH index and sugar content in grapes is presented as a starting point do develop flexible models with large generalization capacity to estimate oenological parameters.

In this context, in order to evaluate the generalization capacity of the machine learning procedures, a comparison with current state of the art approaches and between three different methods, Neural Networks (NNs), Decision Trees (DTs) and Support Vector Regression (SVR), when combined with hyperspectral images, was performed to predict the anthocyanin concentration, pH index and sugar content and support the adequate monitoring of wine quality.

The models were trained with six whole grape berries for each sample, using different approaches of cross-validation and data pre-processing. The oenological parameters were estimated using models trained with the spectra of 2012, 2013 and 2014 samples from the Touriga Franca variety, and the generalization capacity was tested using 2013 samples of the Tinta Barroca and Touriga Nacional varieties.

The results suggest that combining hyperspectral images with appropriate data analysis tools achieve accurate predictions. The machine learning methods were able to predict the values of oenological parameters without significant differences, improving the state of the art results.

Good indicators were obtained in the generalization capacity of the models, suggesting that a robust model capable of predicting oenological parameters on different varieties and harvest years of wine grapes can be obtained without additional training. An environmentally-friendly, fast and low-cost approach is therefore achievable and should be the subject of future testing.

Keywords: Hyperspectral Imaging, Neural Networks, Decision Trees, Support Vector Regression, Pre-Processing, Generalization.

RESUMO

O potencial das imagens hiperespectrais combinado com algoritmos de aprendizagem máquina para prever a concentração de antocianinas, o índice pH e o teor de açúcar em uvas é apresentado, como um ponto de partida para desenvolver modelos de estimação flexíveis e com grande capacidade de generalização para estimar parâmetros enológicos.

Neste contexto, para avaliar a capacidade de generalização dos procedimentos de aprendizagem máquina, uma comparação com a literatura atual e entre três diferentes métodos, *Neural Networks* (NNs), *Decision Trees* (DTs) e *Support Vector Regression* (SVR), quando combinados com imagens hiperespectrais, foi feita para prever a concentração de antocianinas, o índice pH e o teor de açúcar e suportar a monitorização adequada da qualidade do vinho.

Os modelos foram treinados com seis bagos de uva para cada amostra, utilizando diferentes abordagens de validação cruzada e de pré-processamento dos dados. Os parâmetros enológicos foram estimados utilizando modelos treinados com espectros de amostras de 2012, 2013 e 2014, da variedade de Touriga Franca, e a capacidade de generalização foi testada com recurso a amostras de 2013 das variedades de Tinta Barroca e Touriga Nacional.

Os resultados obtidos sugerem que combinar imagens hiperespectrais com ferramentas de análise de dados apropriadas permite atingir predições precisas, sendo os métodos de aprendizagem máquina capazes de prever os valores dos parâmetros enológicos sem diferenças significativas, melhorando os resultados da literatura atual.

Foram obtidos bons indicadores sobre a capacidade de generalização dos modelos, sugerindo que um modelo robusto capaz de prever parâmetros enológicos sobre diferentes variedades e anos de colheita das uvas pode ser obtido sem treino adicional. Uma abordagem amiga do ambiente, rápida e de baixos custos é assim passível de atingir e deverá ser objeto de testes futuros.

Palavras-Chave: Imagens Hiperespectrais, Neural Networks, Decision Trees, Support Vector Regression, Pré-Processamento, Generalização.

GENERAL INDEX

TABLES INDEX

FIGURES INDEX

GRAPHS INDEX

EQUATIONS INDEX

APPENDICES INDEX

ABBREVIATIONS AND ACRONYMS INDEX

ANOVA – One-Way Analysis of Variance.

DT – Decision Tree.

MSC – Multiplicative Scatter Correction.

MSE – Mean Squared Error.

NIR – Near-infrared.

NN – Neural Network.

PCA – Principal Component Analysis.

PLS – Partial Least Squares.

RMSE – Root Mean Squared Error.

R^2 – Determination Coefficient.

SNV – Standard Normal Variate.

SV – Support Vector.

SVR – Support Vector Regression.

TB – Tinta Barroca.

TF – Touriga Franca.

TN – Touriga Nacional.

CHAPTER I – INTRODUCTION

In an increasingly data-driven agriculture, the systematic evaluation of the quality of grapes is of major importance to the competitive market of wine production, representing an important surplus value. The Oporto wine sector (and Douro in general) has been following this evolution with the introduction of several technologies in different aspects of production, one of which is to assess non-intrusively the quality of the grapes, in particular the anthocyanin concentration, pH index and sugar content, allowing winemakers to obtain insights about their wine grapes more frequently, harvesting them at the optimal point of maturity and selecting them according to some quality features.

In this work, three machine learning models for oenological parameter estimation were implemented, namely Neural Networks (NNs), Decision Trees (DTs) and Support Vector Regression (SVR), with their efficiency and generalization capacity compared to state of the art results obtained with other machine learning algorithms and, more importantly, with purely chemometric methods like Partial Least Squares (PLS) regression. Additionally, methods to pre-process the data, smooth the spectra, reduce its dimensionality and validate the models' results were also implemented and their effects will be discussed.

Thus, this work was split into five chapters: the first, which addresses the research problem and produces an outline of the main objectives to be completed; the second, that provides a complete state of the art review of the methods and results published in the same area of research; the third, that gives a theoretical basis on the concept of hyperspectral images and the experimental setup used, the data pre-processing step, dimensionality reduction of the data, the validation methods used and the machine learning algorithms employed for the oenological parameter estimation; the fourth, where a critical analysis and discussion of the results obtained with each model is presented; and finally, the fifth, that gives general conclusions about the work and discusses possibilities for further research.

1.1. Research Problem

Viticulture, and the entire wine industry, has undergone recent changes. As this market becomes global, competitiveness becomes one of the main challenges faced by producers. In recent years, Portugal has been one of the countries to become very competitive in the production of wines, with special focus on Port wine, whose quality is undeniable and recognized throughout the world. To maintain prominence in today’s markets, it’s extremely important to ensure the high quality of the wines produced and to continue to improve the winemaking process.

The number of wine producers has increased and investment in this market has been encouraged, including the use of new technologies and methodologies to access the optimal point of maturity for grapes’ harvesting, and it’s essential to measure a number of oenological quality parameters in the grapes. In this context the anthocyanin concentration, pH index and sugar content parameters are of most importance since they have a direct influence on the quality of the wine, being related with the degree of ripening, acidity, percentage of alcohol in the wine produced, among others.

The traditional laboratory chemical analysis of grapes to assess ripening is time-consuming, costly, prone to errors and invasive (ultimately destroying grapes). With the sustained growth of computational power, new methodologies arise to deal with this problem.

> “As a fast and easy-to-operate technique, infrared spectroscopy has gained wide industrial acceptance for routine wine analysis […] it is anticipated that in the near future infrared spectroscopy will progressively become a routine method for process monitoring and process control in different stages of grape and wine production” (Dambergs, Gishen, & Cozzolino, 2015, p. 261).

So, hyperspectral image-based systems, coupled with powerful data analysis tools, can be used as a viable alternative that serves, in a more consistent and objective way, the purposes of inspection, evaluation and measurement, as they are defined as fast, cost-effective and non-invasive methods.

Hence, the main problem for this work is: how to evaluate oenological parameters of grapes using environmentally-friendly, fast and cheap methods? To find ways to answer this question, there are other more specific questions that delimit this research, such as: can the machine learning models implemented reliably determine the sugar content, anthocyanin concentration and pH index with a precision similar to traditional chemical analysis?

1.2. Motivation

The present work derived from the opportunity to learn about machine learning, artificial intelligence and data processing and apply this knowledge to a real-world problem, with real applications. The interest in these emerging research fields arose early in my second year of the licentiate degree when I was enrolled in a class about algorithms taught by my advisor, Professor Pedro José de Melo Teixeira Pinto, in which NNs models were one of the final topics, and it grew even more when I investigated these topics for my final licentiate project and was registered for the Machine Learning Course by Stanford University in the online education platform Coursera, taught by Professor Andrew Ng.

1.3. Objectives

The main objectives of this research were to develop three models to predict anthocyanin concentration, pH index and sugar content in grapes, with the pre-processing of the data based on available hyperspectral images, and analyse and compare the performance with the current state of the art approaches. More specifically, other requirements arose during the elaboration of the present work:

- Explain the different requirements for the proper functioning of the models.
- Ensure the validation of the predictive models based on performance estimation techniques.
- Compare the performance between machine learning algorithms and chemometric methods.
- Infer the importance of data processing (dimensionality reduction, scaling, normalization, among others) from the hyperspectral images to allow the correct analysis of the samples by the models implemented.
- Specify how to collect data from the environment setup in order to capture the hyperspectral images.
- Propose possible research tasks for the future.

CHAPTER II – STATE OF THE ART REVIEW

"Hyperspectral imaging (Gowen, O'Donnell, Cullen, Downey, & Frias, 2007; Hall, Lamb, Holzapfel, & Louis, 2002) is a technique that collects information concerning how objects reflect and absorb light as a function of their wavelength" (as cited in Fernandes *et al.*, 2011, p. 216), providing both spatial and spectral information – however, it involves complex data that requires powerful analysis tools to extract the necessary information from the underlying patterns in the spectra. In this chapter, it'll be provided a full review of the state-of-the-art methodologies that combine hyperspectral images and data analysis tools to predict anthocyanin concentration, pH index and sugar content on grapes, and a brief review of similar methodologies that intend to perform predictions on other chemical compounds on grapes and other fruits and vegetables.

Part of this work has been submitted by the author to a scientific journal.

2.1. Prediction Methodologies in Wine Grape Berries

"In the process of analysis and evaluation of wine grapes, anthocyanin concentration, pH index and sugar content are highly researched parameters because they are correlated with the flavour, colour and are good indicators of the grapes' ripeness" (Silva, Gomes, Faia & Melo-Pinto, 2016, para. 3). In the last years, the use of such parameters has been proposed, using different near-infrared (NIR) spectroscopy techniques: transmittance mode, "where the fruit surface viewed by the detector is diametrically opposite to the illuminated surface" (Schaare & Fraser, 2000, p. 175); interactance mode, "where the field of view of the detector is separated from the illuminated surface by a light seal in contact with the fruit surface" (idem, ibidem); and reflectance mode, "where the field of view of the light detector includes parts of the fruit surface directly illuminated by the source" (idem, ibidem).

To ease the comparison between authors, the works reviewed are separated by the spectroscopy technique used:

a) transmittance mode spectroscopy: Fernández-Novales, López, Sánchez, García-Mesa, and González-Caballero (2009) used a miniature fibre-optic NIR spectrometer system on the spectral region of 700-1060nm with a chemometric method, PLS regression, to measure sugar content and pH index; Geraudie and Ojeda (2010) measured the anthocyanin concentration on wine grape berries using a combination of multiple linear regression for prediction and a Principal Component Analysis (PCA) for dimensionality reduction.

b) interactance mode spectroscopy: Herrera, Guesalaga, and Agosin (2003) developed an approach to predict sugar content on wine grape berries, using 146 samples on the 800-1050nm NIR region, alongside models with multiple linear regression and PLS regression; Larraín, Guesalaga, and Agosin (2008) measured sugar content, pH index and anthocyanin concentration, on the NIR region of up to 1100nm, using PLS regression; Geraudie *et al.* (2009) as mentioned previously, built models with multiple linear regression and a PCA for dimensionality reduction, in this study to measure sugar (and water) content.

c) reflectance mode spectroscopy: in the present work, it’s more relevant to analyse previous results in reflectance mode spectroscopy, since it follows the same methodology used in this study. The works in reflectance mode can be further divided into:

    a. reflectance mode for a small number of berries in each sample: Arana, Jarén, and Arazuri (2005) implemented a model with PLS regression to measure the sugar content on the NIR region of 500-800nm; Wu, Huang, and He (2008) built a model combining PLS for dimensionality reduction (extracting the three best principal factors) and NNs for the prediction of sugar content; Cao, Wu, and He (2010) used a genetic algorithm to analyse the sugar content and pH index on wine grapes, processing both the whole spectra and selected wavelengths; Fernandes *et al.* (2015, 2011) used adaptive boosting NNs to measure anthocyanin concentration on a first approach and a classic NNs model to predict anthocyanin concentration, pH index and sugar content on the latest, on the NIR region of 380-1028nm; Gomes, Fernandes, and Melo-Pinto (2017b); Gomes, Fernandes, Faia, and Melo-Pinto (2014a, 2014b) compared the effectiveness of both, NNs and PLS regression, for the prediction of sugar content on the NIR region of 380-1028nm;

    b. reflectance mode for a large number of berries in each sample: Cozzolino *et al.* (2005) used a modified PLS regression to measure the pH index on the

NIR region of 400-2500nm; Janik, Cozzolino, Dambergs, Cynkar, and Gishen (2007) compared the performance of NNs using both a PCA and PLS for dimensionality reduction on the prediction of anthocyanin concentration; Le Moigne *et al.* (2008) measured the anthocyanin concentration (among other parameters) on the NIR region of 250-310nm using PLS regression; Ferrer-Gallego, Hernández-Hierro, Rivas-Gonzalo, and Escribano-Bailón (2011) developed a modified PLS regression model to estimate the anthocyanin concentration on wine grapes; González-Caballero, Pérez-Marín, López, and Sánchez (2011) built a model with modified PLS regression operating on the 380-1700nm NIR region to predict sugar content and pH index; Hernández-Hierro, Nogales-Bueno, Rodríguez-Pulido, and Heredia (2013) used a modified PLS regression algorithm to measure the anthocyanin concentration on the 900-1700nm NIR region; Nogales-Bueno, Hernández-Hierro, Rodríguez-Pulido, and Heredia (2014) measured the sugar content and the pH index using a modified PLS regression model on the 900-1700nm NIR region; Chen *et al.* (2015) built two different models, using PLS regression and SVR, to predict the pH index and anthocyanin concentration on the 900-1700nm NIR region; Fadock, Brown, and Reynolds (2016) used PLS regression on the 350-850nm NIR region to predict sugar content, pH index and anthocyanin concentration.

It’s important to mention that the use of a larger number of berries represents a simpler problem than for a small number of berries, since the variability in berries spectra and reference oenological values evaluated attenuate along with the number of berries.

Table 1 summarizes the most relevant results obtained by each of the aforementioned authors using hyperspectral imaging in reflectance mode.

Table 1 – Literature results for the prediction of oenological parameters on whole grape berries, with hyperspectral imaging performed in reflectance mode

		External Test Set	
		R^2	RMSE
Anthocyanin Concentration	Chen *et al*. (2015) [c]	$^{N.P}$ 0.941	N.C.
	Fadock *et al*. (2016) [c]	0.650	75.000 mg. L^{-1}
	Fernandes *et al*. (2011)	$^{N.P}$ 0.650	N.C.
	Fernandes *et al*. (2015)	0.950	14.000 mg. L^{-1}
	Ferrer-Gallego *et al*. (2011)	$^{N.P}$ 0.970	N.C.
	Hernández-Hierro *et al*. (2013) [c]	$^{N.P}$ 0.860	N.C.
	Janik *et al*. (2007) [b, c]	0.900	N.C.
	Le Moigne *et al*. (2008) [c]	$^{N.P}$ 0.979	N.C.
pH Values	Cao *et al*. (2010)	$^{N.P}$ 0.957	0.126
	Cozzolino *et al*. (2005) [c]	$^{N.P}$ 0.850	0.150
	Fadock *et al*. (2016) [a, c]	0.560	0.050
	Fadock *et al*. (2016) [c]	0.810	0.050
	Fernandes *et al*. (2015)	0.730	0.180
	González-Caballero *et al*. (2011)	$^{N.P}$ 0.870	$^{N.P}$ 0.120
	Nogales-Bueno *et al*. (2014) [c]	$^{N.P}$ 0.940	0.120
Sugar Content	Arana *et al*. (2005)	0.710	1.270 ºBrix
	Cao *et al*. (2010)	$^{N.P}$ 0.820	$^{N.P}$ 0.960 ºBrix
	Fadock *et al*. (2016) [a, c]	0.710	0.870 ºBrix
	Fadock *et al*. (2016) [c]	0.890	0.650 ºBrix
	Fernandes *et al.* (2015)	0.920	0.950 ºBrix
	Gomes *et al*. (2014a)	0.959	1.026 ºBrix
	Gomes *et al*. (2014b)	0.948	0.939 ºBrix
	Gomes *et al*. (2017b) [a]	0.948	1.344 ºBrix
	González-Caballero *et al*. (2011) [c]	$^{N.P}$ 0.910	$^{N.P}$ 1.000 ºBrix
	Nogales-Bueno *et al.* (2014) [c]	$^{N.P}$ 0.990	$^{N.P}$ 1.370 ºBrix
	Wu *et al.* (2008)	0.908	N.C

a: Different vintage used in the external test set (generalization set).
b: Different vintage and variety used in the external test set (generalization set).
c: Large number of berries.
N.P: Not provided for external test set.
N.C: Not comparable.

2.2. Other Relevant Methodologies

Some other works can be found, but to measure different chemical compounds on wine grape berries and other fruits (i.e. phenolic compounds, solid sugar compounds or aroma compounds). Noguerol-Pato, González-Barreiro, Cancho-Grande, Martínez, *et al.* (2012); Noguerol-Pato, González-Barreiro, Cancho-Grande, Santiago, *et al.* (2012); Noguerol-Pato, González-Barreiro, Simal-Gándara, *et al.* (2012) built various approaches to study aroma compounds on different varieties of wine grape berries, using gas chromatography and mass spectrometry to determine the aromatic composition; Tarter and Keuter (2005) research focused on the differences in solid sugar compounds between the berries in different positions (top, middle and bottom) of a cluster.

Additionally, the use and effectiveness of support vector machines combined with hyperspectral imaging has already been tested and widely employed on classification problems; i.e., Melgani and Bruzzone (2004) addressed the problem of classification of hyperspectral remote sensing images comparing the effectiveness of support vector machines in hyperdimensional feature spaces with conventional feature-reduction-based approaches (radial basis function NNs and k-nearest neighbour classifiers); Mercier and Lennon (2003) presented modified kernels that take into consideration the spectral similarity between support vectors, applying them to images of an intensive agricultural region in France, selecting 17 bands from 450-950nm spectral data; Rumpf *et al.* (2010) used support vector machines for the early detection of plant diseases based on hyperspectral images obtained in reflectance mode; but approaches for regression are still slightly uncommon.

CHAPTER III – METHODOLOGY

In this chapter, a clear description of the samples is given alongside a theoretical basis about the experimental setup for hyperspectral images, the data pre-processing, dimensionality reduction and model validation methods used and the machine learning algorithms employed, as a form of exposing and justifying the entire process that leads to the build-up of the prediction models that provide the final results regarding the anthocyanin concentration, pH index and sugar content on wine grape berries.

A One-Way Analysis of Variance (ANOVA) was performed to study whether there are any statistically significant differences between the means of the different sets of samples. For a complete description and mathematical formulation about the ANOVA method, consult Christensen (2011). The prediction models were developed using Matlab software (The Mathworks, 2016) and the descriptive statistics, boxplots and one-way ANOVA tests used to study samples were obtained with Minitab Software (State College PA, 2010).

Part of this work has been submitted by the author to a scientific journal.

3.1. Samples

The main subjects of this study were grape bunches of the Touriga Franca (TF) variety, widely recognized as one of the most important varieties for the production of Port wine in the Douro region due to its resiliency to plant diseases, fruity flavour and intense colour, harvested from the vineyards of Quinta do Bonfim in Pinhão, Portugal, in the years of 2012, 2013 and 2014, which is property of Symington Family Estates. In order to test the generalization capacity of the models (that is, the models' ability to predict values outside the known grape bunches used on the training process), samples from the Tinta Barroca (TB) and Touriga Nacional (TN) varieties were also collected on the year of 2013. To obtain the best possible training and testing setups, with an adequate range of values that represent grapes in different ripening stages, "it's important to test grapes between the beginning of veraison (transition from berry growth to berry ripening) and maturity, and from areas within the same vineyard under different conditions (sun exposition, water availability, soil quality, among others)" (Silva *et al.*, 2016, para. 5): consequently, 240 samples were collected in the year of 2012 (24 per day), 84 in the year of 2013 (12 per day) and 120 in the year of 2014 (12 per day) from the TF variety; 84 samples (12 per day) and 60 samples (12 per day) were collected in the year of 2013 from the TB and TN varieties; all of which from three different regions on the vineyard considering vine trees with small, medium and large vigour.

The hyperspectral image acquisition was performed on fresh grape berries: "each sample measured by hyperspectral imaging was composed of six grape berries randomly collected in a single bunch. The berries were removed from bunches with their pedicel still attached […] all the samples were kept frozen, at -18ºC" (Gomes *et al.*, 2014b, p. 2).

The chemical analysis was carried out with the six grape berries being defrosted at room temperature and then

> "crushed in a buffer solution of tartaric acid (pH 3.2) and ethanol (95%), by macerating, and the resulting mixture was kept overnight at 25ºC (Carbonneau & Champagnol, 1993); a centrifugation (SIGMA centrifuge 3K18, 20 min, 4ºC, spin at 7155g) was applied and a clear extract was collected and mixed with acidified ethanol (0.1% HCL)" (as cited in Gomes *et al.*, 2014b, p. 2).

Riberéau-Gayon and Stonestreet (1965) and Office International de la Vigne et du Vin (1990) indicates that:

> "The total anthocyanin concentration was determined photometrically by SO2 bleaching method (Riberéau-Gayon & Stonestreet, 1965). UV/VIS spectrophotometer (Shimadzu) and 1 cm path length, disposable cells were used for spectral measurements at 520 nm and the pigment content, expressed in mg.L^{-1}, was calculated from a calibration curve of malvidin-glucoside. All determinations were performed in duplicate and the juice released was analysed for the pH contents according to validated standard methods (Office International de la Vigne et du Vin, 1990, as cited in Gomes *et al.*, 2017a, p. 42)".

Tables 2, 3 and 4 provide descriptive statistics regarding the laboratory results of all the samples collected, on anthocyanin concentration, pH index and sugar content, respectively. Appendices A through I contain additional statistics, boxplots and one-way ANOVA tests of these values for a more detailed view of the data behaviour. For the TF variety in the 2014 vintage, there aren't any laboratory results available regarding anthocyanin concentration.

Table 2 – Descriptive statistics for the anthocyanin concentration of the laboratory results

Anthocyanin Concentration (mg.L^{-1})								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2012	240	160.28	(153.05; 167.51)	56.86	(52.19; 62.46)	3.89	173.84	257.82
TF 2013	82	207.18	(195.06; 219.31)	55.18	(47.84; 65.21)	16.28	221.90	269.75
TB 2013	84	173.32	(163.66; 182.97)	44.49	(38.63; 52.46)	50.97	185.29	247.76
TN 2013	60	224.86	(215.25; 234.47)	37.21	(31.54; 45.38)	123.68	236.62	319.90

Regarding the anthocyanin concentration values, Table 2 shows that the range of values in the different datasets are distinct, with very different means and standard deviation interval values in all populations. Appendix A clearly shows that there are outliers (observations that lie an abnormal distance from other values in a random sample from a population) in all datasets, while Appendix B allows to find differences in the centre, shape and variability among all boxplots – conducting an ANOVA test between datasets (Appendix C), it was found that there are significant differences among the means between the TF 2012 and TB 2013 samples and TF 2013 and TN 2013 samples.

Table 3 – Descriptive statistics for the pH index of the laboratory results

pH Index								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2012	240	3.55	(3.51; 3.60)	0.35	(0.32; 0.38)	2.85	3.58	4.23
TF 2013	81	3.72	(3.64; 3.80)	0.35	(0.31; 0.42)	3.05	3.74	4.44
TF 2014	120	3.49	(3.45; 3.54)	0.26	(0.24; 0.30)	2.93	3.51	3.97
TB 2013	84	3.59	(3.52; 3.66)	0.32	(0.28; 0.38)	2.90	3.60	4.48
TN 2013	60	3.59	(3.52; 3.66)	0.29	(0.24; 0.35)	3.00	3.64	4.13

As for the pH index values, Table 3 evidences a small range of values in the different datasets, with similar means and standard deviation interval values for all sets. Appendix D shows that outliers couldn't be found in these samples, while observing Appendix E shows that the centre, shape and variability among all boxplots is very uniform – conducting an ANOVA test between datasets (Appendix F), it was still found that there are significant differences in the means between the TF 2013 samples and the TF 2012 and TF 2014 samples.

Table 4 – Descriptive statistics for the sugar content of the laboratory results

Sugar Content (ºBrix)								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2012	240	16.93	(16.50; 17.35)	3.34	(3.07; 3.67)	9.06	17.06	24.72
TF 2013	82	19.45	(18.66; 20.23)	3.59	(3.11; 4.24)	8.10	20.03	25.00
TF 2014	120	13.55	(12.89; 14.21)	3.66	(3.25; 4.20)	7.87	13.00	25.66
TB 2013	84	22.49	(21.51; 23.47)	4.52	(3.92; 5.32)	11.40	23.44	30.85
TN 2013	60	23.26	(22.63; 23.89)	2.44	(2.07; 2.97)	17.20	23.84	27.20

Finally, studying the sugar content values, it's possible to conclude from Table 4 that all datasets have very unique descriptive statistics, with creased differences in the range of values, means and standard deviation interval values. Appendix G shows that outliers can be found for the datasets composed of TF 2013 and TF 2014 samples, while Appendix H shows very distinct centres, shape and variability among all boxplots – the ANOVA test between datasets (Appendix I) found significant differences in the means between TF 2012, TF 2013 and TF 2014 and all the other datasets, while the TB 2013 and TN 2013 varieties have significant differences in the means when compared to all harvest years of the TF variety.

The ANOVA tests allow an important detail to come into consideration, as the models will operate predictions on datasets with statistically different means and it should influence the result analysis made in Chapter IV. The outliers found in the different datasets were kept as part of future analysis, since it's very likely that outliers will always be found in further testing with new datasets from different varieties and vintages, and the model must be ready to reduce the importance of these values when composing a set of predictions.

3.2. Experimental Setup for Hyperspectral Images

As mentioned in Chapter II, hyperspectral imaging can be performed using different NIR spectroscopy techniques: transmittance mode, "where the fruit surface viewed by the detector is diametrically opposite to the illuminated surface" (Schaare & Fraser, 2000, p. 175); interactance mode, "where the field of view of the detector is separated from the illuminated surface by a light seal in contact with the fruit surface" (idem, ibidem); and reflectance mode, "where the field of view of the light detector includes parts of the fruit surface directly illuminated by the source" (idem, ibidem). In the present work, reflectance mode was chosen over transmittance and interactance mode since, for the same illumination scenario, the intensity of light coming from the grape is stronger, which facilitates measurements.

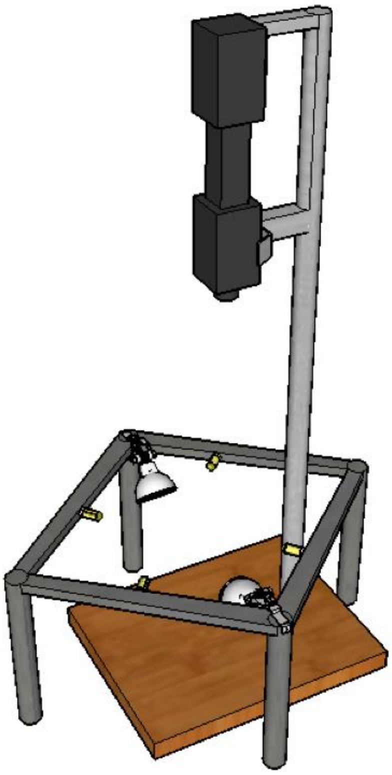
The experimental setup assembled for the images collected was:

> "a hyperspectral camera, composed of a JAI Pulnix (JAI, Yokohama, Japan) black and white camera and a Specim Imspector V10E spectrograph (Specim, Oulu, Finland); lighting, by means of a lamp holder with 300x300x175 mms (length x width x height) that held four 20W, 12V halogen lamps and two 40W, 220V blue reflector lamps (Spotline, Philips, Eindhoven, Netherlands). Both types of lamps were powered by continuous current power supplies to avoid light flickering and the reflector lamps were powered at only 110V to reduce lighting and prevent camera saturation" (Gomes *et al.*, 2014b, p.3).

The resulting hyperspectral images correspond to a single line over the sample and had 1040 wavelengths (ranging between 380 to 1028 nm, with approximately 0.6 nm of width in each channel) x 1392 pixels.

> "The 1392 pixels stand for the spatial dimension over the samples with approximately 110 mm of width. The distance between the camera and the sample base was 420 mm. The camera was controlled with Coyote software from JAI inside a dark room. All the hyperspectral measurements were done at room temperature" (Gomes *et al.*, 2014b, p.3).

Figure 1 illustrates the experimental setup assembled for the hyperspectral image acquisition.



Source: Silva *et al.*, 2016: para. 7.

Figure 1 – Experimental setup for hyperspectral image acquisition

Reflectance is the quotient between the intensity of the light reflected by an object and the light that illuminates that object, being a function of the light wavelength – these reflectance and absorption patterns across wavelengths can uniquely identify chemical compounds and, unlike transmittance and interactance mode, it's possible to perform the imaging without requiring contact between the spectrometer/camera and the sample.

For some position x, and at wavelength λ, the reflectance R can be expressed as:

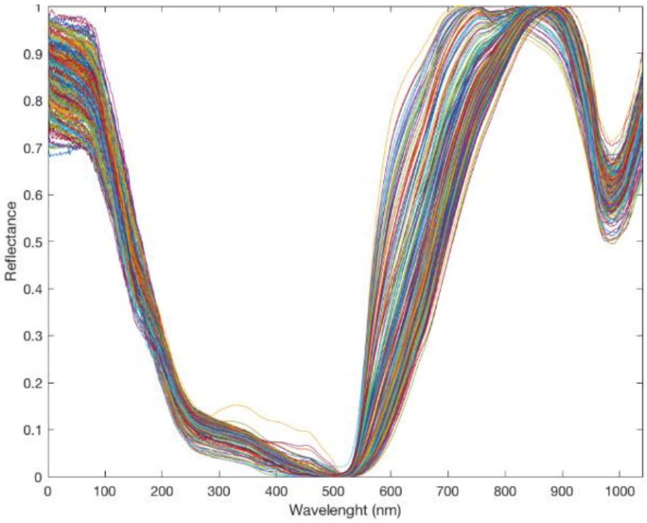
$$R(x,\lambda) = \frac{GI(x,\lambda) - DI(x,\lambda)}{SI(x,\lambda\ ) - DI(x,\lambda)}$$

Equation 1 – Expressing reflectance as a function of some position and wavelength

Where GI is the intensity of light coming from the grape, SI is the intensity of light coming from the Spectralon and DI is the dark current signal, which is electronic noise – this value is measured with the hyperspectral camera lens covered and it must be subtracted from the grape and the Spectralon signal because it's independent of the object being imaged and would distort the calculated reflectance values.

In order to achieve a reduction in measurement noise, 32 hyperspectral images were captured in each grape berry, from six grape berries and for three berry rotations – to create a single reflectance spectrum, all berries' points were averaged over the spatial dimension and rotation. The reflectance measurement was done along the berry "equator" with the pedicel as the pole and the final spectrum was normalized to reduce the noise in the measured light intensities cause by the grape berry size and curvature: the normalization was performed subtracting from each spectrum its minimum values and dividing by the difference between the minimum and maximum values (Gomes *et al.*, 2014b).

Graph 1 shows the final result for the reflectance measurements on the TF 2012 samples. Appendix J contains the results of the reflectance measurements for the remaining varieties and vintages of wine grape berries.



Graph 1 – Reflectance measurements for the TF 2012 samples

3.3. Data Pre-Processing

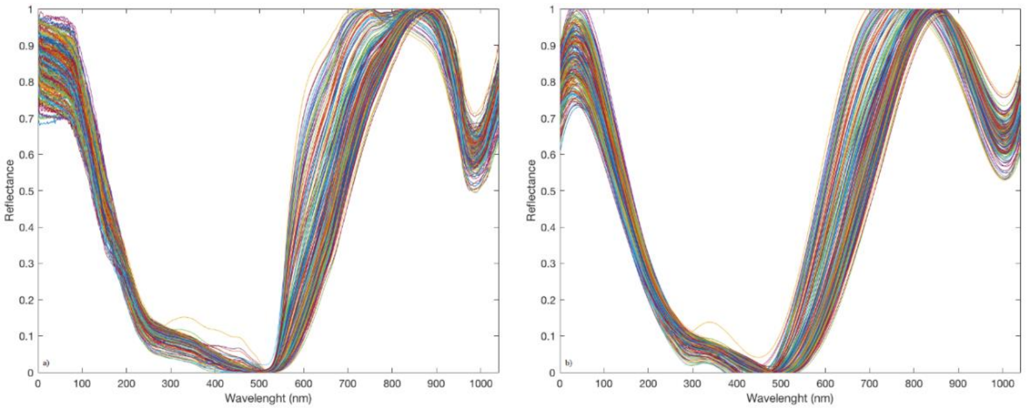
Prior to using the reflectance measurements as an input to the prediction models, it's important to study the possible effects of different pre-processing methods that intend to transform the data to a form more suited for analysis by the machine learning algorithms. In this context, the Standard Normal Variate (SNV) Transformation (Fadock, 2011), Derivatives (Arana *et al.*, 2005; Larraín *et al.*, 2008), Multiplicative Scatter Correction (MSC) (Herrera *et al.*, 2003) and the Savitzky-Golay Filter (Herrera *et al.*, 2003) arise as possible choices for testing, since they are frequently used in spectroscopic measurements for chemical analysis of grapes - however, some authors claim that the use of pre-processing methods is actually detrimental to the final results, with published results by the aforementioned authors questioning the positive effects of Derivatives, MSC or SNV Transformation when applied to chemometric data. In this work, the author chose to implement the Savitzky-Golay Filter for the purpose of smoothing the data and the SNV Transformation to correct scatter and study its effects when compared to a model without any pre-processing methods applied to the reflectance measurements.

As in Savitzky and Golay (1964), the Savitzky-Golay Filter is a technique that attempts to "smooth" the data, that is, to increase the signal-to-noise ratio without distorting the signal. The authors found that a "least squares calculation, may be carried out in the computer by convolution of the data points with properly chosen sets of integers" (idem, p. 1627), without additional computational complexity – in the general case of a group of C values representing any set of convolutional integers, the mathematical description of the convolution process, with an associated normalizing or scaling factor, is:

$$Y_j^* = \frac{\sum_{i=-m}^{i=m} C_i Y_{j+i}}{N}$$

Equation 2 – Description of the convolution process in a general case

With the index j representing the running index of the ordinate data in the original data table and N the normalizing factor. This calculation follows a common criterion to the best fit of least squares, with “the sole function of the computer is to act as a filter to smooth the noise fluctuations and hopefully to introduce no distortions into the recorded data” (idem, p. 1629). Graph 2 shows the reflectance measurements of the TF 2012 samples and the reflectance measurements of the TF 2012 samples after applying a Savitzky-Golay Filter with a frame length of 512nm and a third-degreed polynomial order, side-by-side to ease a comparison:



Graph 2 – Reflectance measurements for the Touriga Franca 2012 samples; a) Original; b) After applying the Savitzky-Golay Filter

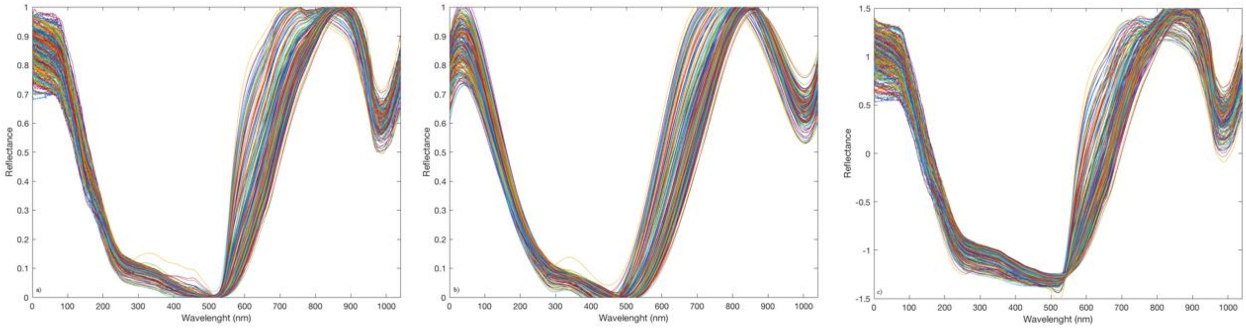
The smoothing and noise reduction of the reflectance measurements is easily observable in Graph 2, especially in the 0-100 and 700-900nm wavelength regions, meaning that the least squares found an apparent good fit to the data.

The SNV transformation, as in Barnes, Dhanoa and Lister is a “mathematical transformation of the $\log(\frac{1}{R})$ spectra by calculation of the standard normal variation at each wavelength […] removes slope variation on an individual sample basis by the use of the following calculation” (1989, p. 772):

$$SNV_{(1-W)} = \frac{(y_{1-W} - \bar{y})}{\sqrt{\frac{\sum(y_{(1-W)} - \bar{y})^2}{n-1}}}$$

Equation 3 – Calculation of the standard normal variation at each wavelength W

Where $SNV_{(1-W)}$ are the individual standard normal variations for W wavelengths, y is the W-wavelength $\log(\frac{1}{R})$ values, and $\bar{y}$ is the mean of the W-wavelength $\log(\frac{1}{R})$ values. This calculation intends to effectively remove the multiplicative interferences of particle size and scatter, since there is a high degree of collinearity between data points in the $\log(\frac{1}{R})$ spectra, which is a function to some extent of scatter and variable path length (idem). Graph 3 shows the reflectance measurements of the TF 2012 samples, the reflectance measurements of the TF 2012 samples after applying the described Savitzky-Golay Filter and the reflectance measurements after applying the SNV transformation to the data (applied to the original reflectance measurements), side-by-side to ease a comparison.



Graph 3 – Reflectance measurements for the TF 2012 samples; a) Original; b) After applying the Savitzky-Golay Filter; c) After applying the SNV transformation

Observing Graph 3, it’s noticeable that the range of reflectance values is different and that some of the graph peeks, either minimum or maximum, have a slightly different representation, possibly meaning that some interference due to scatter or particle size was removed in those regions.

Table 5 contains the test set results obtained with a prediction model with the application of 10-fold cross-validation, PCA and NNs trained via the Levenberg-Marquardt algorithm and random initialization of the weights and bias (methods covered further in this chapter) for the estimation of sugar content on TF 2012 samples, to compare the different pre-processing methods.

Table 5 – Results obtained for the prediction of sugar content on TF 2012 samples with different pre-processing methods

				Test Set		
				R^2	RMSE (ºBrix)	PC
NNs	Sugar Content	TF 2012	Original	0.914	1.340	17
			SG Filter	**0.952**	**0.820**	**18**
			SNV Transformation	0.946	0.988	15

PC: Principal Components used.
SG: Savitzky-Golay.
SNV: Standard Normal Variate.

These results show that, for this particular test, a pre-processing step with the Savitzky-Golay filter is the best choice, since it’s the setup with the least Root Mean Squared Error (RMSE) and the best Coefficient of Determination (R^2). For effects of this work, the author chose to include the Savitzky-Golay Filter in the pre-processing step since it lowers the prediction errors seemingly without adding complexity to the model (similar number of principal components used in all pre-processing methods – further analysis regarding this topic is presented in Chapter IV): however, further investigation should be conducted, since the study of different pre-processing methods and its effects aren’t the main objective of this work and there’s a wide variety of methods and testing setups that could have been used.

To finish the data pre-processing step two widely used operations, mean-centering and auto-scaling, as in Bro and Smilde were employed to the data matrix. “Centering is performed to make interval-scale data behave as ratio-scale data, which is the type of data assumed in most multivariate models” (2003, p. 19) and this operation should allow a “reduced rank of the model", an "increased fit to the data" and the "avoidance of numerical problems" (idem, ibidem). The mean-centering operation can be described by the equation below:

$$MC(y) = y - \bar{y}$$

Equation 4 – Mean-centering operation on a dataset y

Where y represents the original data and $\bar{y}$ the vector with the mean of the dataset. As for scaling, as in Bro and Smilde, this operation is used to adjust for scale differences and to accommodate for heteroscedasticity (sub-populations with different variabilities), which is a very common concern in regression and variance analysis – in this work, the auto-scaling method was applied to let the variance of each variable be identical initially, so that the "subsequent fitting of a model is performed so as to describe as much systematic variation as possible […] every variable has the same initial opportunity of entering the model" (2003, p. 24). The equation below describes the auto-scaling operation:

$$X_f = \frac{X - \bar{X}}{\sigma(X)}$$

Equation 5 – Auto-scaling operation on a dataset *X*

Where X is the original data, $\bar{\bar{X}}$ the vector with the mean of the dataset and $\sigma(X)$ the vector with the standard deviation of the dataset.

3.4. Dimensionality Reduction

The major disadvantage of the hyperspectral imaging technique is the dimensionality of the reflectance spectra, since the resulting matrix has a dimensionality equal to the number of spectral channels measured by the hyperspectral camera (namely, 1040). The difficulties in processing such large multivariate datasets are noticeable and, in order to obtain the maximum performance of the machine learning algorithms employed, a significant reduction of the size of its input is strictly necessary – for this work, a Principal Component Analysis (PCA) was implemented.

As in Wold, Esbensen and Geladi

> “PCA provides an approximation of a data table, a data matrix, X, in terms of the product of two small matrices T and P' that captures the essential data patterns of X, providing the means to significantly reduce the size of a dataset without losing variability in the data. Plotting the columns of T gives a picture of the dominant “object patterns” of X and, analogously, plotting the rows of P' show the complementary «variable patterns»” (1987, p. 37).

The columns in T are called *score vectors* and the rows in P' are called *loading vectors*, while the deviations between projections and the original coordinates are termed the *residuals*, collected in the matrix E (idem). The equation below shows the PCA in matrix form as a least squares model of:

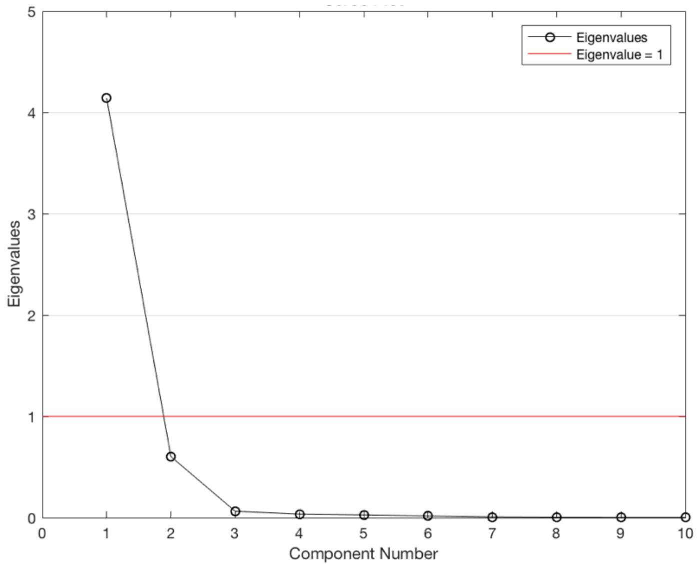
$$X = 1\bar{x} + TP' + \ E$$

Equation 6 – PCA in matrix form

Here the mean vector $\bar{x}$ is explicitly included in the model formulation.

> “A basic assumption in the use of PCA is that the score and loading vectors corresponding to the largest eigenvalues contain the most useful information relating to the specific problem, and that the remaining ones mainly comprise noise: therefore, these vectors are usually written in order of descending eigenvalues” (Wold, Esbensen and Geladi, 1987, p. 42).

Graph 4 shows a scree plot (plot of the eigenvalues in descending order) of the PCA implementation for the TF 2012 reflectance measurements:



Graph 4 – Scree plot of the PCA implementation for the TF 2012 reflectance measurements

In general cases, the number of factors retained for analysis are those with eigenvalues over 1: as seen in the graph, two principal components describe most of the variance in the population (roughly 96.7% of all the variance in a cumulative sum), while the other factors, as stated above, would mainly comprise noise – however, in this case this assumption might not be true due to the highly complex chemical interactions present in the samples that have an impact on the reflectance measurements – there isn't a clear answer to how many factors should be retained for analysis (only general rules of thumb, like the scree plot analysis) and in this work, every model was tested using between 1 and 50 principal components, saving the best result (further analysis is shown in Chapter IV).

Table 6 contains the test results obtained with a prediction model with the application of the Savitzky-Golay Filter, 10-fold cross-validation and NNs trained via the Levenberg-Marquardt algorithm and random initialization of the weights and bias (k-Fold Cross-Validation and NNs will be covered further in this chapter) for the estimation of sugar content on TF 2012 samples, to compare the use of the PCA before applying a machine learning algorithm on the "raw" reflectance measurements.

Table 6 – Results obtained for the prediction of sugar content on TF 2012 samples with and without the application of a PCA

				Test Set		
				R^2	RMSE (ºBrix)	PC
NNs	Sugar Content	TF 2012	**Principal Component Analysis**	**0.952**	**0.820**	**18**
			Original *	0.839	1.678	-

PC: Principal Components used.
*With Savitzky-Golay Filter.

The results presented show that, for this test setup, the NN model takes great benefit from a dimensionality reduction step, specifically with the implementation of a PCA, achieving a significantly better R^2 and a lowest RMSE – additionally, the computational cost when using the PCA before employing the machine learning algorithm is greatly reduced (which is expected, since the algorithm works with a significantly smaller set of inputs). For this work, the author chose to include the PCA as a dimensionality reduction step for every prediction model: however, there is a wide variety of dimensionality reduction methods that can be tested and further investigation in this topic should be conducted.

3.5. Model Validation

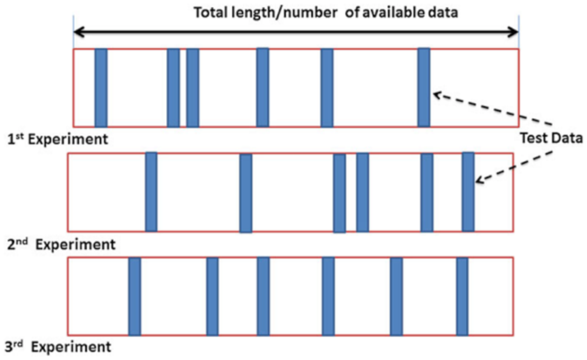
Model validation (or *model selection, model evaluation*)

> “can be understood primarily as a way of measuring the predictive performance of a statistical model, since high values of R^2 don’t necessarily mean a good fit – the model may have introduced too many degrees of freedom and inflate this statistics by overfitting the data, which means that predictions on new data will usually get worse as higher order terms are added. One way to measure the predictive ability of a model is to test it on a set of data not used in the estimation, a «*validation or test set*», instead of the «*training set*» used for estimation: however, there is often not enough data to allow for some of it to be kept back for testing” (Hyndman, 2010).

In this context, cross-validation and bootstrapping methods arise as a viable choice to improve the models’ generalization capacity without adding more samples to a dataset. In this work, the k-Fold Cross-Validation, Monte-Carlo Cross-Validation and Bootstrap methods were implemented and their efficiency compared as to choose the most adequate method to compose the predictive model; their description is found below, as in Lendasse, Wertz and Verleysen (2003).

The consecutive steps of the Monte-Carlo Cross-Validation are:

1. One randomly draws without replacement some elements of the dataset X; these elements form a new learning dataset X_{learn}. The remaining elements of X form the validation set X_{val} (see Figure 2).



Source: Remesan & Mathew, 2014: 64

Figure 2 – Data splitting in the random sub-sampling (Monte-Carlo) approach

2. The training of the model g is done using X_{learn} and the error $E_k(g)$ is calculated according to:

$$E_k(g) = \frac{\sum_{i=1}^{N_{prop}} (g(x_i^{val}) - y_i^{val})^2}{N_{prop}}$$

Equation 7 – Error of the Monte Carlo Cross-Validation method per repetition

With (x_i^{val}, y_i^{val}) the elements of X_{val}, N_{prop} the proportion of the training set chosen and $g(x_i^{val})$ the approximation of y_i^{val} by model g.

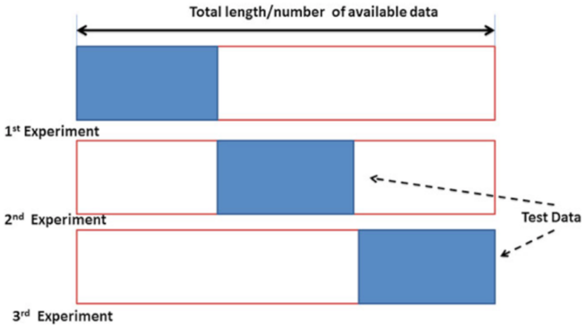
3. Steps 1 and 2 are repeated K times, with K as large as possible. The error $E_k(g)$ is computed for each repetition k. The average error is defined by:

$$\hat{E}_{gen}(g) = \frac{\sum_{k=1}^{K} E_k(g)}{K}$$

Equation 8 – Average generalization error of the Monte-Carlo Cross-Validation

The k-Fold Cross-Validation method is a variant of the Monte-Carlo Cross-Validation method. The consecutive steps of this method are:

1. One divides the elements of the dataset X into K sets of roughly equal size. The elements of k_{th} set form the validation set X_{val}. The other sets form a new learning dataset X_{learn} (check Figure 3).



Source: Remesan & Mathew, 2014: 65

Figure 3 – Data splitting in k-Fold Cross-Validation

2. The training of the model g is done using X_{learn} and the error $E_k(g)$ is calculated according to:

$$E_k(g) = \frac{\sum_{i=1}^{N/K}(g(x_i^{val}) - y_i^{val})^2}{N/K}$$

Equation 9 – Error of the k-Fold Cross-Validation method per set

With $(x_i^{val},\ y_i^{val})$ the elements of X_{val} and $g(x_i^{val})$ the approximation of y_i^{val} by model g.

3. Steps 1 and 2 are repeated for k varying from 1 to K. The average error is computed according to Equation 8.

The consecutive steps of the Bootstrap method are:

1. In the dataset X, one draws randomly N samples with replacement. These new samples form a new learning set X_{learn} with the same size as the original one. The validation set X_{val} is the original learning set X. This process is called re-sampling.
2. The training of the model g is done using X_{learn} and the errors $E_k^{val}(g)$ and $E_k^{learn}(g)$ obtained with this model are calculated according to the following equations:

$$E_k^{learn}(g) = \frac{\sum_{i=1}^{N}(g(x_i^{learn}) - y_i^{learn})^2}{N}$$

Equation 10 – Error of the Bootstrap method for the learning set per experiment

With $(x_i^{learn}, y_i^{learn})$ the elements of X_{learn} and $g(x_i^{learn})$ the approximation of y_i^{learn} obtained by model g;

$$E_k^{val}(g) = \frac{\sum_{i=1}^{N}(g(x_i^{val}) - y_i^{val})^2}{N}$$

Equation 11 – Error of the Bootstrap method for the validation set per experiment

With (x_i^{val}, y_i^{val}) the elements of X_{val} and $g(x_i^{val})$ the approximation of y_i^{val} obtained by model g.

3. The optimism $D_k(g)$, a measure of performance degradation (for the same model) between a learning set and a validation set is computed by subtracting to the error of the validation set, the error of the training set.
4. Steps 1, 2 and 3 are repeated K times, with K as large as possible. The average optimism is computed by:

$$\widehat{D}(g) = \frac{\sum_{k=1}^{K} D_k(g)}{K}$$

Equation 12 – Average optimism computation in the Bootstrap method

5. Once this average optimism is computed, a new training of the model g is done using the initial dataset X; the learning error E_g^I is calculated according to:

$$E_g^I = \frac{\sum_{i=1}^{N} (g(x_i) - y_i)^2}{N}$$

Equation 13 – Learning error in the Bootstrap method

With (x_i, y_i) the elements of X and $g(x_i)$ the approximation of y_i by model g.

6. Step 5 is repeated M times, with M as large as possible. For each repetition m the learning error E_g^I is computed. The apparent error $\hat{E}^I$ is defined as the average of errors E_m^I over the M repetitions. In the case of a linear model g, this repetition is not necessary; learning of a linear model gives a unique set of parameters, making all learning errors E_m^I equal. With nonlinear models, this repetition performs a (Monte-Carlo) estimate of the most probable apparent error obtained after training of g.
7. Finally, with the estimate of the apparent error and of the optimism, their sum gives an estimate of the generalization error.

$$\hat{E}_{gen}(g) = \hat{E}^{I}(g) + \hat{D}(g)$$

Equation 14 – Estimate of the generalization error for the Bootstrap method

Table 7 contains the test results obtained with a prediction model with the application of the Savitzky-Golay Filter, PCA and NNs trained via the Levenberg-Marquardt algorithm and random initialization of the weights and bias (topic covered further in this chapter) for the prediction of sugar content on TF 2012 samples, to compare the use of the different aforementioned model validation methods.

Table 7 – Results obtained for the prediction of sugar content on TF 2012 samples with different model validation methods

				Test Set		
				R^2	RMSE (ºBrix)	PC
NNs	Sugar Content	TF 2012	**k-Fold Cross-Validation**	**0.952**	**0.820**	**18**
			Monte-Carlo Cross-Validation	0.940	1.084	10
			Bootstrap	0.948	0.886	19

PC: Principal Components used.
k-Fold Cross-Validation with k = 10.
Monte-Carlo Cross-Validation with 20% of the samples for validation and K = 1000.
Bootstrap with M = 1000.

These results show that, for this test setup, the application of either model validation method obtains very similar results, with the k-Fold Cross-Validation method obtaining slightly superior values for R^2 and RMSE – farther, this method is by some distance the one with the smallest computation cost, with the Monte-Carlo Cross-Validation and the Bootstrap methods being extremely heavy with the 1000 repetitions chosen (usually a lower limit for the number of repetitions). For this work, the author chose the k-Fold Cross-Validation method as the model selection algorithm.

One important topic to cover in the model validation step is the "*bias-variance trade-off*" or "*bias-variance dilemma*":

"The variance reflects the sensitivity of the function estimate to a training sample. Less sensitivity means that the estimate will be more stable against changes (sampling variations) in the data and thus be less variable under repeated sampling" (Friedman, 1997, p. 60). However, high variance can cause an algorithm to model the random noise in the data, resulting in «*overfitting*»; "the bias reflects sensitivity to the target function - it represents how closely on average the estimate is able to approximate the target function" (idem, ibidem). High bias can cause an algorithm to miss relevant relations between features, resulting in «*underfitting*». So, "it is desirable to have both low bias and low variance, since both contribute to the squared estimation error in equal measure" (idem, ibidem). However,

> "the purpose of training is to gain information concerning the target function from the data: therefore, sensitivity to the training data is essential, and generally more sensitivity results in lower bias; this in turn increases variance, and so there is a natural «*bias-variance trade-off*» associated with function approximation" (idem, ibidem).

It's then worth acknowledging that, for k-Fold Cross-Validation, if the prediction model is stable for a given dataset, the variance of the cross-validation estimates will be very similar independent of the number of folds. k-Fold Cross-Validation with k values between 10-20 reduces the variance while increasing the bias; as k decreases to values between 2-5 and the sample sizes get smaller, there is an increased variance due to the instability of the training sets: in these situations, repeated runs should be performed (Kohavi, 1995).

3.6. Machine Learning Algorithms

Machine Learning is a field of computer science focused on the computers' ability to learn from a set of data, with problems ranging from clustering and dimensionality reduction to unsupervised, reinforcement and supervised learning. In this work, the focus is a problem of supervised learning regression: from a set of inputs, the model should be able to predict outputs based on the data features and relations, and these predictions will be compared to ground-truth results to further optimize the models' structure and parameters.

There's a wide variety of supervised learning algorithms fit for regression and the author chose to study and implement three different algorithms - NNs, DTs and SVR, compare their performance between themselves and with a chemometric method, PLS regression. The NNs algorithm was chosen since it's normally used by reference authors (as seen in Chapter II) due to its ability to model any function of any degree of accuracy (see 3.6.1.); the DTs algorithm (see 3.6.2.), because of its fast predictions, inherent simplified model and its ability to use ensemble methods, that allow to construct more than one DT to boost the model's generalization capacity; finally, the SVR algorithm was chosen (see 3.6.3.) because it has a regularization parameter, making it easier to avoid overfitting, it maps the input vectors to a high-dimensional feature space so that is possible to build expert knowledge about a problem via engineering the kernel function and, most importantly, the algorithm is defined as a convex optimization problem, for which there is no local minima (unlike NNs or PLS regression), using a subset of training points in the decision function, named *support vectors*, making the computational cost significantly smaller.

3.6.1. Neural Networks

The term "*neural network*" has its origins in attempts to find mathematical representations of information processing in biological systems (McCulloch & Pitts, 1943; Widrow & Hoff, 1960; Rosenblatt, 1962; Rumelhart, Hinton, & Williams, 1986), more specifically, trying to mimic what the human brain does. There are a wide number of NNs in use today, but in this work, we will only discuss feedforward NNs for supervised learning problems, since these are the most used type of networks in literature. A description of feedforward network functions is given with wording from Bishop (2006): a complete mathematical formulation can be found in the aforementioned book chapter.

Linear models for regression and classification are based on linear combinations of fixed nonlinear basis functions $\phi_j(x)$ and take the form:

$$y(x, w) = f\left(\sum_{j=1}^{M} w_j \phi_j(x)\right)$$

Equation 15 – Linear models for regression and classification

Where $f(\cdot)$ is a nonlinear activation function in the case of classification and is the identity in the case of regression. The goal is to extend this model by making the basis functions $\phi_j(x)$ depend on parameters and then to allow these parameters to be adjusted, along with the coefficients $\{w_j\}$, during training. There are many ways to construct parametric nonlinear basis functions. NNs use basis functions that follow the same form as Equation 15, so that each basis function is itself a nonlinear function of a linear combination of the inputs, where the coefficients in the linear combination are adaptive parameters.

This leads to the basic NN model, which can be described as a series of functional transformations. First, the construction of M linear combinations of the input variables $x_1, \ldots, x_D$ in the form:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}$$

Equation 16 – Expression of the quantities known as activations in the first layer of the network

Where $j = 1, \ldots, M$, and the superscript (1) indicates that the corresponding parameters are in the first "layer" of the network. The parameters $w_{ji}^{(1)}$ are referred as *weights* and the parameters $w_{j0}^{(1)}$ as *biases*. The quantities a_j are known as *activations*.

Each of them is then transformed using a differentiable, nonlinear activation function $h(\cdot)$ to give:

$$z_j = h(a_j)$$

Equation 17 – Expression of the hidden units

These quantities correspond to the outputs of the basis functions in Equation 15 that, in the context of NNs, are called *hidden units*. The nonlinear functions $h(\cdot)$ are generally chosen to be sigmoidal functions such as the logistic sigmoid or the hyperbolic tangent function. Following Equation 15, these values are again linearly combined to give output unit activations:

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)}$$

Equation 18 – Expression of the quantities known as activations in the second layer of the network

Where $k = 1, \dots, K$ and K is the total number of outputs. This transformation corresponds to the second layer of the network, and again the $w_{k0}^{(2)}$ are bias parameters. Finally, the output unit activations are transformed using an appropriate activation function to give a set of network outputs y_k. The choice of activation function is determined by the nature of the data and the assumed distribution of target variables.

Thus, for standard regression problems, the activation function is the identity so that $y_k = a_k$. A combination of these various stages give the overall network function that, for sigmoidal output unit activation functions, takes the form:

$$y_k(x, w) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

Equation 19 – Overall network function for sigmoidal output unit activation functions

Where the set of all weight and bias parameters have been grouped together into a vector w. Thus, the NNs model is simply a nonlinear function from a set of input variables $\{x_i\}$ to a set of output variables $\{y_k\}$ controlled by a vector w of adjustable parameters.

The bias parameters in Equation 16 can be absorbed into the set of weight parameters by defining an additional input variable x_0 whose value is clamped at $x_0 = 1$, so that Equation 16 takes the form:

$$a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i.$$

Equation 20 – Expression of the quantities known as activations after absorbing the bias parameters into the set of weight parameters in the first layer of the network

Similarly, the second-layer biases can be absorbed into the second-layer weights, so that the overall network function becomes:

$$y_k(x, w) = \sigma\left(\sum_{j=0}^{M} w_{kj}^{(2)} h\left(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\right)\right)$$

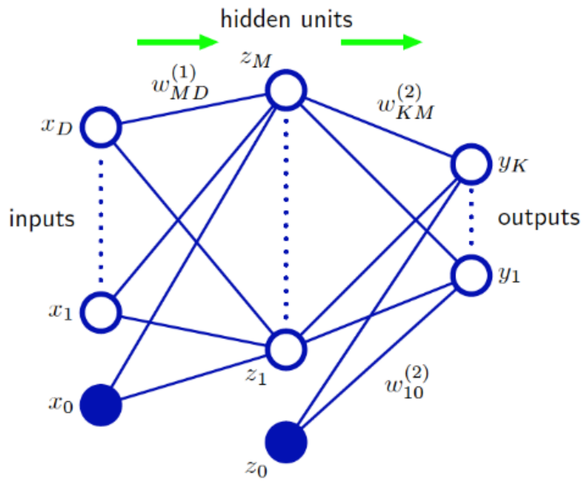
Equation 21 – Overall network function for sigmoidal output unit activation functions after absorbing the bias parameters into the set of weight parameters

The NN model comprises two stages of processing, each of which resembles a perceptron model and for this reason the NN is also known as the *multilayer perceptron*. A key difference compared to the perceptron, however, is that the NN uses continuous sigmoidal nonlinearities in the hidden units, whereas the perceptron uses step-function nonlinearities. This means that the NN function is differentiable with respect to the network parameters, and this property will play a central role in network training.

Because there is a direct correspondence between a network diagram and its mathematical function, one can develop more general network mappings by considering more complex network diagrams. However, these must be restricted to a *feed-forward architecture*, in other words to one having no closed directed cycles, to ensure that the outputs are deterministic functions of the inputs. This is illustrated with a single example in Figure 4. Each (hidden or output) unit in such a network computes a function given by:

$$z_k = h\left(\sum_{j} w_{kj} z_j\right)$$

Equation 22 – Hidden or output unit function in a network



Source: Bishop, 2006: 228

Figure 4 – Example of a NN having a feed-forward topology

Where the sum runs over all units that send connections to unit k (and a bias parameter is included in the summation). For a given set of values applied to the inputs of the network, successive application of Equation 22 allows the activations of all units in the network to be evaluated including those of the output units.

The next task is finding a weight vector w which minimizes the chosen error function $E(w)$. Because the error $E(w)$ is a smooth continuous function of w, its smallest value will occur at a point in weight space such that the gradient of the error function vanishes, so that:

$$\nabla E(w) = 0$$

Equation 23 – Condition to find the smallest value of the error function $E(w)$

Because there is clearly no hope of finding an analytical solution to Equation 23, one should resort to iterative numerical procedures. The optimization of continuous nonlinear functions is a widely-studied problem and there exists an extensive literature on how to solve it efficiently. Most techniques involve choosing some initial value $w^{(0)}$ for the weight vector and then moving through weight space in a succession of steps of the form:

$$w^{(\tau+1)} = w^{(\tau)} + \Delta w^{(\tau)}$$

Equation 24 – Technique to optimize continuous nonlinear functions

Where τ labels the iteration step. Different algorithms involve different choices for the weight vector update $\Delta w^{(\tau)}$. Many algorithms make use of gradient information and therefore require that, after each update, the value of $\nabla E(w)$ is evaluated at the new weight vector $w^{(\tau+1)}$.

The simplest approach to using gradient information is to choose the weight update in Equation 24 to comprise a small step in the direction of the negative gradient, so that:

$$w^{(\tau+1)} = w^{\tau} - \eta \nabla E(w^{\tau})$$

Equation 25 – Approach to using gradient information comprising a small step in the direction of the negative gradient

Where the parameter $\eta > 0$ is known as the *learning rate*. After each such update, the gradient is re-evaluated for the new weight vector and the process repeated. Note that the error function is defined with respect to a training set, and so each step requires that the entire training step be processed in order to evaluate ∇E. Techniques that use the whole data set at once are called *batch methods*. At each step the weight vector is moved in the direction of the greatest rate of decrease of the error function, and so this approach is known as *gradient descent* or *steepest descent*. Although such an approach might intuitively seem reasonable, in fact it turns out to be a poor algorithm. There is, however, an on-line version of gradient descent that has proved useful in practice for training NNs on large datasets (LeCun *et al.*, 1989). Error functions based on maximum likelihood for a set of independent observations comprise a sum of terms, one for each data point:

$$E(w) = \sum_{n=1}^{N} E_n(w)$$

Equation 26 – Error function based on maximum likelihood for a set of independent observations

On-line gradient descent, also known as *sequential gradient descent* or *stochastic gradient descent*, makes an update to the weight vector based on one data point at a time, so that:

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E_n(w^{(\tau)})$$

Equation 27 – Approach to use gradient information comprising a small step in the direction of the negative gradient for an on-line approach

This update is repeated by cycling through the data either in sequence or by selecting points at random with replacement. There are of course intermediate scenarios in which the updates are based on batches of data points.

The final goal is to find an efficient technique for evaluating the gradient of an error function $E(w)$ for a feed-forward NN. This can be achieved using a local message passing scheme in which information is sent alternately forwards and backwards through the network and is known as error backpropagation. This procedure can therefore be summarized as follows:

1. Apply an input vector x_n to the network and forward propagate through the network using Equation 28 and Equation 29 to find the activations of all the hidden and output units.

$$a_j = \sum_i w_{ji} z_i$$

Equation 28 – Weighted sum of the inputs in a general feed-forward NN

$$z_j = h(a_j)$$

Equation 29 – Transformation of the weighted sum of the inputs by a nonlinear activation function $h(\cdot)$

2. Evaluate the δ_k for all the output units using Equation 30.

$$\delta_k = y_k - t_k$$

Equation 30 – Evaluation of the derivatives for all the output units according to the errors δ

3. Backpropagate the δ's using Equation 31 to obtain δ_j for each hidden unit in the network.

$$\delta_j = h'(a_j)\sum_k w_{kj}\delta_k$$

Equation 31 – Backpropagation formula

(The value of δ for a particular hidden unit can be obtained by propagating the δ's backwards from units higher up in the network. This is only possible because the NN function is differentiable).

4. Use Equation 32 to evaluate the required derivatives.

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

Equation 32 – Evaluation of the derivatives for all the hidden units according to the errors δ

(This Equation tells that the required derivative is obtained simply by multiplying the value of δ for the unit at the output end of the weight by the value of z for the unit at the input end of the weight).

For batch methods, the derivative of the total error E can then be obtained by repeating the above steps for each pattern in the training set and then summing over all patterns:

$$\frac{\partial E}{\partial w_{ji}} = \sum_n \frac{\partial E_n}{\partial w_{ji}}$$

Equation 33 – Evaluation of the derivatives for each pattern in the training set

Concluding the description of the feed-forward NN model and its learning process, the final step is to choose the NN structure, training algorithm and form of initialization of the vector of weights and bias.

Regarding the NN structure, the activation function chosen for the hidden units was the bipolar sigmoidal function (related to *tangent hyperbolic*), described in Equation 34:

$$\sigma = \frac{2}{1 + e^{-u}} - 1$$

Equation 34 – Activation function chosen for the hidden units (bipolar sigmoidal function)

According to Kecman:

> "One of the first decisions to be made is how many hidden layers are needed in order to have a good model. First, it should be stated that there is no need to have more than two hidden layers. This answer is supported both by the theoretical result and by many simulations in different engineering fields […]. The real issue at present is whether one or two hidden layers should be used" (2001, p.267).

Referring to Hayashi, Sakata, and Gallant (1990), one should never try a multilayer model for fitting data until it was first tried a single-layer model, and this claim was somehow softened by calling it a rule of thumb: it's true that the simplest model possible should always be chosen but in this work, after multiple experiments on both setups, the NNs model will have two hidden layers.

Also, acknowledging Kecman:

> "The number of neurons in a hidden layer is one the most important design parameters with respect to the approximation capabilities of a NN. Recall that both the number of input components (*features*) and the number of output neurons is in general determined by the nature of the problem […]. In the case of general nonlinear regression […] the main task is to model the underlying function between the given inputs and outputs by filtering out the disturbances contained in the noisy training data set. By changing the number of hidden layer nodes, two extreme solutions should be avoided: filtering out the underlying function (not enough hidden layer neurons) and modelling of noise or overfitting the data (too many hidden layer neurons). In mathematical statistics, these problems are discussed under the rubric "*bias-variance dilemma*" […]. One of the statistical tools to resolve the trade-off between bias and variance is the cross-validation technique […]. In practical applications of NNs, one should build and train with cross-validation many differently structured NNs that differ in bias-variance and then pick the best one" (2001, p.268-271).

This is the procedure that the author ran in this work, and the best structure found regarding the number of neurons in a hidden layer is two neurons per hidden layer.

The training algorithm chosen was the Levenberg-Marquardt Error Backpropagation algorithm, described in (Levenberg, 1944) and further optimized in (Marquardt, 1963). The error function for stopping the learning function is the Mean Squared Error (MSE), and the number of epochs (iterations for weights adjustment) is 25.

Finally, for the form of initialization of the vector with the weights and bias, Kecman mentions that:

> “Initialization by using random numbers is very important in avoiding the effects of symmetry in the network. In other words, all the hidden layer neurons should start with guaranteed different weights. If they have similar (or, even worse, the same) weights, they will perform similarly (the same) on all data pairs by changing weights in similar (the same) directions. This makes the whole learning process unnecessarily long (or learning will be the same for all neurons, and there will practically be no learning)” (2001, p.292).

On a different note, there’s also the Nguyen-Widrow algorithm introduced in Nguyen and Widrow (1990), that generates initial weight and bias values for a layer so that the active regions of the layer’s neurons are distributed approximately evenly over the input space, and it’s said to make the training process faster. The author chose to implement both the random initialization and the Nguyen-Widrow methods and compare their efficiency. Table 8 contains the test results obtained with a prediction model with the application of the Savitzky-Golay Filter, PCA and NNs trained via the Levenberg-Marquardt algorithm for the prediction of anthocyanin concentration, pH index and sugar content on TF 2012 samples, to compare the use of both initialization approaches:

Table 8 - Results obtained for the prediction of sugar content, pH index and anthocyanin concentration on TF 2012 with different NNs initialization approaches

				Test Set			
				R^2	RMSE	PC	ET
NNs	TF 2012	Anthocyanin Concentration	Random Initialization	0.953	15.967 mg.L^{-1}	16	≈ 3 min.
			Nguyen-Widrow	**0.962**	**13.095 mg.L^{-1}**	**14**	≈ 90 sec.
		pH Index	**Random Initialization**	**0.871**	**0.147**	**11**	≈ 3 min.
			Nguyen-Widrow	0.840	0.160	11	≈ 90 sec.
		Sugar Content	**Random Initialization**	**0.952**	**0.820 ºBrix**	**18**	≈ 3 min.
			Nguyen-Widrow	0.917	1.108 ºBrix	11	≈ 90 sec.

PC: Principal Components used.
ET: Execution Time.
k-Fold Cross-Validation with 10 folds.

As seen in Table 8, for this test setup, the random initialization of the weights and bias achieved better results for the prediction of sugar content and pH index, while the Nguyen-Widrow algorithm had better results for the prediction of anthocyanin concentration – however, it can be considered that there isn't a significant difference between both methods (except on the sugar content parameter) and, although the Nguyen-Widrow algorithm reduces the execution time to around half, in this work it doesn't provide significant improvements for the computational cost. With this in mind, the author chose to implement a random initialization of the weights and bias in all NN models.

3.6.2. Decision Trees

DTs are predictive modelling algorithms that belong to the tree models class and, in machine learning, they're usually employed in classification problems – a DT can be easily identified as a flowchart-like structure with internal nodes, branches and leaf nodes. However, these algorithms can also be used for regression tasks, in which the DTs take the name of *regression trees.* According to Rokach and Maimon:

> "Regression trees are DTs that deal with a continuous target. The basic idea is to combine DTs and linear regression to forecast numerical target attributes based on a set of input attributes. These methods perform induction by means of an efficient recursive partitioning algorithm. The choice of the best split at each node of the tree is usually guided by a least squares criterion" (2015a, p.85).

A description of the regression trees algorithm is given with the words from Shalizi (2009), based on Hand, Mannila, and Smyth (2001a, 2001b). A complete mathematical formulation can be found in Rokach and Maimon (2015b).

In simple linear regression, a real-valued dependent variable Y is modelled as a linear function of a real-valued independent variable X plus noise:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Equation 35 – Linear regression modelling of the dependent variable *Y*

In multiple regression, there are multiple independent variables $X_1, X_2, \dots, X_p = X$:

$$Y = \beta_0 + \beta^T X + \epsilon$$

Equation 36 – Multiple regression modelling of the dependent variable *Y*

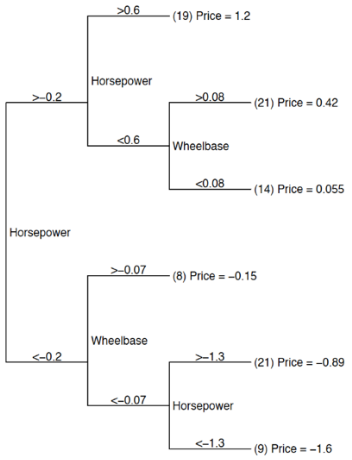
Linear regression is a global model, where there is a single predictive formula holding over the entire data-space. When the data has lots of features which interact in complicated, nonlinear ways, assembling a single global model can be very difficult, and hopelessly confusing when one succeeds. An alternative approach to nonlinear regression is to sub-divide, or *partition*, the space into smaller regions, where the interactions are more manageable. Then, those sub-divisions are partitioned again – this is called *recursive partitioning* – until finally one gets chunks of the space which are so tame that one can fit simple models to them. The global model thus has two parts: one is just the recursive partition, the other is a simple model for each cell of the partition.

Prediction trees use the tree to represent the recursive partition. Each of the *terminal nodes*, or *leaves*, of the tree represents a cell of the partition, and has attached to it a simple model which applies in that cell only. A point x belongs to a leaf if x falls in the corresponding cell of the partition. To figure out which cell one is, one starts at the *root node* of the tree, and goes through a sequence of "questions" about the features, since the *interior nodes* are "labelled with questions", and the *edges* or *branches* between them are "labelled by the answers". Notice that this basic idea of the recursive partition is based on discrete or categorical variables (classification problems), but in regression the variables will typically be continuous. As for the simple local models, for classic regression trees, the model in each cell is just a constant estimate of Y. That is, supposing points $(x_i, y_i), (x_2, y_2), \dots, (x_c, y_c)$ are all the samples belonging to the leaf-node l, the model for l is just:

$$\hat{y} = \frac{1}{c} \sum_{i=1}^{c} y_i$$

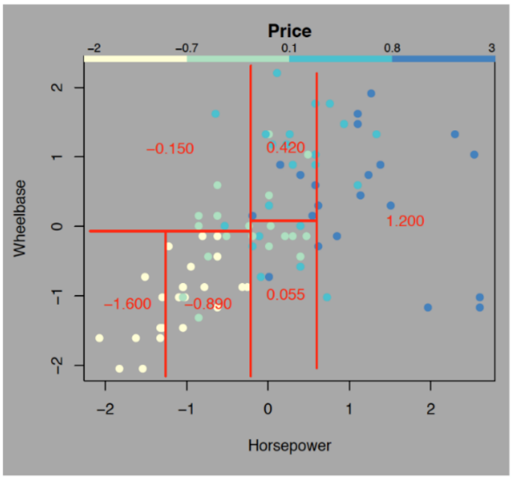
Equation 37 – Modelling of a leaf-node

The sample mean of the dependent variable in that cell. This is a piecewise-constant model - there are several advantages to this: predictions are fast, since there are no complicated calculations, just looking up constants in the tree; it's easy to understand what variables are important in making the prediction, just look at the tree; the model gives a jagged response, so it can work when the true regression surface is not smooth; and there are fast, reliable algorithms to learn these trees. Figure 5 shows an example of a regression tree which predicts the price of cars, while Figure 6 shows the partition of the data implied by the regression tree from Figure 5 (note that all the diving lines are parallel to the axes, because each internal node checks whether a single variable is above or below a given value.



Source: Shalizi, 2009: 3

Figure 5 – Regression tree for predicting the price of 1993-model cars



Source: Shalizi, 2009: 4

Figure 6 – The partition of the data implied by the regression tree from Figure 5

Once the tree is fixed, the local models are completely determined and easy to find (just compute the average), so all the effort should go into finding a good tree, which is to say into finding a good partitioning of the data. With regression trees, the goal is to maximize $I[C; Y]$, where Y is the dependent variable and C is the variable that keeps the information about the leaf position. A direct maximization can't be applied, so a *greedy search* is employed: one starts by finding the binary question which maximizes the information about Y; this gives the root node and two daughter nodes. At each daughter node, the initial procedure is repeated, asking which question would maximize the information about Y, given that one is already in the tree – this process is repeated recursively. However, one could just end up putting every point in its own leaf-node, which would not be very useful. A typical *stopping criterion* is to stop growing the tree when further splits gives less than some minimal amount of extra information, or when they would result in nodes containing less that a certain percent of the total data. So, the sum of squared errors of a tree T is:

$$S = \sum_{c \in leaves(T)} \sum_{i \in C} (y_i - m_c)^2$$

Equation 38 – Sum of squared errors of a tree *T*

Where:

$$m_c = \frac{1}{n_c} \sum_{i \in C} y_i$$

Equation 39 – Prediction for leaf *c*

Equation 38 can then be re-written as:

$$S = \sum_{c \in leaves(T)} n_c V_c$$

Equation 39 – Sum of squared errors of a tree *T* re-written to include the within-leave variance of a leaf *c*

Where V_c is the within-leave variance of leaf c. So, the tree splits will happen with the goal of minimizing S.

The regression-tree-growing algorithm can then be expressed as follows:

1. Start with a single node containing all points. Calculate m_c and S.
2. If all the points in the node have the same value for all the independent variables, stop. Otherwise, search over all binary splits of all variables for the one which will reduce S as much as possible. If the largest decrease in S would be less than some threshold δ, or one of the resulting nodes would contain less than q points, stop. Otherwise, take that split, creating two new nodes.
3. In each node, go back to step 1.

Trees use only one predictor (independent variable) at each step. If multiple predictors are equally good, which one is chosen is basically a matter of chance. One problem with the straight-forward algorithm presented is that it can stop too early, for example, there can be variables which are not very informative themselves, but that lead to very informative subsequent splits. This suggests a problem with stopping when the decrease in S becomes less than some δ. Similar problems can arise from arbitrarily setting a minimum number of points q per node.

A more successful approach to finding regression trees uses the idea of cross-validation: the data is randomly divided into a training and a validation set, the basic tree-growing algorithm is applied to the training data only, with $q = 1$ and $\delta = 0$ - that is, one grows the largest tree possible. This is generally going to be too large and will overfit the data, but then the cross-validation procedure is used to *prune* the tree. At each pair of leaf nodes with a common parent, the error is evaluated on the testing data and see whether the sum of squares would be smaller by removing those two nodes and making their parent a leaf. This is repeated until pruning no longer improves the error on the testing data. In each individual tree, this approach was employed.

However, for this work, an individual DT wasn't used since, despite all methods, they tend to still overfit - instead, a *bagging* (bootstrap aggregated) algorithm of DTs was implemented, based on Breiman (1996), that states that:

> "Bagging predictors is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the training set and using these as new training sets" (1996, p.123).

This technique reduces the effects of overfitting and improves generalization (idem). So, the DTs model presented in this work is actually a bagging of a number of individual DTs, each one implemented with the regression-tree-growing algorithm and cross-validation method described. There is no clear way of finding the optimum number of trees to use in the bagging algorithm, with a general "rule of thumb" being a number between 64-128 trees. For this work, the author chose 100 trees for each DTs model.

3.6.3. Support Vector Regression

In order to introduce the SVR model, a brief description of the Support Vector (SV) algorithm is given below, extracted from Smola and Sch$\ddot{o}$lkopf (2004). A complete mathematical formulation can be found in the aforementioned paper and also in Basak, Pal, and Patranabis (2007).

Suppose a given training data $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \chi \in \mathbb{R}$, where χ denotes the space of the input patterns – for instance, $\mathbb{R}^d$. In ε-SV regression as in (Vapnik, 1995), the goal is to find a function $f(x)$ that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time, is as flat as possible. In other words, one does not care about errors as long as they are less than ε, but will not accept any deviation larger than this.

In the case of linear functions f taking the form:

$$f(x) = \langle w, x \rangle + b, \qquad w \in \chi, \qquad b \in \mathbb{R}$$

Equation 40 - Example of a linear function f.

Where $\langle \cdot, \cdot \rangle$ denotes the dot product in χ. Flatness in the case of Equation 40 means that one seeks small w. One way to ensure this is to minimize the Euclidean norm, i.e. $\|w\|^2$. Formally, this problem can be written as a convex optimization problem by requiring:

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$
$$\text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases}$$

Equation 41 - Convex optimization problem of minimizing the Euclidean norm

The tacit assumption in Equation 41 was that such a function f actually exists that approximates all pairs (x_i, y_i) with ε precision, or in other words, that convex optimization problem is *feasible*. Sometimes, however, this may not be the case, or one may also want to allow for some errors. Analogously to the "soft margin" loss function in (Cortes and Vapnik, 1995), one can introduce slack variables ξ_i, ξ_i^* to cope with otherwise infeasible constraints of the optimization problem in Equation 41. Hence, one arrives at the formulation stated in (Vapnik, 1995):

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$
$$\text{subject to} \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Equation 42 – Convex optimization problem of minimizing the Euclidean norm with slack variables to cope with otherwise infeasible constraints

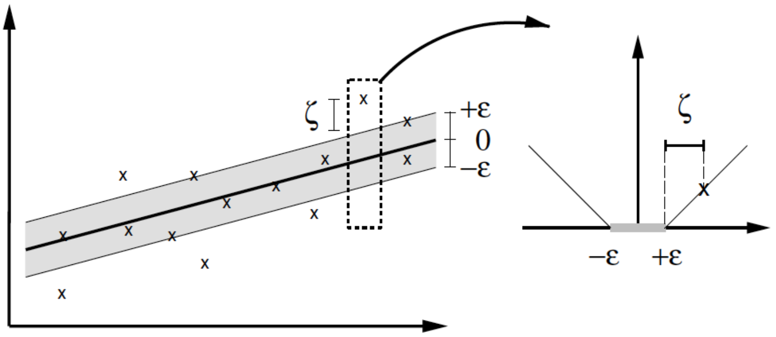
The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. The formulation above corresponds to dealing with a so called ε-insensitive loss function $|\xi|_\varepsilon$ described by:

$$|\xi|_\varepsilon := \begin{cases} 0, & if\ |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & otherwise. \end{cases}$$

Equation 43 – Description of the ε-insensitive loss function

An alternative to Vapnik's ε-SV Regression was introduced by (Chalimourda, Sch$\ddot{o}$lkopf, & Smola, 2004), named υ-SV Regression, where ε is not defined a priori but is itself a variable, its value traded off against model complexity and slack variables by means of a constant $\upsilon \in [0,1]$.

Figure 7 depicts the situation in Vapnik's algorithm graphically.



Source: Smola and Schӧlkopf, 2004: 200

Figure 7 – The soft margin loss setting corresponds for a linear SV machine

Only the points outside the shaded region contribute to the cost insofar, as the deviations are penalized in a linear fashion. The optimization problem in Equation 42 can be solved more easily in its dual formulation. Moreover, the dual formulation provides the key for extending support vector machines to nonlinear functions.

The key idea is to construct a Lagrange function from both the objective function (it will be called the *primal objective function* from so on) and the corresponding constraints, by introducing a dual set of variables. Hence, one proceeds as follows:

$$
\begin{aligned}
L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\
- \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\
- \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*)
\end{aligned}
$$

Equation 44 – Dual formulation via a Lagrange function of the objective function and the corresponding constraints

It is understood that the dual variables in Equation 44 have to satisfy positivity constraints, i.e. $\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$. It follows from the saddle point condition that the partial derivatives of L with respect to the primal variables (w, b, ξ_i, ξ_i^*) have to vanish for optimality – substituting this into Equation 44 yields the dual optimization problem:

$$\text{maximize} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \end{cases}$$

$$\text{subject to} \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases}$$

Equation 45 – Dual optimization problem

In deriving Equation 45 one eliminates the dual variables η_i, η_i^* due to the equalization of its partial derivatives to 0, as these variables did not appear in the dual objective function anymore but were present in the dual feasibility conditions. The so-called *Support Vector expansion* is then obtained following:

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i \ and\ therefore\ f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b$$

Equation 46 – Support Vector expansion

Where w can be completely described as a linear combination of the training patterns x_i. In a sense, the complexity of a function's representation by support vectors is independent of the dimensionality of the input space χ, and depends only on the number of support vectors. Moreover, the complete algorithm can be described in terms of dot products between the data. Even when evaluating $f(x)$ one doesn't need to compute w explicitly (although this may be computationally more efficient in the linear setting). These observations will come handy for the formulation of a nonlinear extension.

The next step is to make the support vector algorithm nonlinear. A computationally cheap way is to map the input vectors into a high-dimensional feature space through some nonlinear mapping, and then solve the optimization problem in that feature space. With the use of a suitable function k such as:

$$[\phi(x_i)\ \cdot\ \phi(x)] = k(x_i, x)$$

Equation 47 - Mapping the input vectors into a high-dimensional feature space.

Where $[\phi(x_i)\ \cdot\ \phi(x)]$ is the dot product of the input vectors' icons in a feature space $\mathcal{F}$, one obtains the nonlinear regression functions of the form:

$$f(x) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\ \cdot k(x_i, x) + b$$

Equation 48 – Expressing nonlinear regression functions

With the nonlinear function k being called a *kernel*. According to Smits and Jordaan:

> "[…] there are two main types of kernels, namely *Local* and *Global* kernels. In local kernels only the data that are close or in the proximity of each other have an influence on the kernel values. In contrast, a global kernel allows data points that are far away from each other to have an influence on the kernel values as well." (2002, p.2786).

In the present work, the author tested different configurations for the SVR model: an implementation with Vapnik's ε–SV Regression and Chalimourda's υ–SV Regression was conducted (different loss functions); and four types of kernels were on trial, namely linear, sigmoid, polynomial (global kernels) and gaussian radial basis (local kernel) kernels, described by Equation 49, 50, 51 and 52, respectively.

$$k(x, y) = x^T y + c$$

Equation 49 – Linear kernel function

$$k(x, y) = \tanh(\alpha x^T y + c)$$

Equation 50 – Sigmoid kernel function

$$k(x,y) = (\alpha x^T y + c)^q$$

Equation 51 – Polynomial kernel function

$$k(x,y) = \exp(-\gamma\|x-y\|^2)$$

Equation 52 – Gaussian radial basis kernel function

Where c is a constant term, α is the slope and γ is the free parameter of the gaussian radial basis function that controls the shape of the gaussian distribution.

Table 9 shows the results obtained with a prediction model with the application of the Savitzky-Golay Filter, PCA and SVR with a gaussian radial basis kernel optimized via random search algorithm (this subject will be introduced below) to compare the two different loss functions, on the estimation of anthocyanin concentration, pH index and sugar content on TF 2012 samples.

Table 9 – Results obtained for the prediction of anthocyanin concentration, pH index and sugar content on TF 2012 with different SVR loss functions

				Test Set		
				R^2	RMSE	PC
SVR	TF 2012	Anthocyanin Concentration	**Vapnik's ε-Regression**	**0.968**	**15.683 mg.L^{-1}**	**8**
			Chalimourda's υ-Regression	0.943	17.153 mg.L^{-1}	17
		pH Index	**Vapnik's ε-Regression**	**0.887**	**0.142**	**15**
			Chalimourda's υ-Regression	0.869	0.144	12
		Sugar Content	**Vapnik's ε-Regression**	**0.964**	**0.943 ºBrix**	**19**
			Chalimourda's υ-Regression	0.944	0.969 ºBrix	16

PC: Principal Components used.
k-Fold Cross-Validation with 10 folds.

Observing Table 9 one can conclude that for this test setup, despite very similar results, Vapnik's ε-SV Regression always gets the best values for R^2 and RMSE: consequently, the author chose to implement Vapnik's loss function on every SVR model.

Table 10 shows the results obtained with a prediction model with the application of the Savitzky-Golay Filter, PCA and Vapnik's ε-SV Regression to compare the use of the aforementioned kernels, with their free parameters optimized via random search algorithm (subject introduced below), on the estimation of sugar content, pH index and anthocyanin concentration on TF 2012 samples.

Table 10 – Results obtained for the prediction of anthocyanin concentration, pH index and sugar content on TF 2012 samples with different kernel functions on the SVR model

				Test Set		
				R^2	RMSE	PC
SVR	TF 2012	Anthocyanin Concentration	**Gaussian Radial Basis**	**0.968**	**15.683 mg.L^{-1}**	**8**
			Linear	0.943	16.740 mg.L^{-1}	18
			Sigmoid	0.943	18.100 mg.L^{-1}	6
			Polynomial	0.932	15.721 mg.L^{-1}	14
		pH Index	**Gaussian Radial Basis**	**0.887**	**0.142**	**15**
			Linear	0.820	0.164	10
			Sigmoid	0.872	0.116	8
			Polynomial	0.867	0.132	19
		Sugar Content	**Gaussian Radial Basis**	**0.964**	**0.943 ºBrix**	**19**
			Linear	0.937	1.176 ºBrix	8
			Sigmoid	0.910	1.131 ºBrix	19
			Polynomial	0.927	1.371 ºBrix	20

PC: Principal Components used.
k-Fold Cross-Validation with 10 folds.

As easily seen in Table 10, for this test setup the gaussian radial basis gets the best R^2 and RMSE values on every prediction: for this reason, the author chose to implement the gaussian radial basis kernel on every SVR model.

To finalize the SVR model there's still one important subject to consider: the free parameters (or *hyper parameters*, as they're commonly named) in the kernel function. In the case of the gaussian radial basis kernel, parameters C, ε (inherent to the SV algorithm) and γ (inherent to the kernel function) are of foremost importance to prevent the model from under or overfitting, since they control the "*bias-variance trade-off*" (mentioned in 3.5).

According to Cherkassky and Mulier, "the coefficient C affects the trade-off between complexity and proportion of the non-separable samples and must be selected by the user" (1998, p.366) while Alpaydin states that:

> "It is critical here, as in any regularization scheme, that a prover value is chosen for C, the penalty factor. If it is too large, one has a high penalty for non-separable point and it may store too many support vectors and overfit. If it is too small, one may have underfitting" (2004, p.224).

Horváth mentions that:

> “For a support vector machine the value of ε in the insensitive loss function should also be selected. ε has an effect on the smoothness of the support vector machine response and it affects the number of support vectors, so both the complexity and the generalization capability of the network depend on its value” (2003, p.392).

Referring to kernel parameters, Wang, Xu, Lu, and Zhang (2003) report that for regression problems, based on scale space theory, they demonstrate the existence of a certain range of σ, within which the generalization performance is stable. Seeing the importance of properly optimizing these parameters to maximize the SVR performance, the author chose to implement three different optimization methods and compare their performance: a random search algorithm based in Rastrigin (1963), a grid-search algorithm based in Zhang, Chen, Qu, Zhao, and Guo (2014) and a genetic algorithm based in Huang and Wang (2006). Table 11 shows the results obtained with a prediction model with the application of the Savitzky-Golay Filter, PCA and SVR on the estimation of anthocyanin concentration, pH index and sugar content on TF 2012 samples to compare the different optimization methods.

Table 11 – Results obtained for the prediction of anthocyanin concentration, pH index and sugar content on TF 2012 samples with different optimization methods for the parameters on the SVR model

				Test Set		
				R^2	RMSE	PC
SVR	TF 2012	Anthocyanin Concentration	**Random Search**	**0.968**	**15.683 mg.L^{-1}**	**8**
			Grid Search	0.904	15.546 mg.L^{-1}	14
			Genetic Algorithm	0.935	14.354 mg.L^{-1}	16
		pH Index	Random Search	0.887	0.142	15
			Grid Search	0.804	0.184	12
			Genetic Algorithm	**0.893**	**0.127**	**14**
		Sugar Content	**Random Search**	**0.964**	**0.943**	**19**
			Grid Search	0.933	0.973	14
			Genetic Algorithm	0.945	0.818	12

PC: Principal Components used.
k-Fold Cross-Validation with 10 folds.

Observing Table 11, one can see that for this test setup, the random search algorithm and the genetic algorithm obtain the best values for R^2 and RMSE on all predictions: however, considering the computational cost, the grid search and the genetic algorithms are extremely expensive, in contrast to the simple random search algorithm which additional computational time introduced to the model is practically irrelevant. Thus, the author chose the random search algorithm as the optimization method for the parameters on all SVR models. After several experiments, it was defined that $\varepsilon = 0.001$ and that the range of values for optimization was $C \ \in [80; 120]$ and $\gamma \ \in [0.00001; 0.001]$.

3.7. Final Prediction Models

To finish Chapter 3, Figure 8 shows a pipeline explaining the complete prediction models that will be used to estimate the anthocyanin concentration, pH index and sugar content of the different vintages and varieties of wine grape berries.

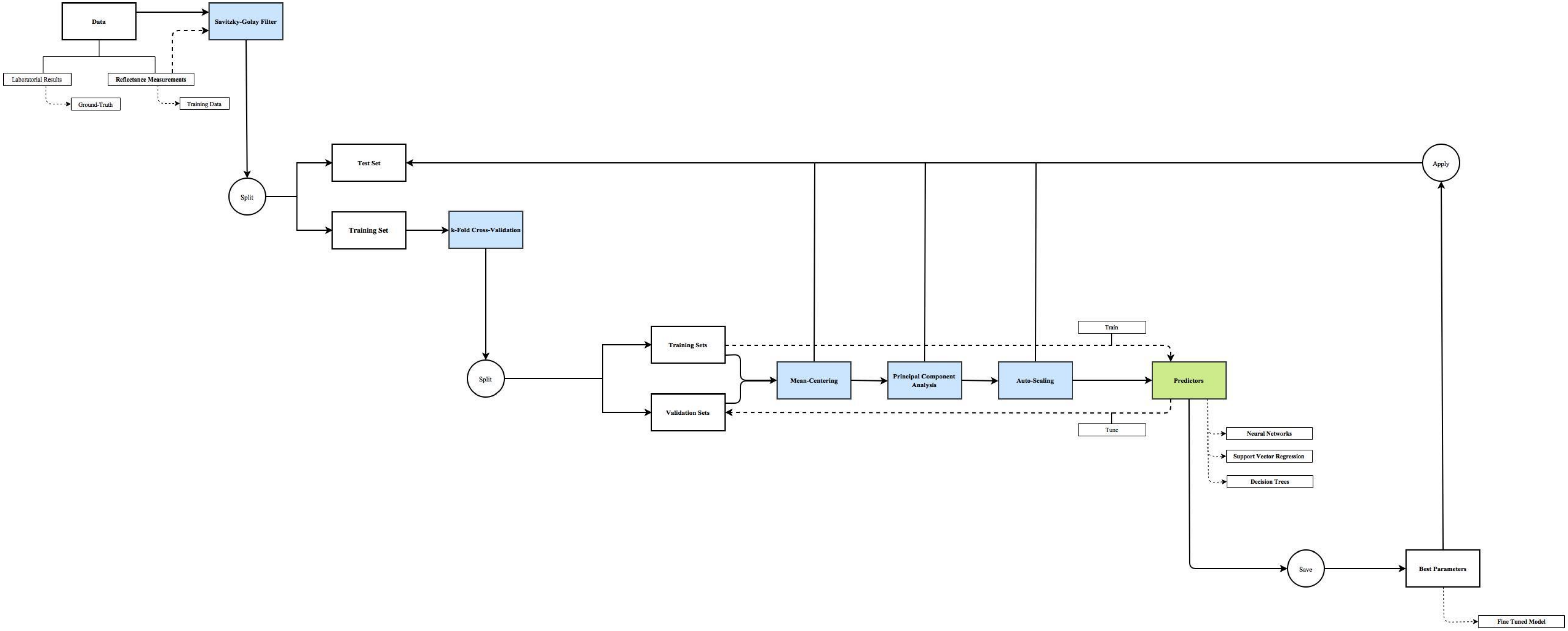


Figure 8 – Pipeline explaining the entire methodology for the models used for prediction

Referring to Figure 8, one starts with the data matrix, composed of the ground-truth laboratorial results and the reflectance measurements that will be used to train the predictor. Then:

1. Apply the Savitzky-Golay Filter (as seen in 3.3) to the reflectance measurements.
2. Divide the samples (includes the ground-truth results and the reflectance measurements after processing) into an independent test set and a training set.
3. k-Fold Cross-Validation is applied (check 3.5) dividing the training samples into training and validation folds.
4. Mean centering, PCA and auto-scaling are applied to the training and validation folds to reduce their dimensionality and facilitate the analysis (as seen in 3.3 and 3.4) – the reason why mean centering is used before the PCA and auto-scaling is used after is because "under the assumption that the PCA is obtained from the covariance matrix, the resulting principal components will be the same regardless of whether mean centering was performed or not as long as the covariance matrix stays the same" (Raschka, 2014, p.1), but "in contrast to mean centering, scaling does have an effect on the covariance matrix and therefore influences the results of a PCA" (idem, p.2).
5. The results of the pre-processing in 3 and 4 will serve as input to the machine learning algorithms, namely NNs (see 3.6.1), DTs (as in 3.6.2) and SVR (check 3.6.3), that will be trained with the data on the training folds and its parameters tuned by the data in the validation folds, in all k iterations.
6. Finally, the best parameters found in the training phase will be saved, which is commonly named as keeping the *fine-tuned model*, and this fine-tuned model will be applied to generate predictions on the test set, after the same pre-processing operations in 4 are applied to the samples.

CHAPTER IV – CRITICAL ANALYSIS AND DISCUSSION

In this chapter, the results of the different prediction models are studied to obtain possible insights about the quality of the models' fits and ways to improve their ability to learn from the data, comparing the results with the implemented state of the art approaches to measure oenological parameters on wine grape berries.

The descriptive statistics, regression plots and residuals plots used for regression analysis were obtained with Minitab Software (State College PA, 2010)

Part of this work has been submitted by the author to a scientific journal.

4.1. Experimental Outline

The experiments conducted were divided into two main test setups:

a) the test sets (with 10% of the samples left out for the test set), in which each wine grape variety and vintage data has its own training, validation and independent test sets with the fine tuning of the SVR free parameters occurring for all varieties;

b) the generalization sets (with 30% of the samples left out for the test set), composed of two different experiments – the first, with different vintages of wine grape berries employed on the independent test sets; the second, with different vintages and varieties of wine grape berries composing the independent test sets; with the fine tuning of the SVR free parameters occurring only for the vintage and variety of wine grape berries present on the training and validation sets, so that the true generalization capacity of the prediction model can be analysed.

The generalization capacity of a model is the ability of the learned model to fit unseen instances: the ultimate goal of machine learning is to achieve prediction models that can fit unseen instances with accuracy. A study of the models' generalization capacity in this work is of major importance, since the prediction of oenological parameters on wine grape berries can be considered an extremely complex case of generalization – for instance, in Portugal alone the number of different autochthonous wine grape varieties is close to 300, which highlights the importance of achieving models with a robust generalization capacity.

It is noteworthy to mention that: the NNs and DTs share the same topology (or structure) for all datasets, since optimizing it for every experiment could be understood as an intermediate step that doesn't allow for a true assessment of the models' generalization capacity (comparing this case with the SVR model, the range of values for the C and γ optimization also remains the same for all experiments, or else it could also be considered an intermediate step); for each oenological parameter, the lowest and the highest value obtained by laboratorial analysis is always present on the independent test sets, with the remaining samples for the training, validation and test sets chosen at random – this decision arose to attempt to have an independent test set that is an accurate representation of the population in study, or else it would be possible to have a skewed model obtaining good fits (that is, it would be possible for a model that couldn't actually learn from the data to have an apparently good generalization capacity because the samples chosen at random benefited the model's skewed predictions).

Table 12 provides detailed information about the experiences conducted and described in the two main test setups.

Table 12 – Outline of the different experiments performed in the sections below

	Anthocyanin Concentration		pH Index		Sugar Content	
	Train. / Val. Set	Test Set	Train. / Val. Set	Test Set	Train. / Val Set	Test Set
Test Sets	TF 2012	TF 2012	TF 2012	TF 2012	TF 2012	TF 2012
	TF 2013	TF 2013	TF 2013	TF 2013	TF 2013	TF 2013
	M.L.R.	M.L.R.	TF 2014	TF 2014	TF 2014	TF 2014
	TB 2013	TB 2013	TB 2013	TB 2013	TB 2013	TB 2013
	TN 2013	TN 2013	TN 2013	TN 2013	TN 2013	TN 2013
Generalization Sets	TF 2012	TF 2013	TF 2012	TF 2013	TF 2012	TF 2013
	M.L.R.	M.L.R.	TF 2012 & 2013	TF 2014	TF 2012 & TF 2013	TF 2014
	TF 2012 & 2013	TB 2013	TF 2012, 2013 & 2014	TB 2013	TF 2012, 2013 & 2014	TB 2013
	TF 2012 & 2013	TN 2013	TF 2012, 2013 & 2014	TN 2013	TF 2012, 2013 & 2014	TN 2013

M.L.R: Missing Laboratorial Results.

A descriptive statistical analysis has been performed to study the laboratorial results of the samples chosen to compose the generalization sets, so that a comparison can be made with respect to their similarity to the entire set of the same vintage and variety (check Appendices K through M).

In order to proceed to the regression analysis and a comparison between the present results and state of the art publications, two different indicators for the test setups and an extra indicator for the generalization sets were chosen: the quality of the fit, given by the R^2 (see Equation 53); the errors, expressed by the RMSE (see Equation 54); and, on the generalization sets, the residuals, with visual interpretation from the residuals vs fit values plots with the goal of finding possible ways to improve the prediction models and their generalization capacity. For more information about how to interpret residuals plots, please see Rawlings, Pantula and Dickey (1998).

$$R^2 = \left(\frac{\sigma_{y\hat{y}}}{\sigma_y \sigma_{\hat{y}}}\right)^2$$

Equation 53 – Calculation of R^2

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N-1}}$$

Equation 54 – Calculation of RMSE

Where y_i is the reference value, $\hat{y}_i$ is the model estimate, $\sigma_{y\hat{y}}$ is the covariance between y and $\hat{y}$ and σ_y, $\sigma_{\hat{y}}$ are the respective standard deviations.

As mentioned in 3.4, the number of principal components used as input for each model was chosen experimentally to yield maximum performance, with tests being performed using between 1 and 50 principal components. The number of folds, k, in the k-Fold Cross-Validation procedure takes the value of 5 (for the test setups with least samples, namely on the test sets experiments with the TF 2013, TF 2014, TB 2013 and TN 2013 samples) and 10 (for the remaining test setups, specifically the test experiments with the TF 2012 samples and all the generalization sets experiments): in the cases with only 5 folds, repeated runs where executed due to the increase in variance (as mentioned in 3.5).

4.2. Neural Networks

4.2.1. Test Sets

The validation and test set results obtained by the NNs model (one for each variety and vintage) for the prediction of anthocyanin concentration are presented in Table 13. As mentioned in 3.1, the TF variety on the vintage year of 2014 doesn't have any laboratory results available, preventing the development of a model for that particular set of samples.

Table 13 – Results for the determination of anthocyanin concentration on the test sets using NNs

		Validation Set		Test Set		
		R^2	RMSE(mg.L^{-1})	R^2	RMSE (mg.L^{-1})	PC
Anthocyanin Concentration	TF 2012	0.832	22.734	0.953	15.967	16
	TF 2013	0.802	23.306	0.968	15.463	8
	TB 2013	0.592	26.911	0.965	21.560	7
	TN 2013	0.759	16.927	0.821	27.471	4

PC: Principal Components used.

Observing the results, some remarks can be made: the NNs model shows accurate predictions with a small error rate for the test sets, but the R^2 and RMSE values on the validation sets might indicate that it suffers from a certain degree of underfitting for the case of the TB 2013 set of samples, since it has somewhat poor results on the training/validation step, but it has good results with a small error rate on the test set; the model accentuates the difficulties in having a quality training step and predictions for the datasets with the least standard deviations and the smallest range of values in their populations, namely the TB 2013 and TN 2013 datasets (Table 2): this might show that the model has some problems in capturing the patterns in the spectra when the prediction intervals are smaller and the standard deviations have low values; for the case of the TN 2013 dataset, the slight decrease in the model's performance can also be explained by the small number of samples (only 60 samples, the smallest dataset used); regarding the number of principal components used, it's observable that for the set of predictions that acquire the worst results (TB 2013 and TN 2013 datasets) the numbers used are the lowest, which indicates that the difficulties in learning can't be overcome by adding more principal components as input (probably because they mainly comprise noise).

Comparing the results with similar works from the literature that train and predict on the same varieties and vintages of wine grape berries, Fernandes *et al.* (2015) had the best results while using a machine learning algorithm (also NNs), obtaining a R^2 of 0.950 and a RMSE of 14.000 mg.L^{-1}: for three out of the four test sets used in the present work, superior results for the R^2 were obtained (R^2 of 0.953, 0.968 and 0.965 for the TF 2012, TF 2013 and TB 2013 test sets, respectively), while for the RMSE the values obtained are similar but slightly inferior for two out of the four test sets (TF 2012 and TF 2013, with RMSE of 15.967 and 15.463 mg.L^{-1}, respectively), but reasonably higher for the remainder; as for the authors using chemometric methods, Le Moigne *et al.* (2008) achieved the best results with his PLS regression model, with a R^2 of 0.979 (the errors are not comparable) for the training step, but no results were available for test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), also with a PLS regression model, with a R^2 of 0.650 and a RMSE of 75.000 mg.L^{-1}: in this work, superior results were obtained for both the R^2 and the RMSE on all the test sets.

The validation and test set results achieved with the NNs model (again, one for each variety and vintage) for the determination of pH index are shown in Table 14.

Table 14 – Results for the determination of pH index on the test sets using NNs

		Validation Set		Test Set		
		R^2	RMSE	R^2	RMSE	PC
pH Index	TF 2012	0.757	0.168	0.871	0.147	11
	TF 2013	0.515	0.241	0.834	0.175	5
	TF 2014	0.651	0.156	0.709	0.163	2
	TB 2013	0.419	0.235	0.746	0.253	2
	TN 2013	0.572	0.176	0.752	0.212	19

PC: Principal Components used.

Analysing the results, it’s noticeable that the model underperforms when predicting the pH index (in comparison to the anthocyanin concentration): as mentioned when studying the anthocyanin concentration, this could be due to the fact that the datasets have very small values for the standard deviations (close to 0) and a small range of values for the prediction intervals (Table 3); another possible explanation is that the greatest variation in the pH patterns is reflected on the model’s training step, which has difficulties to capture such relationships in the data without a greater number of samples for all datasets. Besides that, there’s an additional challenge in measuring the pH on wine grape berries, since the acidity is sensible to small changes in the condition of the sample (water content, temperature, etc.); other indicators seen in the anthocyanin concentration results can be recognized, namely the model underfitting on some of the vintages and varieties (for the pH index results, this is more noticeable on the TF 2013, TB 2013 and TN 2013 datasets) and the fact that the datasets in which the results obtained were the lowest, also have the smallest values of principal components used (for the pH index results the TN 2013 dataset is an exception, but this could be a consequence of choosing the test set samples at random).

Comparing the results with those published in literature for training and prediction on the same varieties and vintages of wine grape berries, Cao *et al.* (2010) had the best results using a non-chemometric method (genetic algorithm) with a R^2 of 0.957 and a RMSE of 0.126 for the training step, but no results can be found for the test sets – the highest values obtained that allow a direct comparison (test set values) are those obtained by Fernandes *et al.* (2015) with his NNs model, with a R^2 of 0.730 and a RMSE of 0.180: for four out of the five test sets used in this work, superior results for the R^2 were achieved (R^2 of 0.871, 0.834, 0.746 and 0.752 for the TF 2012, TF 2013, TB 2013 and TN 2013 test sets, respectively), while for the RMSE the values obtained are better for three out of the five test sets (TF 2012, TF 2013 and TF 2014, with RMSE of 0.147, 0.175 and 0.163, respectively), but higher for the remainder; as for the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.940 and a RMSE of 0.120 for the training step, but there aren't results available for the test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), with a PLS regression model, with a R^2 of 0.810 and a RMSE of 0.050: in this investigation, superior results were obtained for the R^2 of two out of the five test sets (TF 2012 and TF 2013, with R^2 of 0.871 and 0.834, respectively) but the RMSE values are slightly inferior for all experiments: therefore, it's implied that the difficulties in building a prediction model for the pH index are transversal to the other works published in literature.

The validation and test set results obtained with the NNs model (one for each variety and vintage) for the estimation of sugar content are presented in Table 15.

Table 15 – Results for the determination of sugar content on the test sets using NNs

		Validation Set		Test Set		
		R^2	RMSE(ºBrix)	R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012	0.912	0.972	0.952	0.820	18
	TF 2013	0.849	1.338	0.963	1.314	20
	TF 2014	0.746	1.883	0.915	1.216	20
	TB 2013	0.700	2.447	0.907	1.879	19
	TN 2013	0.656	1.302	0.852	1.552	9

PC: Principal Components used.

Examining Table 15 one can see that the results obtained are precise, with good values for the R^2 and RMSE on all test sets: nevertheless, as it was stated on the anthocyanin concentration and pH index results analysis, the model suffers from a certain degree of underfitting on some of the vintages and varieties (in this case, it’s more noticeable on the TB 2013 and TN 2013 datasets) and the datasets in which the results obtained were the less satisfactory, the number of principal components is also the lowest (for the sugar content the TB 2013 dataset is an exception but, as mentioned previously, this could be due to the fact that the test set samples are chosen at random); for the case of the TN 2013 dataset, the decrease on the model’s accuracy can also be explained by the fact that this dataset has a very small number of samples (at 60, is the dataset with the smallest number) with small standard deviation between samples and a small range of values for the prediction interval; it should also be highlighted the model’s capacity to achieve accurate predictions for all datasets with the same NN topology when the ANOVA tests showed that there are significant differences in the means between almost every set of samples (as seen in Appendix I).

Hence, comparing the results with the ones published in literature for training and prediction on the same varieties and vintages of wine grape berries, Gomes *et al.* (2014a) had the best results using a machine learning algorithm (also NNs) with a R^2 of 0.959 and a RMSE of 1.026 ºBrix: for the test set composed of TF 2013 samples superior results for the R^2 were obtained (R^2 of 0.963) but the remainder were somewhat inferior, while for the RMSE the test set with TF 2012 samples had a better RMSE value (RMSE of 0.820 ºBrix) but the rest were slightly worst; regarding the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.990 and a RMSE of 1.370 ºBrix for the training step, but the results for the test sets can’t be found – considering authors that allow a direct comparison of the results (test set results), Gomes *et al.* (2014b) with a PLS regression model obtained the best values for R^2 and RMSE, with 0.948 and 0.939 ºBrix, respectively: in this work, superior values for the R^2 were achieved on the TF 2012 and TF 2013 datasets (R^2 of 0.963 and 0.952, respectively) but for the remaining datasets the values were inferior, while for the RMSE the test set with TF 2012 samples had a better RMSE value (RMSE of 0.820 ºBrix) but the TF 2013, TF 2014, TB 2013 and TN 2013 datasets all got worst results on this parameter.

Overall, the NNs model achieved either superior or comparable results for the prediction of all oenological parameters in comparison to the state of the art approaches. Despite suffering from underfitting for some of the vintages and varieties (mostly the ones with a small number of samples for analysis), changes to the NN topology will only be considered after analysis of the generalization results, since the main goal is to obtain a model that can give accurate predictions for different vintages and varieties without fine tuning the NN structure for all datasets available.

4.2.2. Model Generalization

As mentioned previously, in order to study the models' generalization capacity two different experiments were used: the first, that applies a different vintage of the same varieties that compose the training and validation sets to the test sets (since the only variety that contains different vintage years is the TF, train and validation will occur on one or more vintages of TF and the test set will be composed of samples from the next vintage year); the second, that employs a different vintage and variety on the test set (in this case, all the TF vintage years will be used on the training and validation sets, while the test sets will be composed by the TB or TN samples).

In these experiments only the test set results are presented since it's not of major importance to understand how good the models' fit is early in the training process: it will have to generalize to distinct samples on the test set.

4.2.2.1. Different Vintages

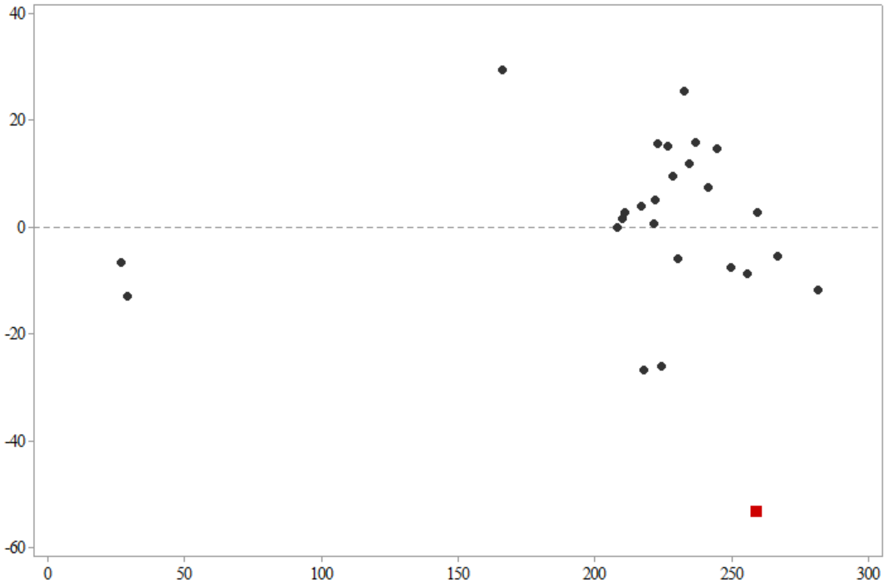
The test set results obtained by the NNs model for the prediction of anthocyanin concentration on different vintages are presented in Table 16. As mentioned previously, since the TF variety on the vintage year of 2014 doesn't have any laboratory results available, a model composed of TF 2012 and 2013 samples to predict TF 2014 values on the test set couldn't be built.

Table 16 – Results for the prediction of anthocyanin concentration on different vintages with NNs

		Test Set		
		R^2	RMSE ($mg.L^{-1}$)	PC
Anthocyanin Concentration	TF 2012 - TF 2013	0.922	20.504	10

PC: Principal Components used.

Observing the results presented, good indicators for a robust model with capacity to learn from wine grapes of different vintages are shown: there is a high correlation between the predictions and ground-truth results, as stated in the R^2 parameter, with only a small decrease of performance when compared to the single variety and vintage models and despite an increase in the RMSE, the value obtained is still similar to the single variety and vintage models that have a smaller number of samples (namely the TB 2013 and TN 2013 datasets); the number of principal components used as input also continues to be in the range of those used on the single test sets, indicating that there still isn't a necessity to increase the percentage of variance explained by the PCA to capture the relationships between the data; the descriptive statistics of the independent test set (see Appendix K) show that the mean and standard deviation values fit the initial 95% confidence intervals determined for the TF 2013 samples (on Table 2), which means that the independent test set is a good representation of the overall population. Graph 5 shows the residuals vs fit values plot for the prediction of anthocyanin concentration on the TF 2013 samples by the NNs model.



Graph 5 – Residuals vs fit values plot for the prediction of anthocyanin concentration on the TF 2013 samples by the NNs model.

Analysing Graph 5, despite some high valued residuals these are pretty evenly symmetrically distributed between the lower and higher digits of the y-axis and in general, no clear patterns can be found: however, some outliers are easily detected – assuming the outlying data points are legitimate, one could assess the impact of that data point on the regression, filtering it out and evaluating possible changes on the models' fit after that – but as mentioned in 3.1, the outliers will be kept part of the model since it's very likely that outliers will always be found in further testing with new datasets (e.g. samples are collected from the same area of a vineyard, all exposed to pretty similar conditions, but one of the branches has a significantly lower amount of sun exposition time due to a specific slope on the terrain: the values obtained on laboratorial analysis will probably compose an outlier) and the model must be ready to reduce the importance of these values when composing a set of predictions.

Analysing the results published in literature it was found that there isn't any work attempting to predict different vintages of wine grape berries on training and testing for the anthocyanin concentration; Janik *et al.* (2007) used not only different vintages but also different varieties on the test set, so a more adequate comparison will be made further in this chapter; comparing the results with the single vintage models, Chen *et al.* (2015), Ferrer-Gallego *et al.* (2011) and Le Moigne *et al.* (2008) had superior R^2 values (R^2 of 0.941, 0.970 and 0.979, respectively) with their PLS (and variants) regression models on the training step, but no results were published for the test set: Fernandes *et al.* (2015) is the only work with test set results published in which the R^2 and RMSE values are superior than those presented, with a R^2 of 0.950 (the R^2 shown is 0.922) and a RMSE of 14.000 $mg.L^{-1}$ (the RMSE shown is 20.504 $mg.L^{-1}$), applying a NNs model – however, considering that the results presented in Table 16 have different vintages applied on the testing phase, the small decrease in performance can be considered acceptable, as well as the values obtained can be considered very satisfactory, since they are still comparable with the remaining state of the art approaches.

The test set results obtained by the NNs model for the prediction of pH index on different vintages are presented in Table 17.

Table 17 – Results for the prediction of pH index on different vintages with NNs

		Test Set		
		R^2	RMSE	PC
pH Index	TF 2012 - TF 2013	0.773	0.204	15
	TF 2012 & 2013 - TF 2014	0.831	0.217	36

PC: Principal Components used.

Examining Table 17, once again it's noticeable that the model underperforms when compared to the results obtained for the anthocyanin concentration (highly expected, since despite adding more samples these are from different populations and the difficulties in estimating the pH index will still be the same); however, two important details arise that are of foremost importance: the model actually obtains better results when predicting for the TF 2014 dataset than for the TF 2013 set of samples, with very similar error measures, which can indicate that the initial diagnosis was correct – the model has difficulties in capturing the relationships between the patterns in the spectra for the pH index but, by adding a greater number of samples (even if these are from different populations) it will be able to achieve better results; and also, the model took a significantly greater number of principal components as input to operate on the TF 2014 test set, which may indicate that when the number of samples starts to grow, the number of factors on the PCA will not mainly comprise noise after the eigenvalues are over 1 but instead contain important information that can ease the model's learning step. The descriptive statistics of the independent test sets (see Appendix K) show that the mean and standard deviation values fit the initial 95% confidence intervals determined for both the TF 2013 and TF 2014 samples (on Table 3), meaning that the independent test sets are a good representation of the overall populations; additionally, the ANOVA tests (in Appendix F) previously shown that the TF 2013 dataset has a significant difference in the mean when compared to the TF 2012 and TF 2014 sets, which in a way praises the model's generalization capacity since it was able to predict for the TF 2014 set while learning from the TF 2012 and TF 2013 set of samples; analysing the residuals vs fit values plots (Appendix N and Appendix O), one can see that for the TF 2013 test set there are a slightly greater number of outliers when compared to the TF 2014 test set but overall, both plots have the residuals evenly symmetrically distributed and clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable, which aids the assumption that the NNs predictions for these setups are not skewed in any way or in need of adding/transforming some input variables.

Comparing these results with the ones published in literature for different vintages of wine grape berries employed on the training and testing phases, Fadock *et al.* (2016) is the only work published that meets this particular experimental outline, obtaining a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model: for both test sets in this work the R^2 values are superior (R^2 of 0.773 and 0.831 for the TF 2013 and TF 2014 test sets, respectively) but the RMSE on both setups is inferior (RMSE of 0.204 and 0.217 for the same test sets, respectively): however, regarding Fadock *et al.* (2016) results, it's rather questionable that a model that obtained such a low R^2 score has simultaneously such a low error measure, especially since it shares the same value than that obtained for the test set with only one vintage and variety employed on the training and testing phases (as seen in Table 1). Comparing the results with the ones presented for the single models in 4.1, a small decrease on the performance is noticeable when measured against the results obtained for the TF 2012 and TF 2013 datasets, but they're superior when compared to the remaining single vintage and variety models and the result obtained for the test set of TF 2014 samples after training on the TF 2012 and TF 2013 datasets (R^2 of 0.831 and RMSE of 0.217) is still superior when compared to the best test set results for single vintage and variety models previously published in literature [Fadock *et al.* (2016), with a PLS regression model obtaining a R^2 of 0.807 and RMSE of 0.050], which indicates that the NNs model has a very reasonable generalization capacity.

Table 18 presents the results for the prediction of sugar content on different vintages obtained by the NNs model.

Table 18 – Results for the prediction of sugar content on different vintages with NNs

		Test Set		
		R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012 - TF 2013	0.913	2.383	23
	TF 2012 & 2013 - TF 2014	0.863	3.968	50

PC: Principal Components used.

Interpreting the results in Table 18, positive indicators for a model with capacity to generalize from a set of training examples to a testing set with different vintages of wine grape berries are shown: high correlation between the predictions and ground-truth results is achieved, as stated by the R^2 parameter, but there's a rather significant increase on the error measures for both setups, especially for the generalization set with the most different vintages employed on the training set – this might be explained by the results of the ANOVA tests mentioned in 3.1 (see Appendix I), since pretty much all datasets have significant differences in the means in comparison to the remaining vintages and varieties, indicating that these are populations with rather different patterns in the spectra to capture in the learning process, which in turn makes the generalization step harder to carry without an increase on the uncertainty of the predictions (despite the increase in the number of samples up for analysis) – nevertheless, the model shows accurate predictions; similarly to the analysis made for the pH index results, it's observable that the number of principal components increases with the variability of the data used (that is, the larger the number of samples, the more principal components are chosen) and that increase may be crucial to the model's adaptability to the differences in the variance of the datasets allowing for more stable predictions; the descriptive statistics of the independent test sets (see Appendix

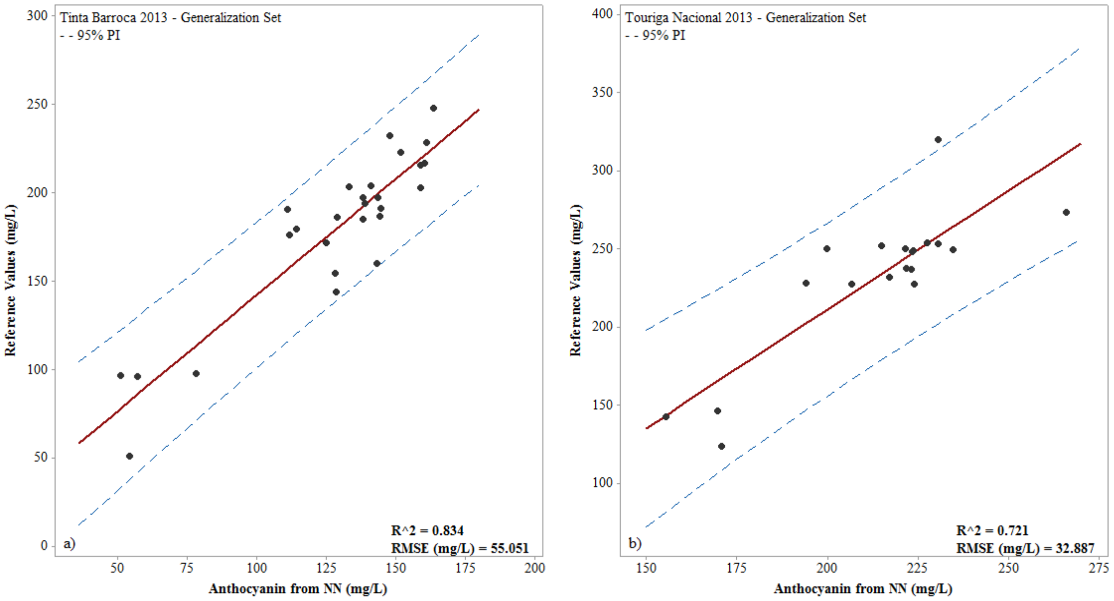
K) show that, similarly to the other oenological parameters, the mean and standard deviation values fit the initial 95% confidence intervals determined for the TF 2013 and TF 2014 samples (on Table 4), which means that the independent test set is a good representation of the overall populations; analysing the residuals vs fit values plots (Appendix N and Appendix O), despite some outliers, both plots have the residuals evenly symmetrically distributed and clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable, which aids validating the model's predictions.

Regarding the comparison with similar results in literature, Gomes *et al.* (2017b) had the best results for predictions on different vintages of wine grape berries with both, a machine learning algorithm (also NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344º Brix, respectively for the mentioned models: in this work, a similar R^2 value (R^2 of 0.913) was obtained for the test set composed of TF 2013 samples when in comparison to the NN model, with a rather significant increase on the error measure; however, for both test setups, the results are worse than those published when obtained by the PLS model, making this the first case of analysis in this work where a chemometric method obtained fairly superior results to a machine learning algorithm.

Overall, the NNs model achieved superior results for the prediction of anthocyanin concentration and pH index in comparison to the state of the art approaches, but inferior results for the prediction of sugar content when compared to a chemometric method (PLS regression). In spite of increasing the number of principal components used as input and a small detriment on performance, the model was still able to capture most of the relations between the data and achieve accurate predictions for different vintages, evidencing a good generalization capacity – consequently, changes to the NN topology (as discussed in 4.2.1) might not be necessary, but in the future some experiments should be performed with different topologies for the prediction of sugar content, to assess if the model can obtain similar or superior results when compared to the PLS regression results published in literature.

4.2.2.2. Different Vintages and Varieties

Graph 6 shows the results obtained by the NNs model for the prediction of anthocyanin concentration on different varieties and vintages of wine grape berries. Since there aren't any laboratorial results for the TF 2014 dataset, only the TF 2012 and TF 2013 set of samples compose the training and validation sets. The number of principal components used was 5 and 19 for the TB 2013 and TN 2013 generalization sets, respectively.

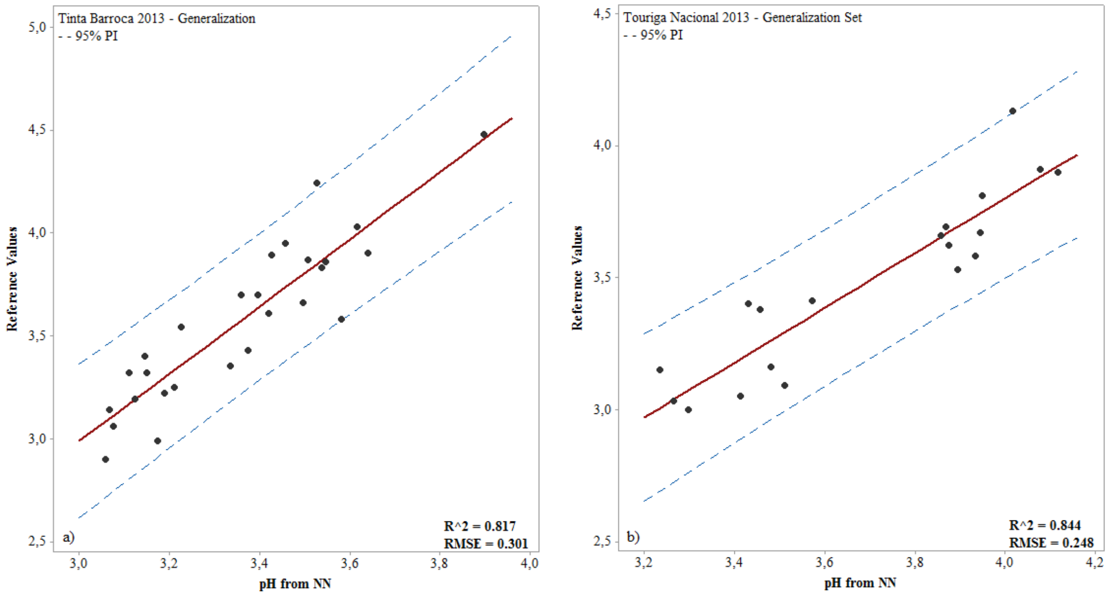


Graph 6 – Results for the estimation of anthocyanin concentration on different vintages and varieties with NNs; a) TB 2013 generalization set; b) TN 2013 generalization set

Observing Graph 6, the decrease in the accuracy of the NNs predictions is clear: the error measure suffers from a large increase (from RMSE on average between 15-25 mg.L^{-1} it goes as high as 55.051 and 32.887 mg.L^{-1} for the TB 2013 and TN 2013 datasets, respectively) and the R^2 values, naturally, decay as well (from R^2 usually above 0.90, the model obtained 0.834 and 0.721 for the TB 2013 and TN 2013 set of samples, respectively) - a decrease on the model's performance was expected, but what is important is to assess if the prediction algorithm can handle the variations in the grapes' oenological patterns that are known to occur between years and varieties, or the models will become more complex since it will be required a new model to be used in every different year or for every different variety (and that is exactly what needs to be avoided) – analysing exclusively the R^2 and RMSE parameters and acknowledging that a decrease in performance is always expected for a test setup of this nature, one can consider that the NNs achieved a rather accurate set of predictions, but the uncertainty of the predictions might be too high for these to be considered good indicators of the NNs generalization capacity; however, the results of the ANOVA tests mentioned in 3.1 (see Appendix C) pointed out that there are significant differences in the means between the TF 2012 and TB 2013 samples and the TF 2013 and TN 2013 datasets, which aids providing an explanation to the increase on the error measures: the populations in the training set have rather different patterns in the spectra to capture in the learning process than those on the test set, making the generalization step harder to carry without an increase on the uncertainty of the predictions; additionally, and contrary to the indicators on the generalization sets with different vintages, for these sets the number of principal components used was significantly smaller which means that the NNs couldn't find important information about the data on the remaining factors determined by the PCA; the descriptive statistics of the independent test sets (see Appendix K) show that the mean and standard deviation values of the TB 2013 dataset fit the initial 95% confidence intervals determined for the overall population (on Table 2): for the TN 2013 set of samples, the mean fits the initial 95% confidence interval but the standard deviation is over the higher limit, which might indicate that the test set with TN 2013 samples is not a good representation of the overall population; analysing the residuals vs fit value plots (Appendix P and Appendix Q), the TB 2013 plot exhibits slight indicators of heteroscedasticity, meaning that the residuals get larger as the predictions move from small to large - heteroscedasticity usually indicates that either an input variable is missing (that is, the model needs more information to be able to identify the patterns in the training set) or, in the most frequent cases, that a transformation to one of the input variables is necessary (because regression models usually work better with variables that have a symmetrical or bell-shaped distribution, it's common to find input variables with an asymmetrical distribution and apply, e.g., a log transform). As for the TN 2013 plot, despite some outliers and a small indicator that the y-axis is unbalanced, one can consider that the residuals are symmetrically distributed and that they don't follow any specific pattern.

Comparing these results with those published in literature for predicting anthocyanin concentration on different vintages and varieties of wine grape berries, Janik *et al.* (2007) has the best (and only) results with a R^2 of 0.900 for his NNs model with PLS scores as input, with a non-comparable error measure: these results are rather superior to the ones obtained in this work (R^2 of 0.834 and 0.721 for the TB 2013 and TN 2013 datasets, respectively) but it's important to mention that this author used a much higher number of samples for the training and validation sets (3134 samples obtained from 4 different vintages and 9 different varieties, while in this work there are 332 samples from 2 different vintages) and the test sets (250 samples from 1 vintage and 9 different varieties, while in this work there are 27 and 19 samples from 1 vintage and 1 variety for the TB 2013 and TN 2013 datasets, respectively) – additionally, for Janik *et al.* (2007) all the different varieties of wine grape berries on the test sets are also a part of the training and validation sets (contrary to this work, in which the TB and TN varieties are only present on the test sets), so it can be considered that the model only generalizes for different vintages since the training and validation sets are composed by samples from 1999 to 2003 of 9 varieties and the test set has the same 9 varieties but only for the vintage year of 2004 – if a comparison is made with the results presented in this work for generalization on different vintages (see 4.2.2.1), one can see that the results shown are superior with significantly fewer samples from fewer harvest years.

Graph 7 presents the results for the prediction of pH index on different vintages and varieties of wine grape berries by the NNs model. The number of principal components used was 7 and 32 for the TB 2013 and TN 2013 datasets, respectively.

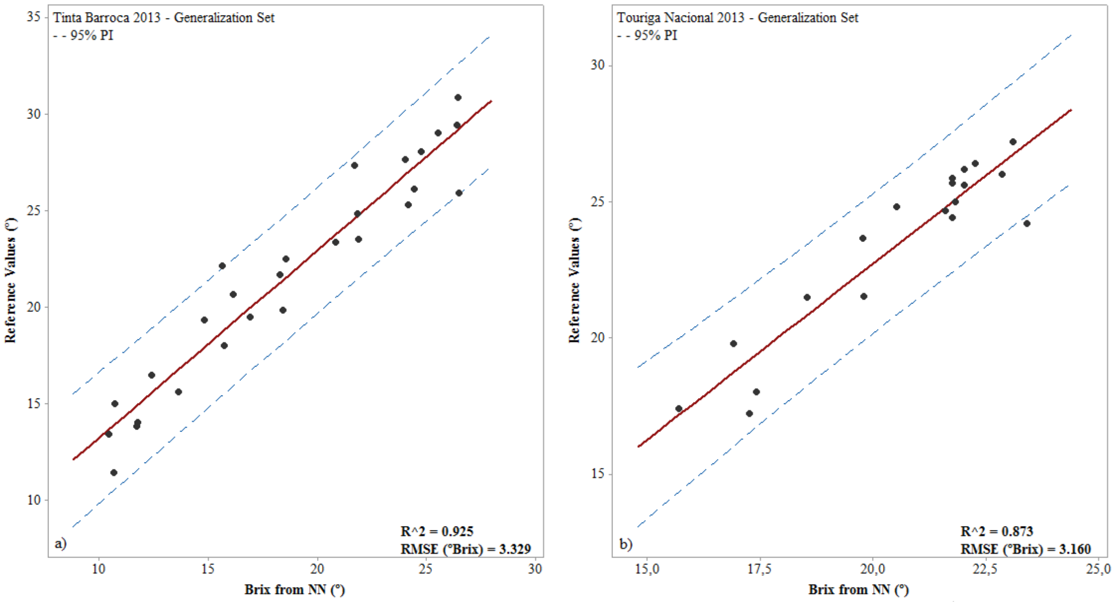


Graph 7 – Results for the estimation of pH index on different vintages and varieties with NNs; a) TB 2013 generalization set; b) TN 2013 generalization set

Inspecting Graph 7 it's clear that the NNs model obtained very satisfactory results: the error measures are comparable to the ones obtained for the single variety and vintage models and the R^2 values are even superior to some of those obtained in 4.2.1. There wasn't a significant decrease on the model's performance when comparing these results with those obtained for the single variety and vintage models, which gives positive indicators of the model's generalization capacity; the results of the ANOVA tests mentioned in 3.1 (see Appendix F) noted significant differences in the means between the TF 2014 dataset and the TF 2012 and TF 2013 samples, which means that the populations in the training set had rather different patterns in the spectra to be captured in the learning process, but the model was able to overcome this difficulty without adding significantly to the uncertainty of the predictions; as seen previously, the number of principal components used grew to allow for more information to be gathered and to achieve more stable predictions, but only for the TN 2013 dataset – for the TB 2013 samples the number of principal components used was rather small, but it might be a consequence of choosing the independent test set samples randomly; however, the descriptive statistics of the independent test sets (see Appendix K) show that the standard deviation value of the TB 2013 dataset doesn’t fit the initial 95% confidence interval (on Table 3), which might indicate that this test set is not a good representation of the overall population; analysing the residuals vs fit values plots (Appendix P and Appendix Q), despite some outliers, both plots seem to have the residuals following an evenly symmetrical distribution clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable.

Regarding the comparison with the current literature, there aren’t any works published that predict pH index on different varieties and vintages of wine grape berries: comparing the results with authors that employed only different vintages on the test sets (as seen in 4.2.2.1), Fadock *et al.* (2016) obtained a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model – despite the fact that the test sets compose not only different vintages but also different varieties of wine grape berries, the results published in this work (R^2 of 0.817 and 0.844, RMSE of 0.301 and 0.248 for the TB 2013 and TN 2013 datasets, respectively) can be considered significantly better.

Graph 8 shows the results for the estimation of sugar content on different vintages and varieties of wine grape berries by the NNs model. The number of principal components used was 45 and 15 for the TB 2013 and TN 2013 datasets, respectively.



Graph 8 – Results for the estimation of sugar content on different vintages and varieties with NNs; a) TB 2013 generalization set; b) TN 2013 generalization set

Examining Graph 8, similarly to the pH index analysis, it's clear that the NNs model achieves very positive indicators regarding its generalization capacity: the error measures had a rather significant increase when compared to the single variety and vintage models, but the R^2 values obtained are even superior to some of those present in 4.2.1. Comparing these results with those obtained for training, validation and testing exclusively on the TB 2013 and TN 2013 varieties one can see that this model actually had a more adequate fit with training and validation on different varieties and vintages of wine grape berries, which praises the model's generalization ability; the results of the ANOVA tests mentioned in 3.1 (see Appendix I) noted significant differences in the means between almost every single variety and vintage, providing somewhat of an explanation to the increase on the degree of uncertainty, but the model was able to overcome this difficulty providing an accurate set of predictions; similarly to the previous analysis, the number of principal components used grew, but only for the TB 2013 dataset – for the TN 2013 samples the number of principal components used was small, but once again, it might be a consequence of choosing the independent test samples randomly; however, the descriptive statistics of the independent test sets (see Appendix K) show that the standard deviations values for both datasets don't fit the initial 95% confidence intervals (on Table 4), which might indicate that these test sets are not a good representation of the overall population; analysing the residuals vs fit values plots (Appendix P and Appendix Q), despite some outliers, both plots seem to have the residuals following an evenly symmetrical distribution clustering towards the middle of the plot and towards the lower digits of the y-axis, with no clear patterns identifiable.

As for the comparison with current literature, similarly to the pH index analysis, there aren't any works published that predict sugar content on different varieties and vintages of wine grape berries, not allowing for a direct comparison to be made: considering the results for authors that employed only different vintages on the test sets (as seen in 4.2.2.1), Gomes *et al.* (2017b) had the best results with both, a machine learning algorithm (also NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344º Brix, respectively for the mentioned models - in this work, despite having test sets composed not only of different vintages but also of different varieties of wine grape berries, for the TB 2013 dataset a superior fit (R^2 of 0.925) was found when compared to the author's NNs model but (naturally) with inferior error measures.

Overall, despite a slight drop on the models' performance on the generalization sets and a higher error rate in the predictions (which can be considered as a reasonable outcome due to the fact that these varieties and vintages can't be found on the training steps, increasing the uncertainty), these results are very good indicators in respect to the NNs generalization capacity, since they indicate it might not be necessary to build models who require a yearly update of samples, or new samples for each variety: however, further tests with different topologies can be made for the prediction of anthocyanin concentration, where the drop on the quality of the fits and increase on the error measures was actually rather significant.

4.3. Decision Trees

4.3.1. Test Sets

The validation and test set results obtained by the DTs model (one for each variety and vintage) for the prediction of anthocyanin concentration are presented in Table 19. As mentioned in 3.1, the TF variety on the vintage year of 2014 doesn't have any laboratory results available, preventing the development of a model for that particular set of samples.

Table 19 – Results for the determination of anthocyanin concentration on the test sets using DTs

		Validation Set		Test Set		
		R^2	RMSE(mg.L^{-1})	R^2	RMSE (mg.L^{-1})	PC
Anthocyanin Concentration	TF 2012	0.811	25.667	0.942	20.964	9
	TF 2013	0.701	33.393	0.921	44.788	1
	TB 2013	0.623	28.556	0.916	31.811	14
	TN 2013	0.605	21.748	0.872	32.054	4

PC: Principal Components used.

Observing the results one can conclude that the DTs model shows good fits but with a big error rate for the test sets when compared to the NNs models presented in 4.2.1. The R^2 and RMSE values on the validation sets might indicate that it suffers from a certain degree of underfitting for the TF 2013, TB 2013 and TN 2013 set of samples, since it has poor results on the training/validation step, but accurate predictions on the test set (despite the high degree of uncertainty); similarly to the analysis made for the NNs model, the model accentuates the difficulties in having a quality training step and predictions for the datasets with the least standard deviations and the smallest range of values in their populations, namely the TB 2013 and TN 2013 datasets (Table 2); for the case of the TN 2013 dataset, the decrease in the model's performance can also be explained by the small number of samples (only 60 samples, the smallest dataset used); as for the number of principal components used, it's observable that a rather small number was chosen for all datasets (except for the TB 2013 set of samples, but that might be due to the fact that the samples for the independent test set are chosen at random) which indicates that the difficulties in learning can't be overcome by adding more principal components as input (probably because they mainly comprise noise).

Comparing the results with similar works from the literature that train and predict on the same varieties and vintages of wine grape berries, Fernandes *et al.* (2015) had the best results while using a machine learning algorithm (NNs), obtaining a R^2 of 0.950 and a RMSE of 14.000 mg.L^{-1}: for the DTs model, despite having a test set with a similar quality of fit with a higher error measure (R^2 of 0.942 and RMSE of 20.964 mg.L^{-1} for the TF 2012 dataset), overall all test sets obtain worse results when compared not only with this author, but also with the results presented for NNs model in 4.1; as for the authors using chemometric methods, Le Moigne *et al.* (2008) achieved the best results with his PLS regression model, with a R^2 of 0.979 (the errors are not comparable) for the training step, but no results were available for test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), also with a PLS regression model, with a R^2 of 0.650 and a RMSE of 75.000 mg.L^{-1}: in this work, the DTs model had superior results for the both the R^2 and the RMSE on all the test sets.

The validation and test set results achieved with the DTs model (again, one for each variety and vintage) for the determination of pH index are shown in Table 20.

Table 20 – Results for the determination of pH index on the test sets using DTs

		Validation Set		Test Set		
		R^2	RMSE	R^2	RMSE	PC
pH Index	TF 2012	0.735	0.179	0.838	0.181	13
	TF 2013	0.524	0.248	0.839	0.252	17
	TF 2014	0.619	0.172	0.869	0.159	14
	TB 2013	0.558	0.206	0.721	0.344	14
	TN 2013	0.699	0.186	0.888	0.202	17

PC: Principal Components used.

Analysing the results, and similarly to the analysis in 4.1, the model slightly underperforms when predicting the pH index (in comparison to the anthocyanin concentration) but the decrease in performance is not so clear as it was in the NNs models (perfectly normal, since the anthocyanin concentration results for the DTs were worse than the ones obtained by the NNs): as mentioned previously, this could be due to the fact that the datasets have very small values for the standard deviations (close to 0) and a small range of values for the prediction intervals (Table 3); another possible explanation (that was addressed and scrutinized in the NNs model analysis) is that the greatest variation in the pH patterns is reflected on the model's training step, which has difficulties to capture such relationships in the data without a greater number of samples for all datasets. Besides that, there's an additional challenge in measuring the pH on wine grape berries, since the acidity is sensible to small changes in the condition of the sample; another indicator seen in the anthocyanin concentration results (and in the pH index results for the NNs models) can be recognized, namely the model underfitting on some of the vintages and varieties (for the pH index results, this is more noticeable on the TF 2013, TF 2014 and TB 2013 datasets); regarding the number of principal components used, a difference can be spotted – contrary to the NNs models, which only used a bigger number of principal components as input when different vintages composed the training and validation sets and the number of samples grew, the DTs models started using a bigger number of principal components as input on average for the prediction of pH index in single vintage models – this might indicate that the DTs model needs more information from its inputs to ease the learning process and that the number of factors on the PCA may not necessarily be composed by noise after the eigenvalues are over 1, but instead contain important information to capture the patterns in the spectra.

Comparing the results with those published in literature for training and prediction on the same varieties and vintages of wine grape berries, Cao *et al.* (2010) had the best results using a non-chemometric method (genetic algorithm) with a R^2 of 0.957 and a RMSE of 0.126 for the training step, but no results can be found for the test sets – the highest values obtained that allow a direct comparison (test set values) are those obtained by Fernandes *et al.* (2015) with his NNs model, with a R^2 of 0.730 and a RMSE of 0.180: for four out of the five test sets used in this work for the DTs model, superior results for the R^2 were achieved (R^2 of 0.838, 0.839, 0.869 and 0.888 for the TF 2012, TF 2013, TF 2014 and TN 2013 test sets respectively), while for the RMSE the values obtained are better for only one of the five test sets (TF 2014, with RMSE of 0.159) and higher for the remainder; as for the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.940 and a RMSE of 0.120 for the training step, but there aren’t results available for the test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), with a PLS regression model, with a R^2 of 0.810 and a RMSE of 0.050: in this investigation, superior results for the DTs model were obtained for the R^2 of four out of the five test sets (TF 2012, TF 2013, TF 2014 and TN 2013, with a R^2 of 0.838, 0.839, 0.869 and 0.888, respectively) but the RMSE values are inferior for all experiments: therefore, as mentioned in 4.1, it’s implied that the difficulties in building a prediction model for the pH index are transversal to the other works published in literature. Performing a comparison with the results presented in this work with the NNs models, it’s noticeable that the DTs models achieve a slightly better quality of fit for the predictions, but with higher error measures – both models achieve very similar results.

The validation and test set results obtained with the DTs model (one for each variety and vintage) for the estimation of sugar content are presented in Table 21.

Table 21 – Results for the determination of sugar content on the test sets using DTs

		Validation Set		Test Set		
		R^2	RMSE(ºBrix)	R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012	0.857	1.315	0.928	1.488	16
	TF 2013	0.775	1.980	0.930	2.870	14
	TF 2014	0.608	2.419	0.893	2.432	16
	TB 2013	0.680	2.692	0.904	3.478	20
	TN 2013	0.711	1.333	0.870	1.979	18

PC: Principal Components used.

Examining Table 21 one can see that the set of predictions are accurate, with good values for the R^2 on all test sets, but when compared to the results obtained in 4.1 with the NNs models, an increase on the RMSE is noted: as it was stated primarily on the anthocyanin concentration results analysis, the model suffers from a certain degree of underfitting on some of the vintages and varieties (in this case, it’s more noticeable on the TF 2014, TB 2013 and TN 2013 datasets); regarding the principal components, and similarly to what was seen on the pH index results analysis, the DTs models use a higher number of principal components as input aiding the assumption that the models need more information from its inputs to ease the learning process; for the case of the TN 2013 dataset, the decrease on the model’s accuracy can also be explained by the fact that this dataset has a very small number of samples (at 60, is the dataset with the smallest number) with small standard deviation between samples and a small range of values for the prediction interval; despite obtaining inferior results when compared to the ones obtained by the NNs models, the model’s capacity to achieve accurate predictions for all datasets with the same structure should be highlighted, since the ANOVA tests (as seen in Appendix I) showed that there are significant differences in the means between almost every set of samples.

Hence, comparing the results with the ones published in literature for training and prediction on the same varieties and vintages of wine grape berries, Gomes *et al.* (2014a) had the best results using a machine learning algorithm (NNs) with a R^2 of 0.959 and a RMSE of 1.026 ºBrix: in this work, for the DTs models, the results obtained for all test sets are somewhat inferior; regarding the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.990 and a RMSE of 1.370 ºBrix for the training step, but the results for the test sets can’t be found – considering authors that allow a direct comparison of the results (test set results), Gomes *et al.* (2014b) with a PLS regression model obtained the best values for R^2 and RMSE, with 0.948 and 0.939 ºBrix, respectively: in this work, the values obtained for the R^2 and RMSE are slightly inferior for all datasets.

Overall, the DTs models achieved similar results for the prediction of all oenological parameters in comparison to the state of the art approaches: the pH index results are rather superior than those published by other authors, but for the anthocyanin concentration and sugar content the results obtained are somewhat inferior. When compared to the results obtained by the NNs models in 4.1, the DTs models had comparable results for the prediction of the pH index, but the performance was also worse when operating on the anthocyanin concentration and sugar content. As in the analysis made in 4.1, despite suffering from underfitting for some of the vintages and varieties, changes to the DTs structure will only be considered after analysis of the generalization results, since the main goal is to obtain a model that can give accurate predictions for different vintages and varieties without fine tuning the DTs structure for all datasets available.

4.3.2. Model Generalization

As mentioned in 4.2.2, in order to study the models' generalization capacity two different experiments were used: the first, that applies a different vintage of the same varieties that compose the training and validation sets to the test sets (since the only variety that contains different vintage years is the TF, train and validation will occur on one or more vintages of TF and the test set will be composed of samples from the next vintage year); the second, that employs a different variety and vintage on the test set (in this case, all the TF vintage years will be used on the training and validation sets, while the test sets will be composed by the TB or TN samples).

In these experiments only the test set results are presented since it's not of major importance to understand how good the models' fit is early in the training process: it will have to generalize to distinct samples on the test set).

4.3.2.1. Different Vintages

The test set results obtained by the DTs model for the prediction of anthocyanin concentration on different vintages are presented in Table 22. As mentioned previously, since the TF variety on the vintage year of 2014 doesn't have any laboratory results available, a model composed of TF 2012 and 2013 samples to predict TF 2014 values on the test set couldn't be built.

Table 22 – Results for the prediction of anthocyanin concentration on different vintages with DTs

		Test Set		
		R^2	RMSE (mg.L^{-1})	PC
Anthocyanin Concentration	TF 2012 - TF 2013	0.916	45.034	20

PC: Principal Components used.

Observing the results presented, positive indicators for a model with capacity to learn from wine grapes of different vintages are shown: there is a high correlation between the predictions and ground-truth results, as stated in the R^2 parameter, with only a small decrease of performance when compared to the single variety and vintage models, and the error measure is quite similar to the one obtained in the single variety and vintage model. However, the degree of uncertainty for the predictions is still rather high when compared to the values obtained for the NNs model in 4.2.2.1; the number of principal components used as input is higher when compared to the number used for the single variety and vintage models, indicating that there is a necessity of increasing the percentage of variance explained by the PCA to obtain a more stable set of predictions; the descriptive statistics of the independent test set (see Appendix L) show that the mean value fits the initial 95% confidence interval determined for the TF 2013 samples (on Table 2) but the standard deviation value is over the higher limit of the confidence interval, which might indicate that the independent test set isn't a good representation of the overall population; analysing the residuals vs fit values plot (Appendix R), one can find some high valued residuals, outliers and slight indicators of heteroscedasticity, meaning that (as seen previously) the residuals get larger as the predictions move from small to large – heteroscedasticity usually indicates that either an input variable is missing (that is, the model needs more information to be able to identify the patterns in the training set) or, in the most frequent cases, that a transformation to one of the input variables is necessary (because regression models usually work better with variables that have a symmetrical or bell-shaped distribution, it's common to find input variables with an asymmetrical distribution and apply, e.g., a log transform).

Analysing the results published in literature it was found that there isn't any work attempting to predict different vintages of wine grape berries on training and testing for the anthocyanin concentration; Janik *et al.* (2007) used not only different vintages but also different varieties on the test set, so a more adequate comparison will be made further in this chapter; comparing the results with the single vintage models, Chen *et al.* (2015), Ferrer-Gallego *et al.* (2011) and Le Moigne *et al.* (2008) had superior R^2 values (R^2 of 0.941, 0.970 and 0.979, respectively) with their PLS (and variants) regression models on the training step, but no results were published for the test set: Fernandes *et al.* (2015) is the only work with test set results published in which the R^2 and RMSE values are superior than those presented, with a R^2 of 0.950 (the R^2 shown is 0.916) and a RMSE of 14.000 mg.L^{-1} (the RMSE shown is 45.034 mg.L^{-1}), applying a NNs model – however, considering that the results presented in Table 22 have different vintages applied on the testing phase, the small decrease on the quality of the fit can be considered acceptable, but the error measure deserves attention, since it is a significantly high increase when compared to the author's values. As it was already mentioned, when comparing these results with the ones obtained by the NNs model in 4.2.2.1, the quality of the fit is very similar but the RMSE obtained is much worse (it doubles the value), and if these increases in the degrees of uncertainty remain for the other oenological parameters, a change in the DTs structure should be considered.

The test set results obtained by the DTs model for the prediction of pH index on different vintages are presented in Table 23.

Table 23 – Results for the prediction of pH index on different vintages with DTs

		Test Set		
		R^2	RMSE	PC
pH Index	TF 2012 - TF 2013	0.831	0.226	12
	TF 2012 & 2013 - TF 2014	0.682	0.194	29

PC: Principal Components used.

Examining Table 23, it's noticeable that the model underperforms when compared to the results obtained for the anthocyanin concentration and one detail arises that is of foremost importance: contrary to what happened for the NNs models in 4.2.2.1, despite obtaining similar results for the prediction of the TF 2013 dataset values, the results obtained for the prediction on the TF 2014 samples suffer from a massive drop in the quality of the fit and not even a significant increase on the number of principal components used as input helped ease the learning process, which reinforces the idea that changes to the DTs structure must be made since the generalization capacity is being highly affected. The descriptive statistics of the independent test sets (see Appendix L) show that the mean and standard deviation values fit the initial 95% confidence intervals determined for both the TF 2013 and TF 2014 samples (on Table 3), meaning that the independent test sets are a good representation of the overall populations; analysing the residuals vs fit values plots (Appendix R and Appendix S), one can see that despite finding some outliers and a slightly unbalanced y-axis, both plots have the residuals symmetrically distributed and clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable.

Comparing these results with the ones published in literature for different vintages of wine grape berries employed on the training and testing phases, Fadock *et al.* (2016) is the only work published that meets this particular experimental outline, obtaining a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model: for both test sets in this work, and despite a huge drop on the performance when predicting for the TF 2014 dataset, the R^2 values are superior (R^2 of 0.831 and 0.682 for the TF 2013 and TF 2014 test sets, respectively) but the RMSE on both setups is inferior (RMSE of 0.226 and 0.194 for the same test sets, respectively): however, as stated in 4.2.2.1, regarding Fadock *et al.* (2016) results, it's rather questionable that a model that obtained such a low R^2 score has simultaneously such a low error measure, especially since it shares the same value than that obtained for the test set with only one vintage and variety employed on the training and testing phases (as seen in Table 1). As it was already mentioned, comparing the results with the ones presented for the single models in 4.2, there is a significant decrease on the performance when measured against the results obtained for the TF 2014 dataset, but the result obtained for the test set of TF 2013 samples after training on the TF 2012 dataset (R^2 of 0.831and RMSE of 0.226) is still superior when compared to the best test set results for single vintage and variety models previously published in literature [Fadock et al (2016), with a PLS regression model obtaining a R^2 of 0.807 and RMSE of 0.050]; comparing these results with the ones obtained by the NNs models in 4.2.2.1 show that the DTs don't have the same generalization capacity, and changes to the DTs structure should be studied.

Table 24 presents the results for the prediction of sugar content on different vintages obtained by the DTs model.

Table 24 – Results for the prediction of sugar content on different vintages with DTs

		Test Set		
		R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012 - TF 2013	0.879	2.304	8
	TF 2012 & 2013 - TF 2014	0.625	4.307	44

PC: Principal Components used.

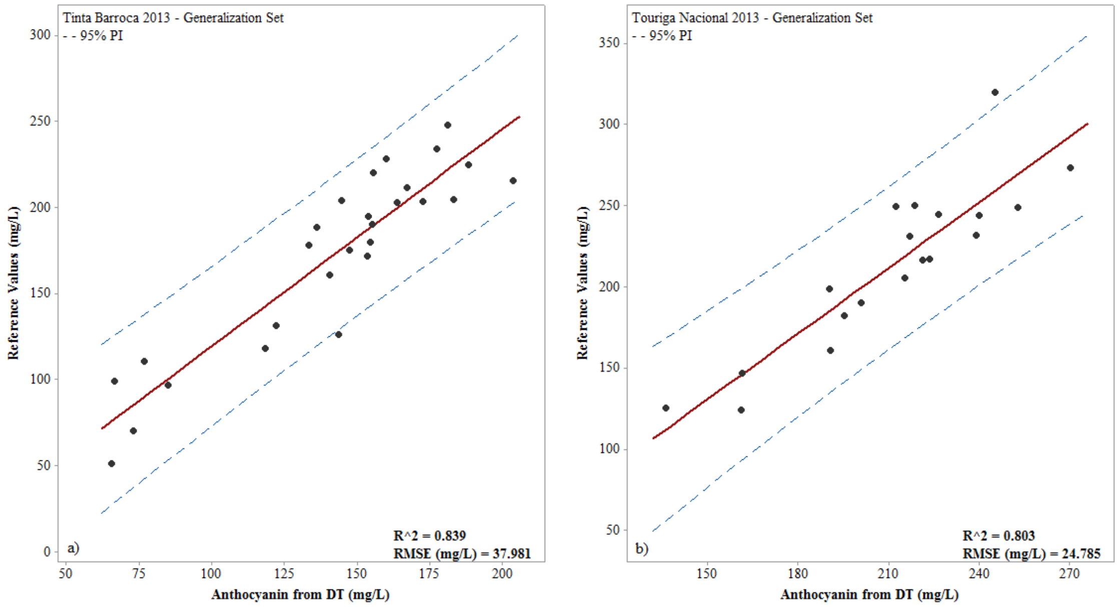
Interpreting the results in Table 18, some negative indicators regarding the model's capacity to generalize from a set of training examples to a testing set with different vintages of wine grape berries are shown: the correlation between the predictions and ground-truth results suffers a rather significant drop for the TF 2014 samples, with an important increase on the error measure – this drop can be eased by the results of the ANOVA tests mentioned in 3.1 (see Appendix I), since pretty much all datasets have significant differences in the means in comparison to the remaining vintages and varieties, indicating that these are populations with rather different patterns in the spectra to capture in the learning process, which in turn makes the generalization step harder to carry without an increase on the uncertainty of the predictions (despite the increase in the number of samples up for analysis) – nevertheless, the model shows significantly worse results in the generalization set composed of TF 2014 samples, despite using a much higher number of principal components as input, meaning that this increase couldn't help the model adapt to the differences in the variance of the datasets; the descriptive statistics of the independent test sets (see Appendix L) show that the mean and standard deviation values for the TF 2013 samples and the standard deviation value for the TF 2014 datasets don't fit the initial 95% confidence intervals, which means that both independent test sets might not be a good representation of the overall populations; analysing the residuals vs fit values plots (Appendix R and Appendix S), despite some outliers (for the case of the TF 2014 datasets, it's quite a higher number), both plots have the residuals pretty evenly symmetrically distributed and clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable.

Regarding the comparison with similar results in literature, Gomes *et al.* (2017b) had the best results for predictions on different vintages of wine grape berries with both, a machine learning algorithm (NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344 ºBrix, respectively for the mentioned models: in this work, the R^2 and RMSE values obtained by the DTs models are significantly inferior, not only comparing with the aforementioned author, but also with the results presented in 4.2.2.1 for the NNs models.

Overall, the DTs models achieved comparable results for the prediction of anthocyanin concentration and pH index in comparison to the state of the art approaches, but inferior results for the prediction of sugar content. When comparing the DTs models with the NNs models in 4.2.2.1, it's noticeable that the generalization capacity of the DTs is significantly worse, with high degrees of uncertainty for the prediction of most of the oenological parameters, indicating that changes to the DTs structure (as discussed in 4.3.1) might be necessary since the model is underfitting when extending his predictions for different vintages. Further discussion on the DTs structure will be made after analysing the results in 4.3.2.2.

4.3.2.2. Different Vintages and Varieties

Graph 9 shows the results obtained by the DTs model for the prediction of anthocyanin concentration on different varieties and vintages of wine grape berries. Since there aren't any laboratorial results for the TF 2014 dataset, only the TF 2012 and TF 2013 set of samples compose the training and validation sets. The number of principal components used was 13 and 41 for the TB 2013 and TN 2013 generalization sets, respectively.

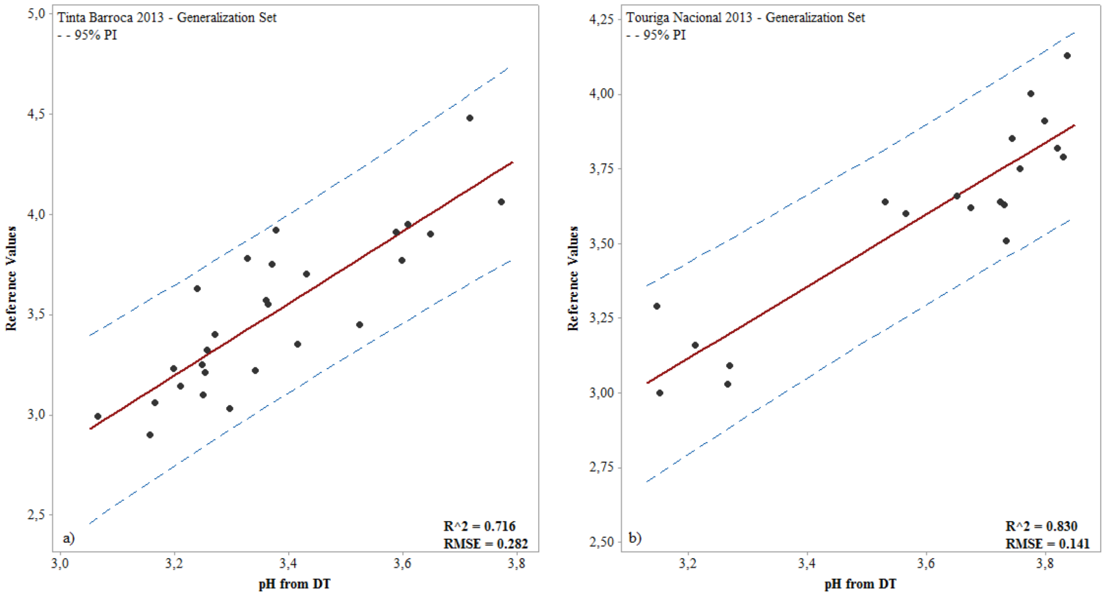


Graph 9 – Results for the estimation of anthocyanin concentration on different vintages and varieties with DTs; a) TB 2013 generalization set; b) TN 2013 generalization set

Observing Graph 9, the decrease in the accuracy of the DTs predictions isn't as clear as it was for the NNs models, since despite having high values for the RMSE, there isn't a significant increase when comparing the results with the single variety and vintage models, because the error measure was already high; the R^2 values show a slight decrease (from R^2 usually above 0.90, the model obtained 0.839 and 0.803 for the TB 2013 and TN 2013 set of samples, respectively) – analysing exclusively the R^2 and RMSE parameters and acknowledging that a decrease in performance is always expected for a test setup of this nature, one can consider that the DTs achieved a rather accurate set of predictions (contradicting the indicators in 4.3.2.1), but the uncertainty of the predictions might be too high for these to be considered good indicators of the DTs generalization capacity; contrary to what happened in the NNs models in 4.2.2.2, the number of principal components used in these generalization sets was significantly bigger when compared to the generalization sets with different vintages, meaning that the DTs found important information about the data on the remaining factors determined by the PCA (except for the TB 2013 dataset, but that might be due to the random choosing of the samples to compose the independent test sets); the descriptive statistics of the independent test sets (see Appendix L) show that the mean and standard deviation values of the TB 2013 dataset fit the initial 95% confidence interval determined for the overall population (on Table 2): for the TN 2013 set of samples, the mean fits the initial 95% confidence interval but the standard deviation is over the higher limit, which might indicate that the test set with TN 2013 samples is not a good representation of the overall population; analysing the residuals vs fit value plots (Appendix T and Appendix U), both plots exhibit indicators of the residuals being unbalanced on the y-axis: the solution to this problem is most of the times transforming the data, typically the response variable (e.g., applying a log transform).

Comparing these results with those published in literature for predicting anthocyanin concentration on different vintages and varieties of wine grape berries, Janik *et al.* (2007) has the best (and only) results with a R^2 of 0.900 for his NNs model with PLS scores as input, with a non-comparable error measure: these results are rather superior to the ones obtained in this work (R^2 of 0.839 and 0.803 for the TB 2013 and TN 2013 datasets, respectively) but it's important to note that, as mentioned previously (in 4.2.2.2) that this author used a much higher number of samples for the training and validation sets (3134 samples obtained from 4 different vintages and 9 different varieties, while in this work there are 332 samples from 2 different vintages) and the test sets (250 samples from 1 vintage and 9 different varieties, while in this work there are 27 and 19 samples from 1 vintage and 1 variety for the TB 2013 and TN 2013 datasets, respectively) – additionally, for Janik *et al.* (2007) all the different varieties of wine grape berries on the test sets are also a part of the training and validation sets (contrary to this work, in which the TB and TN varieties are only present on the test sets), so it can be considered that the model only generalizes for different vintages since the training and validation sets are composed by samples from 1999 to 2003 of 9 varieties and the test set has the same 9 varieties but only for the vintage year of 2004 – if a comparison is made with the results presented in this work for generalization on different vintages (see 4.3.2.1), one can see that the results shown are superior with significantly fewer samples from fewer harvest years. Regarding a comparison with the generalization set results obtained by the NNs model in 4.2.2.2, the DTs models actually obtained superior results, contradicting the prior belief in 4.3.2.1 that the model's generalization capacity was inferior to the one achieved by the NNs: howsoever, changes to the DTs structure should still be considered since these results might constitute a statistical anomaly.

Graph 10 presents the results for the prediction of pH index on different vintages and varieties of wine grape berries by the DTs model. The number of principal components used was 39 and 22 for the TB 2013 and TN 2013 datasets, respectively.

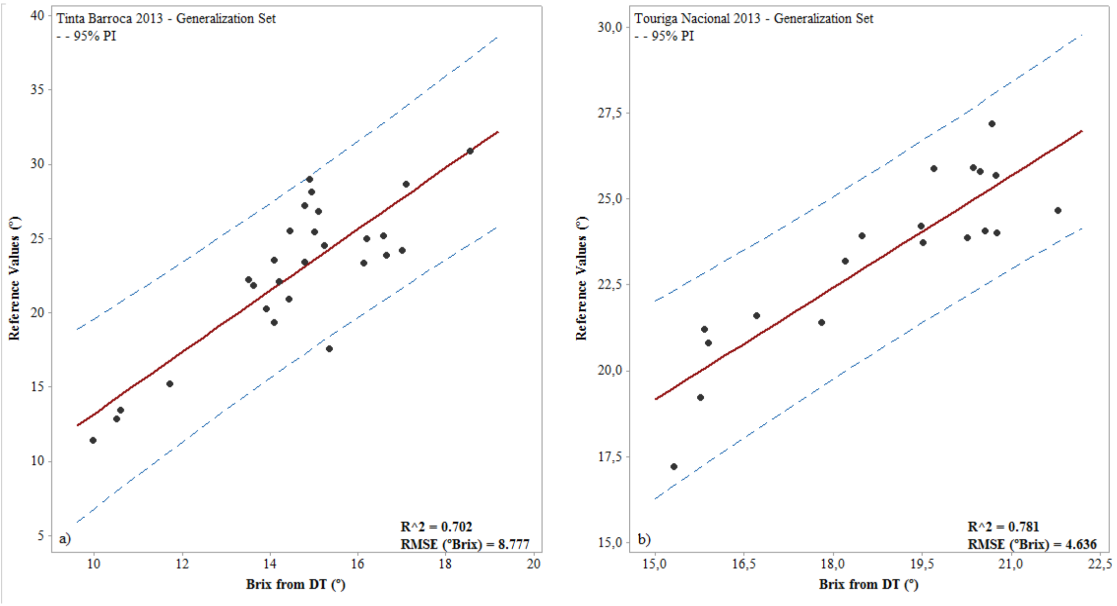


Graph 10 – Results for the estimation of pH index on different vintages and varieties with DTs; a) TB 2013 generalization set; b) TN 2013 generalization set

Inspecting Graph 10 it's clear that the DTs model obtained very satisfactory results for the TN 2013 generalization set: both the error measure and R^2 values are comparable to the ones obtained for the single variety and vintage models in 4.3.1, giving a positive indicator of the model's generalization capacity; however, for the TB 2013 dataset there was a slight decrease in performance. The results of the ANOVA tests mentioned in 3.1 (see Appendix F) noted significant differences in the means between the TF 2014 dataset and the TF 2012 and TF 2013 samples, which means that the populations in the training set had rather different patterns in the spectra to be captured in the learning process, but the model was able to overcome this difficulty without adding significantly to the uncertainty of the predictions; as seen previously, the number of principal components used grew to allow for more information to be gathered and to achieve more stable predictions; the descriptive statistics of the independent test sets (see Appendix L) show that the standard deviation value of the TB 2013 dataset doesn't fit the initial 95% confidence interval (on Table 3), which might indicate that this test set is not a good representation of the overall population; analysing the residuals vs fit values plots (Appendix T and Appendix U), both plots seem to have the residuals following an evenly symmetrical distribution clustering towards the lower single digits of the y-axis, with no clear patterns identifiable.

Regarding the comparison with the current literature, there aren't any works published that predict pH index on different varieties and vintages of wine grape berries: comparing the results with authors that employed only different vintages on the test sets (as seen in 4.3.2.1), Fadock *et al.* (2016) obtained a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model – despite the fact that the test sets compose not only different vintages but also different varieties of wine grape berries, the results published in this work (R^2 of 0.716 and 0.830, RMSE of 0.282 and 0.141 for the TB 2013 and TN 2013 datasets, respectively) can be considered significantly better. Comparing these results with the ones obtained by the NNs models in 4.2.2.2, for the TN 2013 generalization set the values are very similar, but the NNs perform better for the TB 2013 dataset than the DTs model.

Graph 11 shows the results for the estimation of sugar content on different vintages and varieties of wine grape berries by the DTs model. The number of principal components used was 2 and 3 for the TB 2013 and TN 2013 datasets, respectively.



Graph 11 – Results for the estimation of sugar content on different vintages and varieties with DTs; a) TB 2013 generalization set; b) TN 2013 generalization set

Examining Graph 11, contrary to the pH index analysis, it's clear that the DTs model didn't achieve positive indicators regarding its generalization capacity: the error measurements had a significant increase when compared to the single variety and vintage models (especially for the TB 2013 generalization set) and the R^2 values are very low. The results of the ANOVA tests mentioned in 3.1 (see Appendix I) noted significant differences in the means between almost every single variety and vintage, providing somewhat of an explanation to the increase on the degree of uncertainty, but the model still obtained very poor results; contrary to the previous analysis, the number of principal components used was significantly smaller, indicating that adding more principal components didn't aid the model in achieving better fits and capturing the patterns in the spectra; the descriptive statistics of the independent test sets (see Appendix L) show that the mean and standard deviation values for both datasets fit the initial 95% confidence intervals (on Table 4), indicating that these test sets are a good representation of the overall population; analysing the residuals vs fit values plots (Appendix T and Appendix U), despite some outliers and having a slightly unbalanced y-axis (especially for the TB 2013 generalization set), both plots seem to have the residuals following a symmetrical distribution clustering towards the lower digits of the y-axis, with no clear patterns identifiable.

As for the comparison with current literature, similarly to the pH index analysis, there aren't any works published that predict sugar content on different varieties and vintages of wine grape berries, not allowing for a direct comparison to be made: considering the results for authors that employed only different vintages on the test sets (as seen in 4.3.2.1), Gomes *et al.* (2017b) had the best results with both, a machine learning algorithm (NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344º Brix, respectively for the mentioned models – in this work, the results for both test sets are significantly inferior (but it's important to note that the test sets are composed not only of different vintages but also of different varieties of wine grape berries).

Overall, despite showing good indicators regarding the models' generalization capacity for the prediction of anthocyanin concentration and pH index (the results are comparable with the ones achieved by the NNs models in 4.2.2.2), the DTs models are not fully convincing concerning the lack of necessity to build models who require a yearly update of samples, or new samples for each variety: the models show a serious degree of underfitting when attempting to generalize for different vintages, and the degree of uncertainty of the predictions is always extremely high – possible solutions might be adding more individual DTs to the ensemble or even removing the k-Fold Cross-Validation step for the prediction models with DTs: since when using the bagging algorithm, bootstrap replicates of the training set are formed and used as new training sets, the model validation step might be redundant or even detrimental to the outcome, because as seen in 3.5, k-Fold Cross-Validation with k values between 10-20 reduce the variance while increasing the bias, and having a model validation step with k-Fold Cross-Validation and a Bootstrap on the machine learning algorithm might highly increase the bias and result in underfitting.

4.4. Support Vector Regression

4.4.1. Test Sets

The validation and test set results obtained by the SVR model (one for each variety and vintage) for the prediction of anthocyanin concentration are presented in Table 25. As mentioned in 3.1 (and in 4.2.1 and 4.3.1), the TF variety on the vintage year of 2014 doesn't have any laboratory results available, preventing the development of a model for that particular set of samples.

Table 25 – Results for the determination of anthocyanin concentration on the test sets using SVR

		Validation Set		Test Set		
		R^2	RMSE(mg.L^{-1})	R^2	RMSE (mg.L^{-1})	PC
Anthocyanin Concentration	TF 2012	0.840	21.397	0.968	15.683	8
	TF 2013	0.817	22.466	0.979	11.887	10
	TB 2013	0.662	24.315	0.933	22.471	4
	TN 2013	0.649	17.401	0.929	36.860	11

PC: Principal Components used.

Observing the results, it’s clear that the SVR model shows extremely accurate predictions with a relatively small error rate for the test sets, but (as it was seen in the NNs and DTs models) the R^2 and RMSE values on the validation sets might indicate that it suffers from a certain degree of underfitting for the case of the TB 2013 and TN 2013 set of samples, since it has somewhat poor results on the training/validation step, but it has good results with a small error rate on the test set; the model accentuates the difficulties in having a quality training step and predictions for the datasets with the least standard deviations and the smallest range of values in their populations, namely the TB 2013 and TN 2013 datasets (Table 2): this might show that the model has some problems in capturing the patterns in the spectra when the prediction intervals are smaller and the standard deviations have low values; for the case of the TN 2013 dataset, the slight decrease in the model’s performance can also be explained by the small number of samples (only 60 samples, the smallest dataset used); regarding the number of principal components used as input to the model, it’s visible that the number is rather small, indicating that the SVR algorithm can capture the patterns in the spectra without an increase of variance in the inputs when compared to the NNs and DTs models.

Comparing the results with similar works from the literature that train and predict on the same varieties and vintages of wine grape berries, Fernandes *et al.* (2015) had the best results while using a machine learning algorithm (NNs), obtaining a R^2 of 0.950 and a RMSE of 14.000 mg.L^{-1}: for two out of the four test sets used in the present work, superior results for the R^2 were obtained (R^2 of 0.968 and 0.979 for the TF 2012 and TF 2013 datasets, respectively), while for the RMSE the values obtained are inferior in one out of the four test sets (TF 2013, with a RMSE of 11.887 mg.L^{-1}), but higher for the remainder; as for the authors using chemometric methods, Le Moigne *et al.* (2008) achieved the best results with his PLS regression model, with a R^2 of 0.979 (the errors are not comparable) for the training step, but no results were available for test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), also with a PLS regression model, with a R^2 of 0.650 and a RMSE of 75.000 mg.L^{-1}: in this work, superior results were obtained for both the R^2 and RMSE on all the test sets; comparing these results with the ones obtained by the NNs and DTs in 4.2.1 and 4.3.1, it's noticeable that the SVR model has better results than the DTs model and similar or superior results to the NNs model in all test sets, indicating that the learning process in the SVR model worked extremely well.

The validation and test set results achieved with the SVR model (again, one for each variety and vintage) for the determination of pH index are shown in Table 14.

Table 26 – Results for the determination of pH index on the test sets using SVR

		Validation Set		Test Set		
		R^2	RMSE	R^2	RMSE	PC
pH Index	TF 2012	0.730	0.1771	0.887	0.142	15
	TF 2013	0.561	0.225	0.863	0.165	12
	TF 2014	0.703	0.139	0.889	0.123	20
	TB 2013	0.574	0.196	0.864	0.216	5
	TN 2013	0.727	0.141	0.902	0.117	17

PC: Principal Components used.

Analysing the results, similarly to what happened for the NNs and DTs models, the model underperforms when predicting the pH index in comparison to the anthocyanin concentration: this could be due to the fact that the datasets have very small values for the standard deviations (close to 0) and small range of values for the prediction intervals (Table 3); another possible explanation is that the greatest variation in the pH patterns is reflected on the model's training step, which has difficulties to capture such relationships in the data without a greater number of samples for all datasets. Besides that, there's an additional challenge in measuring the pH on wine grape berries, since the acidity is sensible to small changes in the condition of the sample; other indicators seen in the anthocyanin concentration results can be recognized, namely the model underfitting on some of the vintages and varieties (for the pH index results, this is more noticeable on the TF 2013 and TB 2013 datasets); regarding the number of principal components used as input to the model, when in comparison to the other models presented in this work it's noticeable that the SVR model has more similarities with the DTs model, using a high number of principal components in this single variety and vintage model.

Comparing the results with those published in literature for training and prediction on the same varieties and vintages of wine grape berries, Cao *et al.* (2010) had the best results using a non-chemometric method (genetic algorithm) with a R^2 of 0.957 and a RMSE of 0.126 for the training step, but no results can be found for the test sets – the highest values obtained that allow a direct comparison (test set values) are those obtained by Fernandes *et al.* (2015) with his NNs model, with a R^2 of 0.730 and a RMSE of 0.180:
for all five test sets used in this work superior results for the R^2 were achieved, while for the RMSE the values obtained were only inferior for one test set (RMSE of 0.216 for the TB 2013 dataset); as for the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.940 and a RMSE of 0.120 for the training step, but there aren't results available for the test sets – the best results available for direct comparison (test set results) are those obtained by Fadock *et al.* (2016), with a PLS regression model, with a R^2 of 0.810 and a RMSE of 0.050: in this investigation, superior results were obtained for the R^2 of all the five test sets but the RMSE values are slightly inferior for all experiments: therefore, it's implied that the difficulties in building a prediction model for the pH index are transversal to the other works published in literature; comparing these results with the ones obtained by the NNs and DTs model in this work, similarly to what happened in the anthocyanin concentration, the SVR model has superior results for pretty much all the test sets.

The validation and test set results obtained with the SVR model (one for each variety and vintage) for the estimation of sugar content are presented in Table 27.

Table 27 – Results for the determination of sugar content on the test sets using SVR

		Validation Set		Test Set		
		R^2	RMSE(ºBrix)	R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012	0.905	1.009	0.964	0.943	19
	TF 2013	0.850	1.321	0.979	1.760	16
	TF 2014	0.786	1.647	0.926	1.653	17
	TB 2013	0.817	1.853	0.962	1.368	18
	TN 2013	0.669	1.255	0.966	1.925	10

PC: Principal Components used.

Examining Table 27 one can see that the results obtained are extremely robust, with good values for the R^2 and RMSE on all test sets: nevertheless, as it was stated on the anthocyanin concentration and pH index results analysis, the model suffers from a certain degree of underfitting on some of the vintages and varieties (in this case, it's more noticeable on the TF 2014 and TB 2013 datasets); the number of principal components used was again reasonably high, indicating that the SVR model extracts important information to identify the patterns in the spectra in the remaining factors of the PCA after they cross the eigenvalue of 1, suggesting that they don't essentially comprise noise; the model's capacity to achieve accurate predictions for all datasets should be highlighted, since the ANOVA tests showed that there are significant differences in the means between almost every set of samples (as seen in Appendix I).

Hence, comparing the results with the ones published in literature for training and prediction on the same varieties and vintages of wine grape berries, Gomes *et al.* (2014a) had the best results using a machine learning algorithm (NNs) with a R^2 of 0.959 and a RMSE of 1.026 ºBrix: for four out of the five test sets the SVR model had better R^2 values (R^2 of 0.964, 0.979, 0.962, 0.966 for the TF 2012, TF 2013, TB 2013 and TN 2013 datasets, respectively), while for the RMSE the test set with TF 2012 samples had a better value (RMSE of 0.943 ºBrix) but the rest were slightly worst; regarding the authors using chemometric methods, Nogales-Bueno *et al.* (2010) had the best results with his modified PLS regression model, with a R^2 of 0.990 and a RMSE of 1.370 ºBrix for the training step, but the results for test sets can't be found – considering authors that allow a direct comparison of the results (test set results), Gomes *et al.* (2014b) with a PLS regression model obtained the best values for R^2 and RMSE, with 0.948 and 0.939 ºBrix, respectively: in this work, the SVR model once again had superior R^2 values for four out of five test sets and a better RMSE value for the TF 2012 dataset, but the remaining obtained worse error measures; comparing these results with ones obtained by the NNs and DTs models in this work, the SVR repeated the best overall performance, similarly to what happened in the anthocyanin concentration and pH index results.

Overall, the SVR model achieved superior or comparable results for the prediction of all oenological parameters in comparison to the state of the art approaches and the remaining models presented in this work. Despite suffering from underfitting for some of the vintages and varieties (mostly the ones with a small number of samples for analysis), the results presented are very satisfactory and are very positive indicators ahead for the tests on the generalization capacity.

4.4.2. Model Generalization

In order to study the models' generalization capacity, two different experiments were used (as seen in 4.2.2 and 4.3.2): the first, that applies a different vintage of the same varieties that compose the training and validation sets to the test sets (since the only variety that contains different vintage years is the TF, train and validation will occur on one or more vintages of TF and the test set will be composed of samples for the next vintage year); the second, that employs a different vintage and variety on the test set (in this case, all the TF vintage years will be used on the training and validation sets, while the test sets will be composed by the TB or TN samples).

In these experiments only the test set results are presented since it's not of major importance to understand how good the models' fit is early in the training process: it will have to generalize to distinct samples on the test set.

4.4.2.1. Different Vintages

The test set results obtained by the SVR model for the prediction of anthocyanin concentration on different vintages are presented in Table 28. As mentioned previously, since the TF variety on the vintage year of 2014 doesn't have any laboratory results available, a model composed of TF 2012 and 2013 samples to predict TF 2014 values on the test set couldn't be built.

Table 28 – Results for the prediction of anthocyanin concentration on different vintages with SVR

		Test Set		
		R^2	RMSE ($mg.L^{-1}$)	PC
Anthocyanin Concentration	TF 2012 - TF 2013	0.938	28.349	30

PC: Principal Components used.

Observing the results presented, strong indicators for a robust model with capacity to learn from wine grapes of different vintages are shown: there is a high correlation between the predictions and ground-truth results, as stated in the R^2 parameter, with only a small decrease of performance when compared to the single variety and vintage models and despite an increase in the RMSE, the value obtained is still similar to the single variety and vintage models that have a smaller number of samples (specifically, the TB 2013 and TN 2013 datasets); the number of principal components used as input grew even more, indicating that there is a necessity to increase the percentage of variance explained by the PCA to achieve stable predictions; the descriptive statistics of the independent test set (see Appendix M) show that the mean fits the initial 95% confidence interval determined for the TF 2013 samples (on Table 2) but the standard deviation value is over the higher limit, which might indicate that the independent test set isn't a good representation of the overall population; analysing the residuals vs fit values plot (Appendix V), despite some high value residuals and slight indicators of heteroscedasticity, the residuals are rather evenly distributed between the lower and higher digits of the y-axis and in general, no clear patterns can be found.

Analysing the results published in literature it was found that there isn't any work attempting to predict different vintages of wine grape berries on training and testing for the anthocyanin concentration; Janik *et al.* (2007) used not only different vintages but also different varieties on the test set, so a more adequate comparison will be made further in this chapter; comparing the results with the single vintage models, Chen *et al.* (2015), Ferrer-Gallego *et al.* (2011) and Le Moigne *et al.* (2008) had superior R^2 values (R^2 of 0.941, 0.970 and 0.979, respectively) with their PLS (and variants) regression models on the training step, but no results were published for the test set: Fernandes *et al.* (2015) is the only work with test set results published in which the R^2 and RMSE values are superior than those presented, with a R^2 of 0.950 (the R^2 shown is 0.938) and a RMSE of 14.000 mg.L^{-1} (the RMSE shown is 28.349 mg.L^{-1}), applying a NNs model – however, considering that in the results presented in Table 28 have different vintages applied on the testing phase, the small decrease in performance can be considered acceptable, as well as the values obtained can be considered very satisfactory, since they are still comparable with the remaining state of the art approaches; comparing these results with the ones obtained by the NNs and DTs models in this work, it's observable that the SVR had a better value for the R^2 but the error measure is rather inferior – overall, the results are quite similar.

The test set results obtained by the SVR model for the prediction of pH index on different vintages are presented in Table 29.

Table 29 – Results for the prediction of pH index on different vintages with SVR

		Test Set		
		R^2	RMSE	PC
pH Index	TF 2012 - TF 2013	0.833	0.236	10
	TF 2012 & 2013 - TF 2014	0.873	0.275	18

PC: Principal Components used.

Examining Table 29, once again it's noticeable that the model underperforms when compared to the results obtained for the anthocyanin concentration; however, two important details are noteworthy: the model actually obtains better results when predicting for the TF 2014 dataset than for the TF 2013 set of samples, with very similar error measures, which might indicate that the model has difficulties in capturing the relationships between the patterns in the spectra for the pH index but, by adding a greater number of samples (even if these are from different populations) it will be able to achieve better results; and also, the model took a greater number of principal components as input to operate on the TF 2014 test set, which may indicate that when the number of samples starts to grow, the number of factors on the PCA will not mainly comprise noise after the eigenvalues are over 1 but instead contain important information that can ease the model's learning step. The descriptive statistics of the independent test sets (see Appendix M) show that the mean and standard deviation values for the TF 2013 samples fit the initial 95% confidence intervals (on Table 3), but for the TF 2014 dataset the standard deviation value is over the higher limit, which might indicate that this isn't a good representation of the overall population; additionally, the ANOVA tests (in Appendix F) previously shown that the TF 2013 dataset has a significant difference in the means when compared to the TF 2012 and TF 2014 sets, which in a way praises the model's generalization capacity since it was able to predict for the TF 2014 set while learning from the TF 2012 and TF 2013 set of samples; analysing the residuals vs fit values plots (Appendix V and Appendix W), one can see that both plots have the residuals evenly symmetrically distributed and clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable, which aids the assumption that the SVR predictions for these setups are not skewed in any way or in need of adding/transforming some input variables.

Comparing these results with ones published in literature for different vintages of wine grape berries employed on the training and testing phases, Fadock *et al.* (2016) is the only work published that meets this particular experimental outline, obtaining a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model: for both test sets in this work the SVR model obtains superior R^2 values (R^2 of 0.833 and 0.873 for the TF 2013 and TF 2014 test sets, respectively) but the RMSE on both setups is inferior (RMSE of 0.236 and 0.275 for the same test sets, respectively): however, as mentioned in the critical analysis of the NNs and DTs results, the error measure is quite questionable for a model that obtained such a low R^2 score simultaneously; comparing the results with ones presented for the single models in 4.4.1, a small decrease on the performance is noticeable when measured against the results obtained, e.g., for the TN 2013 dataset, but they're superior when compared to the best test set results for single vintage and variety models previously published in literature [Fadock *et al.* (2016), with a PLS regression model obtaining a R^2 of 0.807 and a RMSE of 0.050], which indicates that the SVR model has a very powerful generalization capacity; comparing the results with the ones obtained by the NNs and DTs models in this work, once again the SVR model shows superior results (but with slightly inferior error measures).

Table 30 presents the results for the prediction of sugar content on different vintages obtained by the SVR model.

Table 30 – Results for the prediction of sugar content on different vintages with SVR

		Test Set		
		R^2	RMSE (ºBrix)	PC
Sugar Content	TF 2012 - TF 2013	0.953	0.977	41
	TF 2012 & 2013 - TF 2014	0.829	4.464	49

PC: Principal Components used.

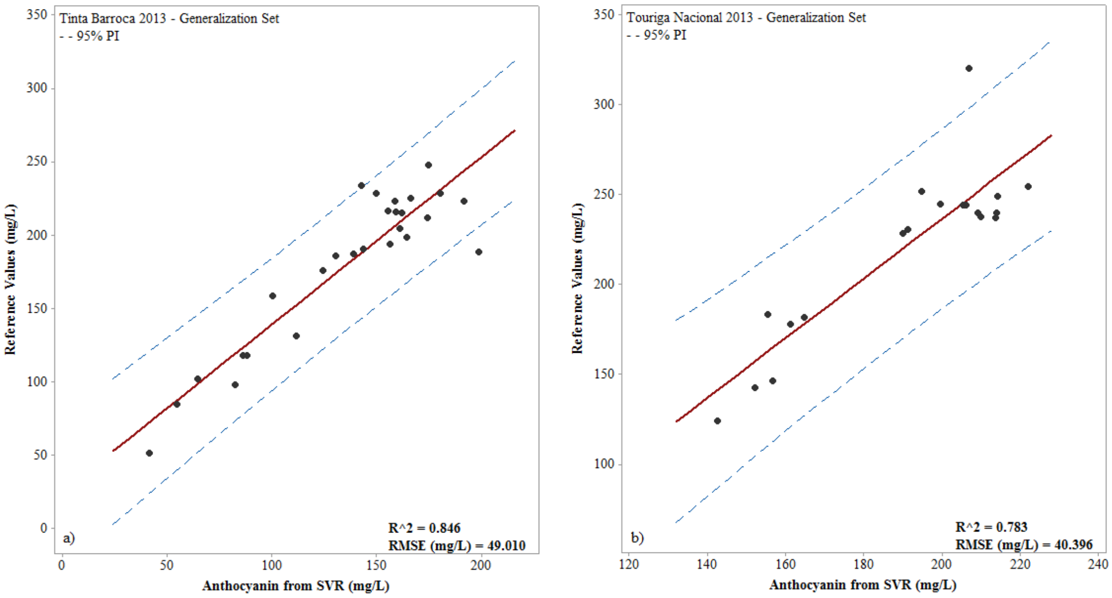
Interpreting the results in Table 30, some mixed indicators for a model with capacity to generalize from a set of training examples to a testing set with different vintages of wine grape berries are shown: there is a high correlation between the predictions and ground-truth results and a low error measure for the TF 2013 dataset but the results for the TF 2014 suffer from quite a degradation, as stated by the R^2 and RMSE parameters – this might be explained by the results of the ANOVA tests mentioned in 3.1 (see Appendix I), since pretty much all datasets have significant differences in the means in comparison to the remaining vintages and varieties, indicating that these are populations with rather different patterns in the spectra to capture in the learning process, which in turn makes the generalization step harder to carry without an increase on the uncertainty of the predictions (despite the increase in the number of samples up for analysis) – nevertheless, the model shows accurate predictions; the number of principal components used as input increases with the variability of the data used (that is, the larger the number of samples, the more principal components are chosen) and that increase may be crucial to the model's adaptability to the differences in the variance of the datasets allowing for more stable predictions; the descriptive statistics of the independent test sets (see Appendix M) show that the mean and standard deviation values fit the initial 95% confidence intervals determined for the TF 2013 and TF 2014 samples (on Table 4), which means that the independent test sets are a good representation of the overall populations; analysing the residuals vs fit values plots (Appendix V and Appendix W), despite some outliers and some residuals unbalanced on the y-axis, both plots have evenly symmetrically distributed residuals, clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable, which aids validating the model's predictions.

Regarding the comparison with similar results in literature, Gomes *et al.* (2017b) had the best results for predictions on different vintages of wine grape berries with both, a machine learning algorithm (NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344 ºBrix, respectively for the mentioned models: in this work, superior results were obtained for the R^2 and RMSE values (0.953 and 0.977 ºBrix, respectively) on the test set composed by TF 2013 samples, but for the TF 2014 test set the results are rather inferior; comparing these results with the ones obtained by the NNs and DTs models in this work, the SVR model achieves rather superior results in the TF 2013 dataset but slightly inferior results for the TF 2014 test set when compared to the NNs model.

Overall, the SVR model achieved superior results for the prediction of all oenological parameters in comparison to the state of the art approaches and the results published by the other models in this work. In spite of increasing the number of principal components used as input and a small detriment on performance, the model was still able to capture most of the relations between the data and achieve very accurate predictions for different vintages, evidencing the best generalization capacity seen so far – consequently, no changes seem to be necessary to the prediction models using the SVR algorithm.

4.4.2.2. Different Vintages and Varieties

Graph 12 shows the results obtained by the SVR model for the prediction of anthocyanin concentration on different varieties and vintages of wine grape berries. Since there aren't any laboratorial results for the TF 2014 dataset, only the TF 2012 and TF 2013 set of samples compose the training and validation set. The number of principal components used was 1 and 7 for the TB 2013 and TN 2013 generalization sets, respectively.

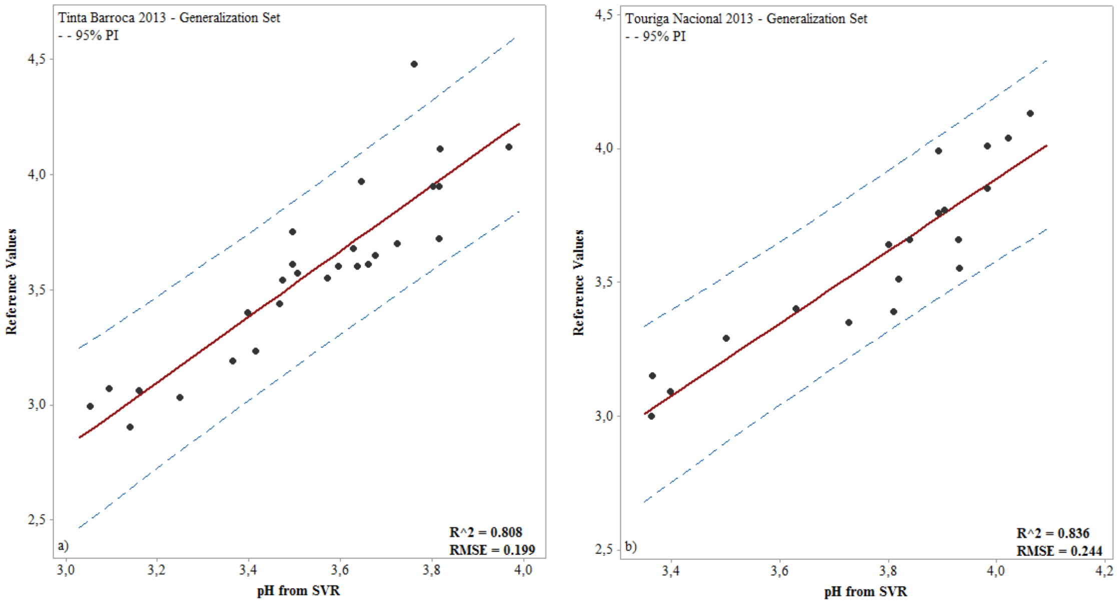


Graph 12 – Results for the estimation of anthocyanin concentration on different vintages and varieties with SVR; a) TB 2013 generalization set; b) TN 2013 generalization set

Observing Graph 12, the decrease in the accuracy of the SVR predictions is clear: the error measure suffers from a large increase (from RMSE on average between 15-30 mg.L^{-1} it goes as high as 49.010 and 40.396 mg.L^{-1} for the TB 2013 and TN 2013 datasets, respectively) and the R^2 values, naturally, decay as well (from R^2 usually above 0.90, the model obtained 0.846 and 0.783 for the TB 2013 and TN 2013 set of samples, respectively) – analysing exclusively the R^2 and RMSE parameters and acknowledging that a decrease on the model's performance is expected for a test setup of this nature, one can consider that the SVR achieved a rather accurate set of predictions, but the uncertainty of the predictions might be too high and should be considered a negative indicator for the model's generalization capacity; however, the results of the ANOVA tests mentioned in 3.1 (see Appendix C) pointed out that there are significant differences in the means between the TF 2012 and TB 2013 samples and the TF 2013 and TN 2013 datasets, which aids providing an explanation to the increase on the error measures: the populations in the training set have rather different patterns in the spectra to capture in the learning process than those on the test set, making the generalization step harder to carry without an increase on the uncertainty of the predictions; additionally, and contrary to the indicators on the generalization sets with different vintages, for these sets the number of principal components used was significantly smaller which means that the SVR model couldn’t find important information about the data on the remaining factors determined by the PCA; the descriptive statistics of the independent test sets (see Appendix M) show that the mean values fit the initial 95% confidence intervals (on Table 2) but the standard deviation values for both datasets are over the higher limit, which might indicate that the independent test sets aren’t a good representation of the overall populations; analysing the residuals vs fit value plots (Appendix X and Appendix Y), both plots have the residuals symmetrically distributed with no clear patterns identifiable.

Comparing these results with those published in literature for predicting anthocyanin concentration on different vintages and varieties of wine grape berries, Janik *et al.* (2007) has the best (and only) results with a R^2 of 0.900 for his NNs model with PLS scores as input, with a non-comparable error measure: but, as seen in 4.2.2.2 and 4.3.2.2, it can be considered that the model only generalizes for different vintages since the training and validation sets are composed by samples from 1999 to 2003 of 9 varieties and the test set has the same 9 varieties but only for the vintage year of 2004 – if a comparison is made with the results presented in this work for generalization on different vintages for the SVR model (see 4.4.2.1), one can see that the results shown are superior with significantly fewer samples from fewer harvest years; comparing the results with the ones obtained by the NNs and DTs models in this work, the SVR model shows (once again) the best results, indicating that it might be the model with the best generalization capacity.

Graph 13 presents the results for the prediction of pH index on different vintages and varieties of wine grape berries by the SVR model. The number of principal components used was 22 and 11 for the TB 2013 and TN 2013 datasets, respectively.

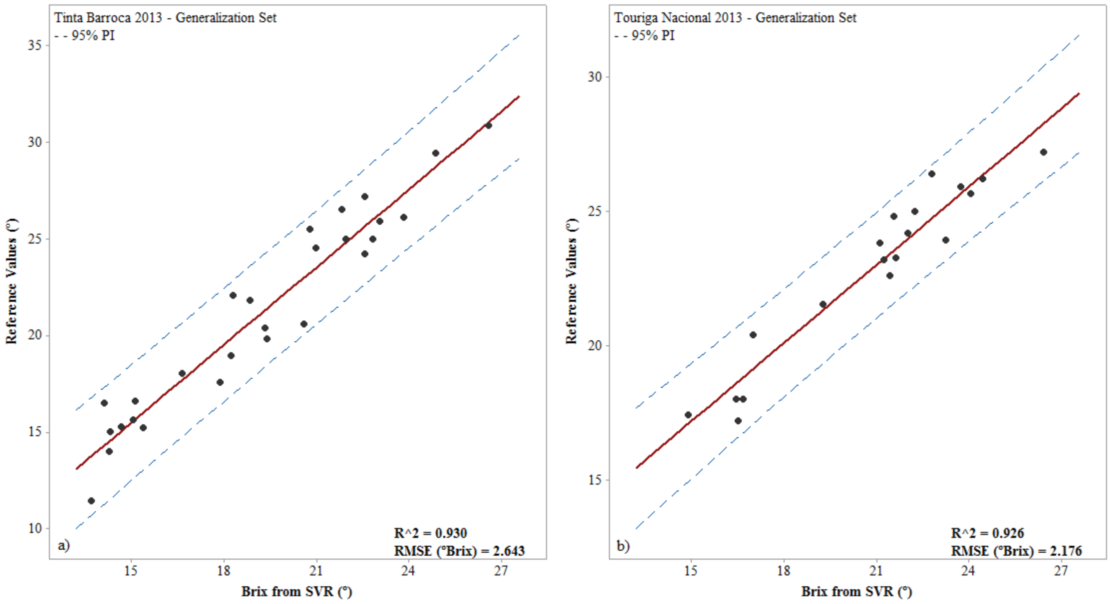


Graph 13 – Results for the estimation of pH index on different vintages and varieties with SVR; a) TB 2013 generalization set; b) TN 2013 generalization set

Inspecting Graph 13 it's clear that the SVR model suffered from a decrease in R^2 values when compared to the single variety and vintage models, with similar error measures: however, these can still be considered positive indicators of the model's generalization capacity, since the NNs and DTs results presented previously in this work only didn't show a decrease in performance because their results for the single variety and vintage models were far inferior when compared to the SVR results; the results of the ANOVA tests mentioned in 3.1 (see Appendix F) noted significant differences in the means between the TF 2014 dataset and the TF 2012 and TF 2013 samples, which means that the populations in the training set had rather different patterns in the spectra to be captured in the learning process, which aids providing an explanation for the decrease in the quality of the fits; the number of principal components used grew for the TN 2013 dataset but for the TB 2013 samples the number chosen was similar to the ones used in the single variety and vintage models – it might be a consequence of choosing the independent test sets randomly; the descriptive statistics of the independent test sets (see Appendix M) show that the standard deviation value for the TB 2013 set of samples doesn't fit the initial 95% confidence interval (on Table 3), which might indicate that this test set is not a good representation of the overall population; analysing the residuals vs fit values plots (Appendix X and Appendix Y), despite some outliers (especially for the TN 2013 test set), both plots seem to have the residuals following a rather evenly symmetrical distribution clustering towards the middle of the plot and towards the lower single digits of the y-axis, with no clear patterns identifiable.

Regarding the comparison with the current literature, there aren’t any works published that predict pH index on different varieties and vintages of wine grape berries: comparing the results with authors that employed only different vintages on the test sets (as seen in 4.4.2.1), Fadock *et al.* (2016) obtained a R^2 of 0.560 and a RMSE of 0.050 with his PLS regression model – despite the fact the test sets compose not only different vintages but also different varieties of wine grape berries, the results published in this work for the SVR model (R^2 of 0.808 and 0.836, RMSE of 0.199 and 0.244 for the TB 2013 and TN 2013 datasets, respectively) can be considered significantly better; comparing the results with the ones presented for the NNs and DTs models, some perform better on one of the test sets than in the other but, overall, all models achieved pretty similar results.

Graph 14 shows the results for the estimation of sugar content on different vintages and varieties of wine grape berries by the NNs model. The number of principal components used was 45 and 35 for the TB 2013 and TN 2013 datasets, respectively.



Graph 14 – Results for the estimation of sugar content on different vintages and varieties with SVR; a) TB 2013 generalization set; b) TN 2013 generalization set

Examining Graph 14, it’s clear the SVR model achieved very positive indicators regarding its generalization capacity: the error measures had a slight increase when compared to the single variety and vintage models, but the R^2 values continue showing a high correlation between the predictions and the ground-truth results; the results of the ANOVA tests mentioned in 3.1 (see Appendix I) noted significant differences in the means between almost every single variety and vintage, providing somewhat of an explanation to the increase on the degree of uncertainty; the number of principal components used grew to allow the model to achieve a more stable set of predictions; the descriptive statistics of the independent test sets (see Appendix M) show that the mean of the TB 2013 dataset and the standard deviation of the TN 2013 set of samples don't fit the initial 95% confidences intervals (on Table 4), which might indicate that these test sets are not a good representation of the overall populations; analysing the residuals vs fit values plots (Appendix X and Appendix Y), both plots seem to have the residuals following an evenly symmetrical distribution clustering towards the middle of the plot and towards the lower digits of the y-axis, with no clear patterns identifiable.

As for the comparison with current literature, similarly to the pH index analysis, there aren't any works published that predict sugar content on different varieties and vintages of wine grape berries, not allowing for a direct comparison to be made: considering the results for authors that employed only different vintages on the test sets (as seen in 4.4.2.1), Gomes *et al.* (2017b) had the best results with both, a machine learning algorithm (NNs) and a chemometric method (PLS regression), with a R^2 of 0.917 and 0.948 and a RMSE of 1.355 ºBrix and 1.344 ºBrix, respectively for the mentioned models – in this work, despite having test sets composed not only of different vintages but also of different varieties of wine grape berries, both test sets had a superior fit (R^2 of 0.930 and 0.926 for the TB 2013 and TN 2013 samples, respectively) when compared to the author's NNs model but (naturally) with inferior error measures; comparing the results with the ones presented for the NNs and DTs models in this work, the SVR achieved (once again) superior results.

Overall, despite a slight drop on the models' performance on the generalization sets and a higher error rate in the predictions (which can be considered as a reasonable outcome due to the fact that these varieties and vintages can't be found on the training steps, increasing the uncertainty), these results are very strong indicators in respect to the SVR generalization capacity, that obtained the overall best results of all the models implemented: this might indicate that, as mentioned previously, it won't be necessary to build models who require a yearly update of samples, or new samples for each variety.

4.5. Results’ Overview

Table 31 summarizes the best results obtained by each of the models presented for the prediction of anthocyanin concentration, pH index and sugar content on:

a) Single variety and vintage test sets.

b) Different vintage test sets (generalization sets).

c) Different variety and vintage test Sets (generalization sets).

Table 31 – Summary of the best results obtained by each of the models presented for the prediction of the oenological parameters in wine grape berries

	DTs						NNs						SVR					
	a)		b)		c)		a)		b)		c)		a)		b)		c)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Anthocyanin Concentration	0.942	20.964	0.916	45.034	0.839	37.981	0.968	15.463	0.922	20.504	0.834	32.887	0.979	11.887	0.938	28.349	0.846	49.010
pH Index	0.888	0.202	0.831	0.226	0.830	0.141	0.871	0.147	0.831	0.217	0.844	0.248	0.902	0.117	0.873	0.275	0.836	0.244
Sugar Content	0.930	2.870	0.879	2.304	0.781	4.636	0.963	1.314	0.913	2.383	0.925	3.329	0.979	1.760	0.953	0.977	0.930	2.643

Anthocyanin Concentration in mg.L^{-1}

Sugar Content in ºBrix

Analysing Table 31 it’s clear that all the models share similar results for the predictions of all the oenological parameters for the single variety and vintage test sets, with exception of the DTs model that achieves poor results for the sugar content prediction. However, for the generalization sets a difference in the capacity of the models to achieve accurate predictions arises, with the SVR model taking a prominent position with the best overall results while, on the other hand, the DTs model reveals the worst outcome indicating that (as mentioned previously) a fine tuning of model’s structure should occur for future works.

CHAPTER V – CONCLUSIONS AND FUTURE WORK

Hyperspectral imaging in reflectance mode was combined with several machine learning algorithms (Neural Networks, Decision Trees and Support Vector Regression) to compose a framework capable of predicting oenological parameters on different varieties and vintages of wine grape berries. This work brings forwards different means to achieve a fast, inexpensive and non-destructive type of analysis that provides an alternative to traditional methods when studying wine grape berries during ripening.

The results obtained represent progress in comparison to current state of the art publications in the prediction of anthocyanin concentration, pH index and sugar content for the majority of the models tested, maintaining a high performance through different varieties and vintages of wine grapes: this represents improvements in terms of the study of the generalization capacity, vital to achieve a model capable of predicting for a wide variety of wine grapes without the need to fine tune the model with new samples every vintage year, or for every different variety. Moreover, the hyperspectral imaging was conducted with a small number of whole berries, which is a setup rarely found in literature.

Conducting the methodology in an incremental manner, it was possible to find that the machine learning algorithms implemented benefit from data pre-processing, dimensionality reduction and model validation steps, easing the learning process and allowing for more complex relations in the patterns of the spectra to be found - for the intermediate test setups performed, the models response found that applying a Savitzky-Golay filter would improve the quality of the fits and reduce the error measures; it found that applying a Principal Component Analysis for dimensionality reduction would improve the models’ performance and decrease the computational cost; and finally, it compared typical model validation algorithms (k-Fold Cross-Validation, Monte-Carlo Cross-Validation and Bootstrap) and concluded that the first obtained similar results while greatly reducing the execution time of the models.

While studying the machine learning algorithms in depth it was possible to understand their learning process and list some of the variations that can be found in the structure of the Neural Networks (different methods to initialize the weights and bias of the network, different activation functions, different training algorithms and different number of neurons and hidden layers to compose the final structure), Decision Trees (choosing the number of individual Decision Trees to compose the bagging algorithm) and Support Vector Regression (different loss functions, kernel functions and optimizing methods for the hyper parameters).

Summarizing the performance of each model individually, the Neural Networks had either similar or superior results to all the machine learning approaches found in literature, revealing a strong generalization capacity for the prediction of all oenological parameters, but presented a slightly high error measure for the anthocyanin concentration and showed inferior results for the prediction of sugar content in different vintages of wine grape berries when compared to a chemometric method (Partial Least Squares regression); the Decision Trees had comparable results in respect to the state of the art approaches and the Neural Networks for the single vintage and variety models, but revealed a poor generalization capacity for different vintages of wine grape berries when compared to the remaining models; the Support Vector Regression model presented superior results to all the machine learning and chemometric approaches found in literature, with the best overall generalization capacity of all the models implemented.

Further works might include the in-depth study of different pre-processing and dimensionality reduction methods, since there is a wide variety of methods and test setups that weren't implemented (it would be relevant to study the effect of different pre-processing and dimensionality reduction methods on the generalization capacity of each model, instead of only testing the effects on single variety and vintage models) that might represent an improvement on the models' capacity to capture the different patterns in the spectra, especially for the estimation of the pH index that represented a decrease in performance for every model; the tuning of the Neural Networks topology and the Decision Trees structure, especially the latter, since despite having inferior results for the generalization capacity is natural for a model of reduced complexity, the underfitting might result from having a double model validation step that increased the bias; test the models accuracy with a higher number of samples of different varieties and vintages, since the results obtained were always referring to a relatively short number of samples when comparing with reference authors; and finally, an effort into introducing a Deep Learning framework would be extremely interesting, since it represents an emerging area of investigation with very good results in a wide variety of areas of application, and these algorithms have a higher capacity to extract complex patterns in the data, which would represent a strong tool for the analysis of the pH index and to improve the generalization capacity of the models.

BIBLIOGRAPHIC REFERENCES

Alpaydin, E. (2004). Linear Discrimination. In *Introduction to Machine Learning* (pp. 209–233). Massachusetts, USA: MIT Press.

Arana, I., Jarén, C., & Arazuri, S. (2005). Maturity, variety and origin determination in white grapes (Vitis vinifera L.) using near infrared reflectance technology. *Journal of Near Infrared Spectroscopy*, *13*(6), 349–357.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*(5), 772–777.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support Vector Regression. *Neural Information Processing - Letters and Reviews*, *11*(10), 203–224.

Bishop, C. M. (2006). Neural Networks. In *Pattern recognition and machine learning* (pp. 225–291). New York, USA: Springer.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*, 123–140.

Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*, 16–33.

Cao, F., Wu, D., & He, Y. (2010). Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy. *Computers and Electronics in Agriculture*, *71*, S15–S18.

Carbonneau, A., & Champagnol, F. (1993). Nouveaux systèmes de culture integré du vignoble. *Programme AIR*.

Chalimourda, A., Schölkopf, B., & Smola, A. J. (2004). Experimentally optimal ν in support vector regression for different noise models and parameter settings. *Neural Networks*, *17*, 127–141.

Chen, S., Zhang, F., Ning, J., Liu, X., Zhang, Z., & Yang, S. (2015). Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging. *Food Chemistry*, *172*, 788–93.

Cherkassky, V., & Mulier, F. (1998). Classification. In *Learning from Data: Concepts, Theory, and Methods* (pp. 340–403). New York, USA: John Wiley & Sons, Inc.

Christensen, R. (2011). One-Way ANOVA. In *Plane Answers to Complex Questions: The Theory of Linear Models* (pp. 91-105). New York, USA: Springer.

Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, *20*, 273–297.

Cozzolino, D., Cynkar, W., Janik, L., Dambergs, B., Francis, I. L., & Gishen, M. (2005). Measurement of colour, total soluble solids and pH in whole red grapes using visible and near infrared spectroscopy. In *Proceedings 12th Australian Wine Industry Technical Conference* (pp. 24–29).

Dambergs, R., Gishen, M., & Cozzolino, D. (2015). A Review of the State of the Art, Limitations, and Perspectives of Infrared Spectroscopy for the Analysis of Wine Grapes, Must, and Grapevine Tissue. *Applied Spectroscopy Reviews*, *50*(3), 261–278.

Fadock, M. (2011). *Non-Destructive VIS/NIR Reflectance Spectrometry for Red Wine Grape Analysis*. Masters Thesis in Applied Science at The University of Guelph, Ontario.

Fadock, M., Brown, R. B., & Reynolds, A. G. (2016). Visible-Near Infrared Reflectance Spectroscopy for Nondestructive Analysis of Red Wine Grapes. *American Journal of Enology and Viticulture*, *67*(1).

Fernandes, A. M., Franco, C., Mendes-Ferreira, A., Mendes-Faia, A., Costa, P. L. da, & Melo-Pinto, P. (2015). Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks. *Computers and Electronics in Agriculture*, *115*, 88–96.

Fernandes, A. M., Oliveira, P., Moura, J. P., Oliveira, A. A., Falco, V., Correia, M. J., & Melo-Pinto, P. (2011). Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks. *Journal of Food Engineering*, *105*(2), 216–226.

Fernández-Novales, J., López, M.-I., Sánchez, M.-T., García-Mesa, J.-A., & González-Caballero, V. (2009). Assessment of quality parameters in grapes during ripening using a miniature fiber-optic near-infrared spectrometer. *International Journal of Food Sciences and Nutrition*, *60 Suppl 7*(915060982), 265–277.

Ferrer-Gallego, R., Hernández-Hierro, J. M., Rivas-Gonzalo, J. C., & Escribano-Bailón, M. T. (2011). Determination of phenolic compounds of grape skins during ripening by NIR spectroscopy. *LWT - Food Science and Technology*, *44*(4), 847–853.

Friedman, J. H. (1997). On Bias, Variance, 0/1 - Loss, and the Curse-of-Dimensionality. *Data Mining and Knowledge Discovery*, *1*(1), 55–77.

Geraudie, V., Roger, J. M., Ferrandis, J. L., Gialis, J. M., Barbe, P., Maurel, V. B., & Pellenc, R. (2009). A revolutionary device for predicting grape maturity based on NIR spectrometry. In *FRUTIC 09: 8th Fruit Nut and Vegetable Production Engineering Symposium*.

Geraudie, V., Roger, J. M., & Ojeda, H. (2010). Développement d'un appareil permettant de prédire la maturité du raisin par spectroscopie proche infra-rouge. (PIR). *Revue Française d'Oenologie*, *240*.

Gomes, V. M., Fernandes, A. M., Faia, A., & Melo-Pinto, P. (2014a). Comparison of different approaches for the Prediction of Sugar Content in Whole Port Wine Grape Berries using Hyperspectral Imaging. In *ENBIS 14: 14th Annual Conference of the European Network for Business and Industrial Statistics*.

Gomes, V. M., Fernandes, A. M., Faia, A., & Melo-Pinto, P. (2014b). Determination of sugar content in whole Port Wine grape berries combining hyperspectral imaging with neural networks methodologies. *IEEE Symposium Series on Computational Intelligence*.

Gomes, V. M., Fernandes, A. M., Martins-Lopes, P., Pereira, L., Faia, A., & Melo-Pinto, P. (2017a). Characterization of neural network generalization in the determination of pH and anthocyanin concent of wine grape in new vintages and varieties. *Food Chemestry*, 218, 40-46.

Gomes, V. M., Fernandes, A. M., & Melo-Pinto, P. (2017b). Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries. *Computers and Electronics in Agriculture*, 140, 244–254.

González-Caballero, V., Pérez-Marín, D., López, M. I., & Sánchez, M. T. (2011). Optimization of NIR spectral data management for quality control of grape bunches during on-vine ripening. *Sensors*, *11*(6), 6109–6124.

Gowen, A., O'Donnell, C., Cullen, P., Downey, G., & Frias, J. (2007). Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, *18*(12), 590–598.

Hall, A., Lamb, D. W., Holzapfel, B., & Louis, J. (2002). Optical remote sensing applications in viticulture - a review. *Australian Journal of Grape and Wine Research*, *8*(1), 36–47.

Hand, D. J., Mannila, H., & Smyth, P. (2001a). A Systematic Overview of Data Mining Algorithms. In *Principles of Data Mining* (pp. 141–165). Massachusetts, EUA: MIT Press.

Hand, D. J., Mannila, H., & Smyth, P. (2001b). Predictive Modeling for Classification. In *Principles of Data Mining* (pp. 327–367). Massachusetts, EUA: MIT Press.

Hayashi, Y., Sakata, M., & Gallant, S. I. (1990). Multi-layer versus single-layer neural networks and an application to reading hand-stamped characters. In *Proceedings of the*

International Conference on Neural Networks (pp. 781–784). Paris.

Hernández-Hierro, J. M., Nogales-Bueno, J., Rodríguez-Pulido, F. J., & Heredia, F. J. (2013). Feasibility study on the use of near-infrared hyperspectral imaging for the screening of anthocyanins in intact grapes during ripening. *Journal of Agricultural and Food Chemistry*, *61*(41), 9804–9809.

Herrera, J., Guesalaga, A., & Agosin, E. (2003). Shortwave near infrared spectroscopy for non-destructive determination of maturity of wine grapes. *Measurement Science and Technology*, *14*(5), 689–697.

Horváth, G. (2003). Neural networks in measurement systems. *Advances in Learning Theory: Methods, Models and Applications*, 375–402.

Huang, C.-L., & Wang, C.-J. (2006). GA-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with Applications*, *31*(2), 231–240.

Hyndman, R. J. (2010). Why every statistician should know about cross-validation. Retrieved September 28, 2017, from https://robjhyndman.com/hyndsight/crossvalidation/

Janik, L. J., Cozzolino, D., Dambergs, R., Cynkar, W., & Gishen, M. (2007). The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and artificial neural networks. *Analytica Chimica Acta*, *594*(1), 107–118.

Kecman, V. (2001). Multilayer Perceptrons. In *Learning and soft computing* (pp. 255–313). Massachusetts, USA: MIT Press.

Kohavi, R. (1995). A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2* (pp. 1137–1143). Quebec, Canada: Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Larraín, M., Guesalaga, A. R., & Agosin, E. (2008). A multipurpose portable instrument for determining ripeness in wine grapes using NIR spectroscopy. *IEEE Transactions on Instrumentation and Measurement*, *57*(2), 294–302.

Le Moigne, M., Dufour, E., Bertrand, D., Maury, C., Seraphin, D., & Jourjon, F. (2008). Front face fluorescence spectroscopy and visible spectroscopy coupled with chemometrics have the potential to characterise ripening of Cabernet Franc grapes. *Analytica Chimica Acta*, *621*(1), 8–18.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lendasse, A., Wertz, V., & Verleysen, M. (2003). Model selection with cross-validations and bootstraps - application to time series prediction with RBFN models. In *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP* (pp. 174–174). Istanbul, Turkey.

Levenberg, K. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics*, *2*, 164–168.

Marquardt, D. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, *11*(2), 431–441.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*, 115–133.

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(8), 1778–1790.

Mercier, G., & Lennon, M. (2003). Support vector machines for hyperspectral image classification with spectral-based kernels. In *Geoscience and Remote Sensing*

Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International (Vol. 1, pp. 288–290).

Nguyen, D., & Widrow, B. (1990). Improving the learning speed of 2-layer neural net works by choosing initial values of the adaptive weights. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 21–26).

Nogales-Bueno, J., Hernández-Hierro, J. M., Rodríguez-Pulido, F. J., & Heredia, F. J. (2014). Determination of technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during ripening by near infrared hyperspectral image: A preliminary approach. *Food Chemistry*, 152, 586-591.

Noguerol-Pato, R., González-Barreiro, C., Cancho-Grande, B., Martínez, M. C., Santiago, J. L., & Simal-Gándara, J. (2012). Floral, spicy and herbaceous active odorants in Gran Negro grapes from shoulders and tips into the cluster, and comparison with Brancellao and Mouratón varieties. *Food Chemistry*, *135*(4), 2771–82.

Noguerol-Pato, R., González-Barreiro, C., Cancho-Grande, B., Santiago, J. L., Martínez, M. C., & Simal-Gándara, J. (2012). Aroma potential of Brancellao grapes from different cluster positions. *Food Chemistry*, *132*(1), 112–124.

Noguerol-Pato, R., González-Barreiro, C., Simal-Gándara, J., Martínez, M. C., Santiago, J. L., & Cancho-Grande, B. (2012). Active odorants in Mouratón grapes from shoulders and tips into the bunch. *Food Chemistry*, *133*(4), 1362–1372.

Office International de la Vigne et du Vin (1990). *Recueil des méthodes internationales d'analyse des vins et des moûts: édition officielle, juin 1990*. Paris: O.I.V.

Raschka, S. (2014). *The Effect of Scaling and Mean Centering Prior to a Principal Component Analysis*. Retrieved from http://nbviewer.jupyter.org/github/rasbt/pattern_classification/blob/master/dimensionality_reduction/projection/scale_center_pca/scale_center_pca.pdf

Rastrigin, L. A. (1963). The convergence of the random search method in the extremal control of a many parameter system. *Automation and Remote Control*, *24*(10), 1337–1342.

Rawlings, J.O., Pantula, S.G., & Dickey, D.A. (1998). Regression Diagnostics. In *Applied Regression Analysis: A Research Tool* (pp. 341-397). New York, USA: Springer.

Remesan, R., & Mathew, J. (2014). Model Data Selection and Data Pre-processing Approaches. In *Hydrological Data Driven Modelling: A Case Study Approach* (pp. 41–67). New York: Springer.

Ribéreau-Gayon, P., & Stonestreet, E. (1965). Determination of anthocyanins in red wine. *Bulletin de la Société chimique de France*, *9*, 2649–52.

Rokach, L., & Maimon, O. (2015a). Beyond Classification Tasks. In *Data Mining with Decision Trees: Theory and Applications* (2nd ed., pp. 85–99). Toh Tuck Link, Singapore: World Scientific Publishing.

Rokach, L., & Maimon, O. (2015b). *Data Mining with Decision Trees: Theory and Applications* (2nd ed.). Toh Tuck Link, Singapore: World Scientific Publishing.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington D.C, USA: Spartan.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, *1*, 318–362.

Rumpf, T., Mahlein, A.-K., Steiner, U., Oerke, E.-C., Dehne, H.-W., & Plümer, L. (2010). Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, *74*(1), 91–99.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639.

Schaare, P. N., & Fraser, D. G. (2000). Comparison of reflectance, interactance and transmission modes of visible-near infrared spectroscopy for measuring internal properties of kiwifruit (Actinidia chinensis). *Postharvest Biology and Technology*, *20*, 175–184.

Shalizi, C. (2009). *Lecture 10: Regression Trees*. Pittsburgh, Pennsylvania, USA.

Silva, R., Gomes, V., Faia, A., & Melo-Pinto, P. (2016). *Support Vector Regression and Hyperspectral Imaging applied to Oenological Parameters Estimation on Different Varieties and Vintages of Wine Grapes*. Manuscript submitted for publication.

Smits, G. F., & Jordaan, E. M. (2002). Improved SVM regression using mixtures of kernels. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)* (Vol. 3, pp. 2785–2790). IEEE.

Smola, A. J., & Schölkopf, B. (2004). A Tutorial on Support Vector Regression. *Statistics and Computing*, *14*(3), 199–222.

State College PA. (2010). Minitab 17 Statistical Software. Minitab, Inc. Retrieved from www.minitab.com

Tarter, M. E., & Keuter, S. E. (2005). Effect of rachis position on size and maturity of Cabernet Sauvignon berries. *American Journal of Enology and Viticulture*, *56*(1), 86–89.

The Mathworks. (2016). MATLAB and Statistics Toolbox Release 2016b. The Mathworks, Inc. Retrieved from www.mathworks.com

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. *Springer* (Vol. 8).

Wang, W., Xu, Z., Lu, W., & Zhang, X. (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, *55*(3), 643–663.

Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *IRE WESCON Convention Record*, *4*, 96–104.

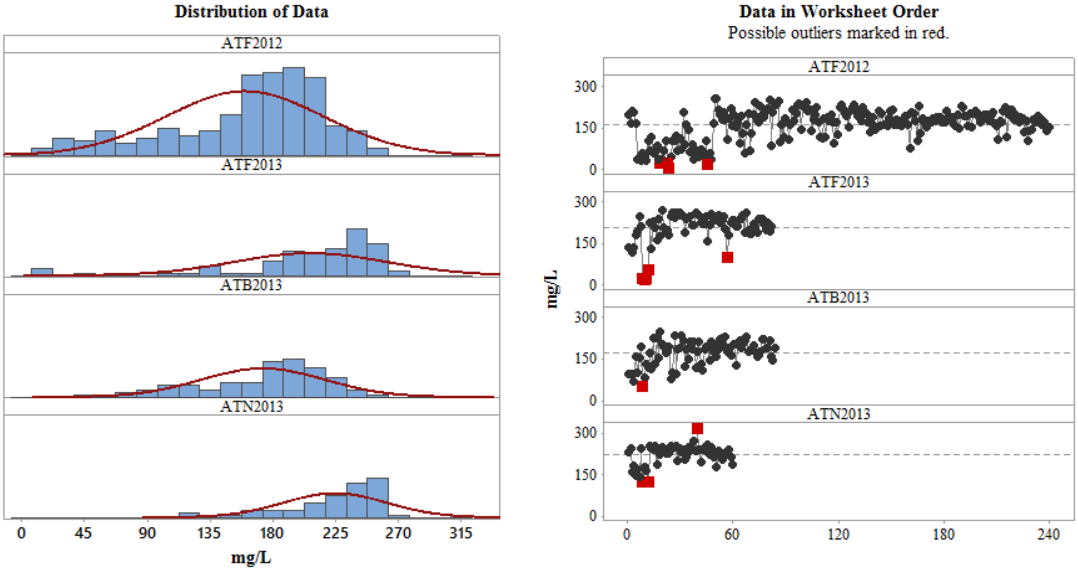
Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52.

Wu, G.-F., Huang, L.-X., & He, Y. (2008). Research on the sugar content measurement of grape and berries by using Vis/NIR spectroscopy technique. *Spectroscopy and Spectral Analysis*, *28*(9), 2090–3.

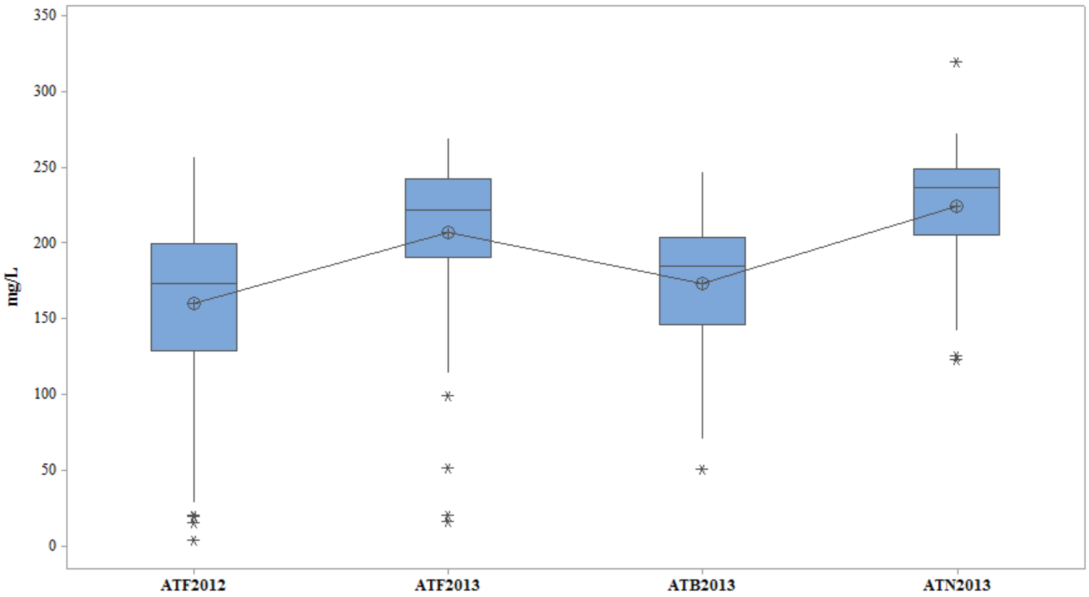
Zhang, H., Chen, L., Qu, Y., Zhao, G., & Guo, Z. (2014). Support vector regression based on grid-search method for short-term wind power forecasting. *Journal of Applied Mathematics*, 11.

APPENDICES

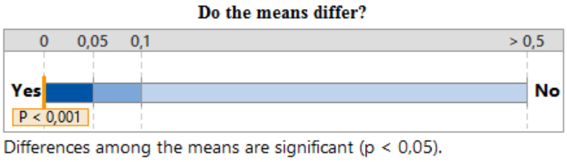
APPENDIX A – Data distribution for the anthocyanin concentration values of the laboratory results



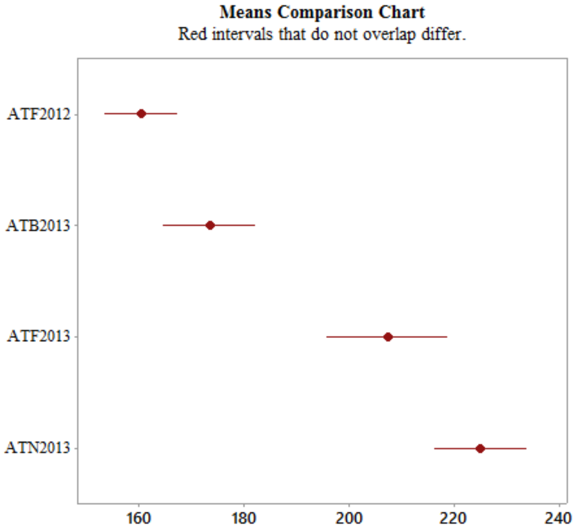
APPENDIX B – Boxplots for the anthocyanin concentration values of the laboratory results



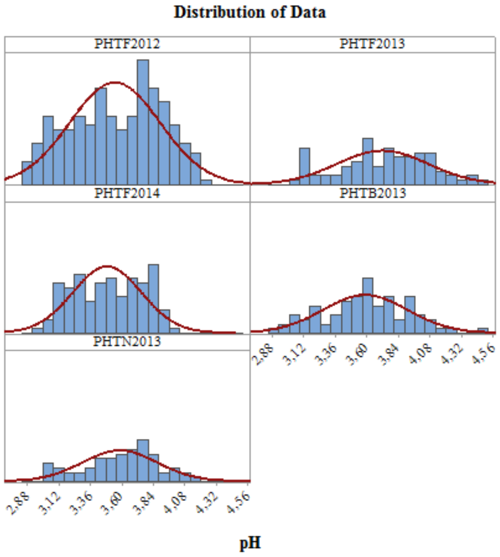
APPENDIX C – Summary report of the One-Way ANOVA tests for anthocyanin values in laboratory results

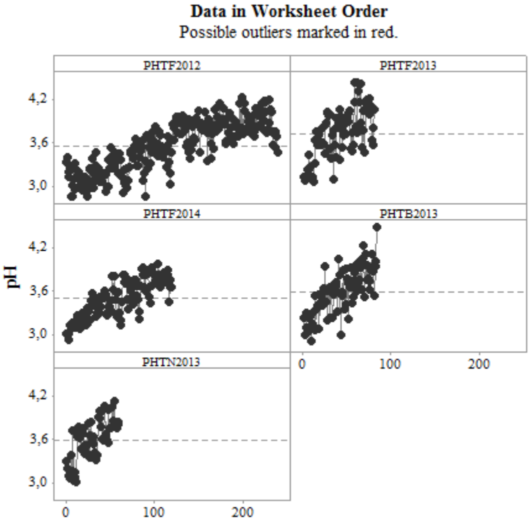

Do the means differ?

Differences among the means are significant (p < 0,05).

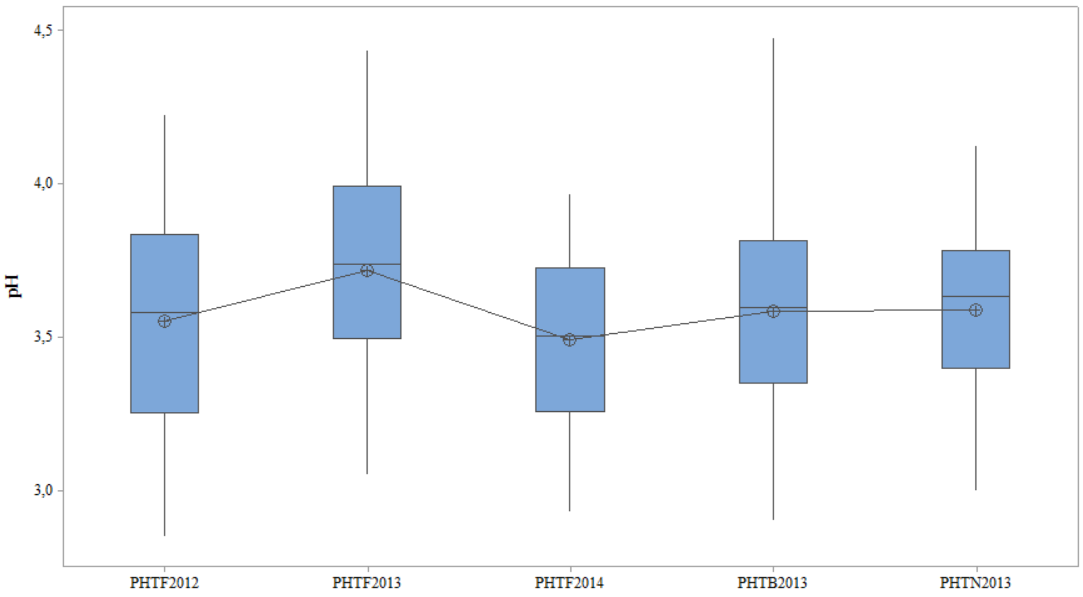

Means Comparison Chart
Red intervals that do not overlap differ.

APPENDIX D – Data distribution for the pH index values of the laboratory results



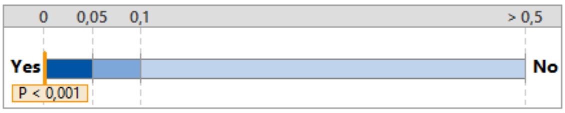


APPENDIX E – Boxplots for the pH index values of the laboratory results



APPENDIX F – Summary report of the One-Way ANOVA tests for pH index values in laboratory results

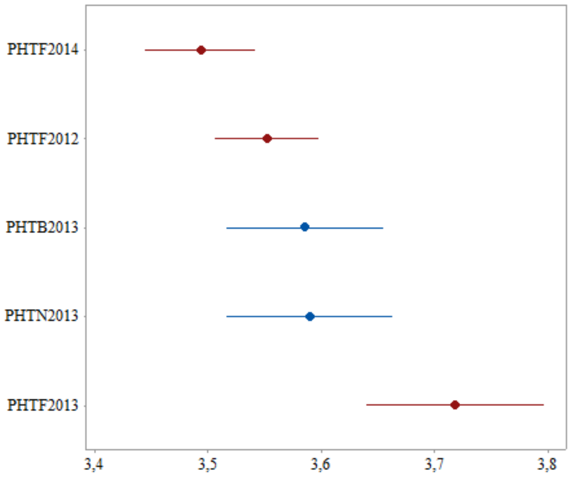
Do the means differ?



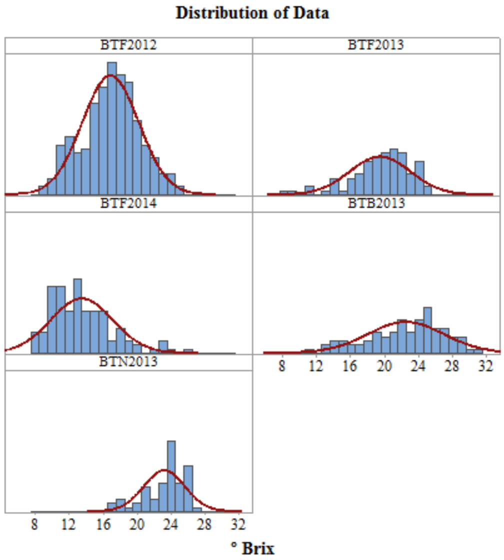
Differences among the means are significant (p < 0,05).

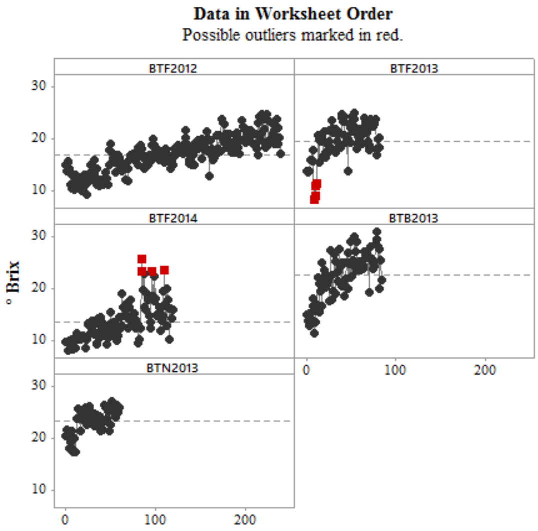
Means Comparison Chart

Red intervals that do not overlap differ.

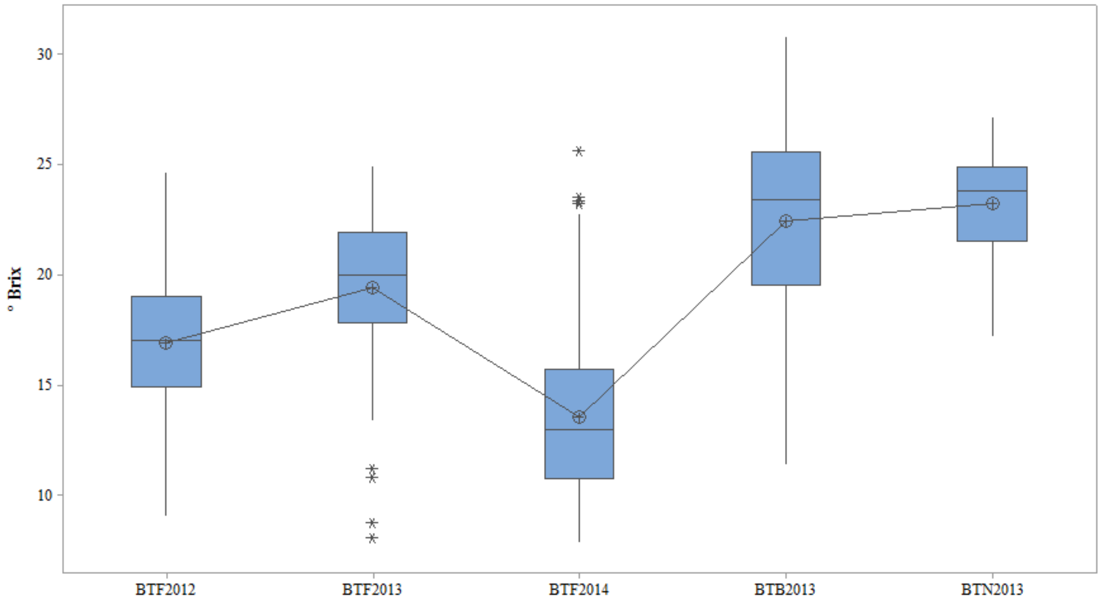


APPENDIX G – Data distribution for the sugar content values of the laboratory results

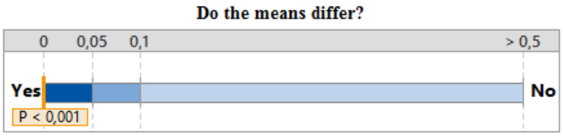




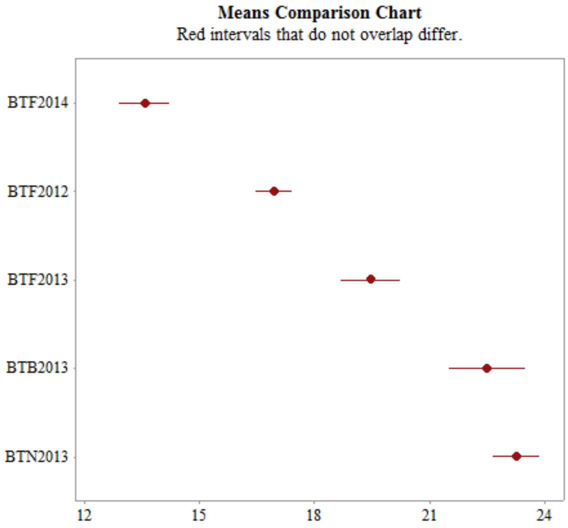
APPENDIX H – Boxplots for the sugar content values of the laboratory results



APPENDIX I – Summary report of the One-Way ANOVA tests for sugar content values in laboratory results

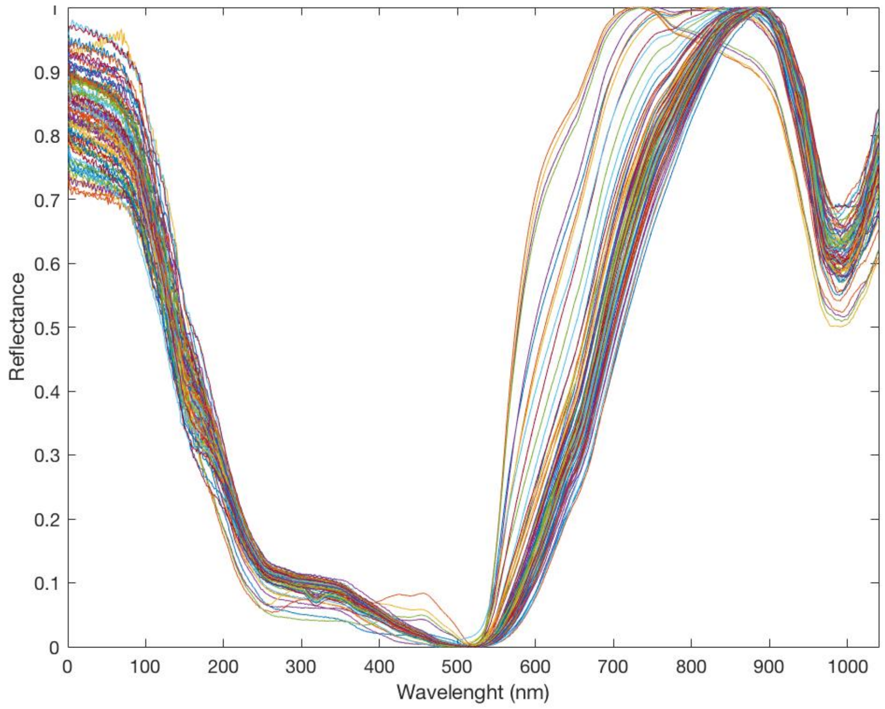


Differences among the means are significant (p < 0,05).

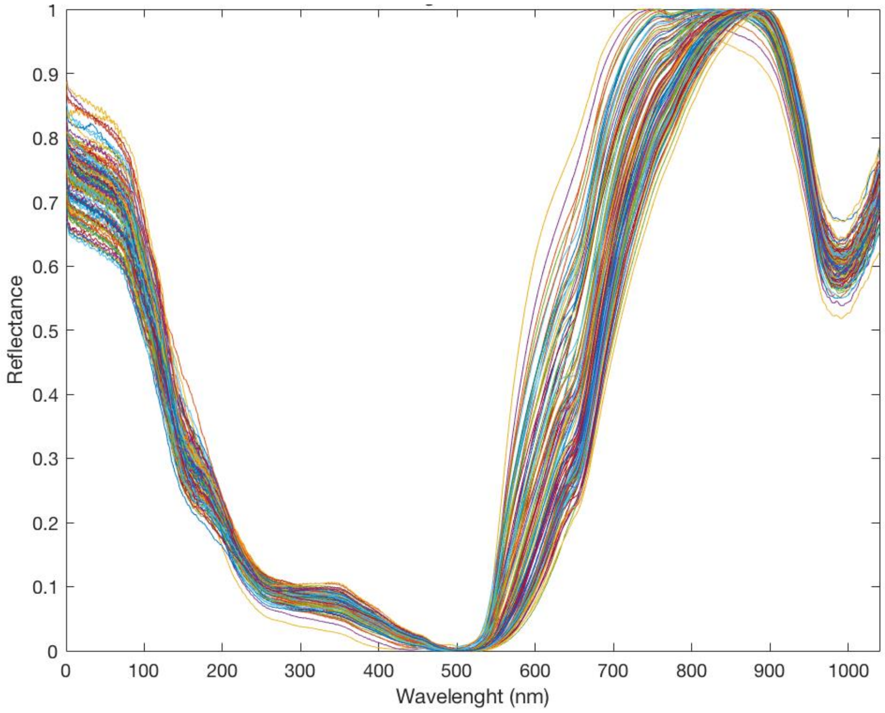


APPENDIX J – Reflectance measurements for the TF 2013, TF 2014, TB 2013 and TN 2013 samples, respectively.

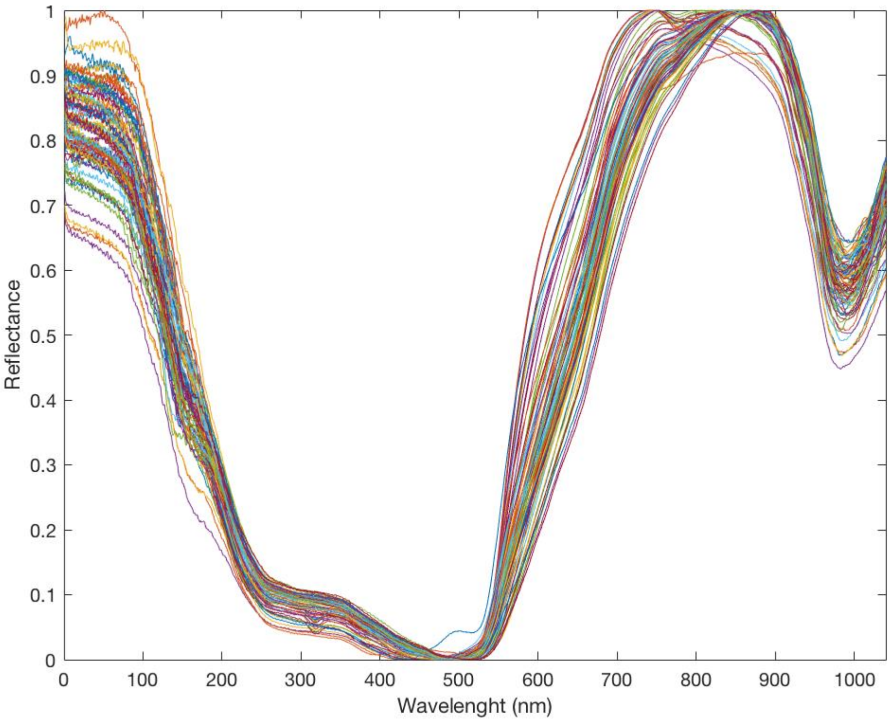
Reflectance measurements for the TF 2013 samples



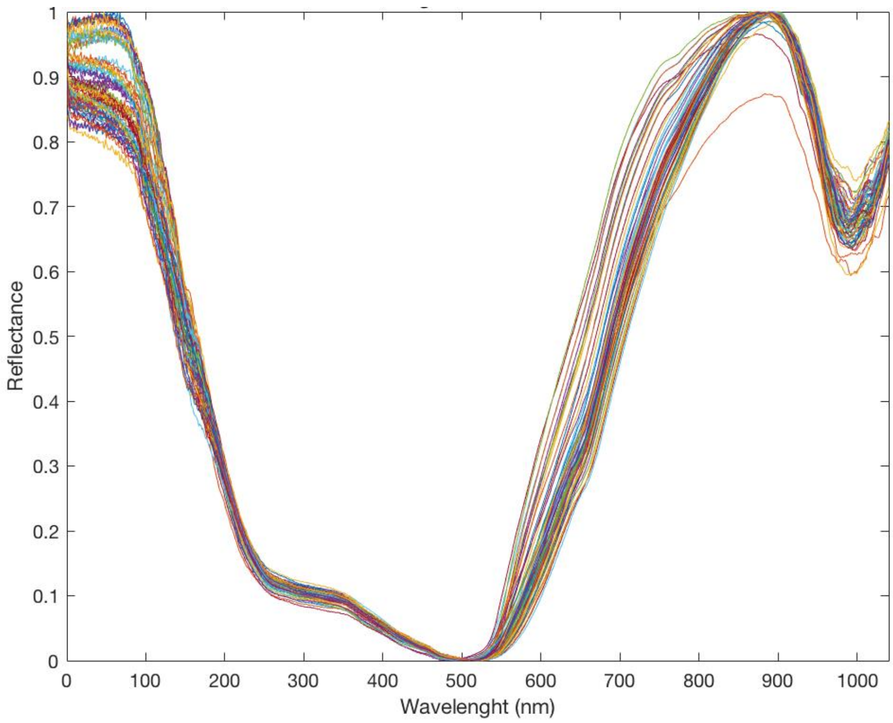
Reflectance measurements for the TF 2014 samples



Reflectance measurements for the TB 2013 samples



Reflectance measurements for the TN 2013 samples



APPENDIX K – Descriptive statistics of the laboratory results of the samples used on the generalization sets for the NN model

Anthocyanin Concentration (mg.L^{-1})

Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	216.34	(191.12; 241.56)	62.44	(48.97; 86.19)	16.28	232.62	269.75
TB 2013	27	178.77	(160.29; 197.25)	46.72	(36.80; 64.03)	50.97	190.31	247.76
TN 2013	19	231.50	(208.95; 254.05)	46.79	(35.35; 69.19)	123.68	248.28	319.90

pH Index

Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	3.70	(3.55; 3.86)	0.39	(0.31; 0.54)	3.05	3.74	4.44
TF 2014	37	3.49	(3.39; 3.59)	0.29	(0.24; 0.38)	2.93	3.49	3.97
TB 2013	27	3.57	(3.42; 3.73)	0.39	(0.31; 0.53)	2.90	3.58	4.48
TN 2013	19	3.48	(3.32; 3.65)	0.34	(0.25; 0.50)	3.00	3.53	4.13

Sugar Content (ºBrix)

Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	19.78	(18.29; 21.27)	3.68	(2.89; 5.09)	8.10	20.74	25.00
TF 2014	37	13.09	(11.90; 14.28)	3.57	(2.90; 4.64)	7.87	12.73	25.66
TB 2013	27	21.65	(19.45; 23.86)	5.57	(4.39; 7.63)	11.40	22.14	30.85
TN 2013	19	23.43	(21.88; 24.97)	3.32	(2.43; 4.75)	17.20	24.67	27.2

APPENDIX L – Descriptive statistics of the laboratory results of the samples used on the generalization sets for the DT model

Anthocyanin Concentration (mg.L^{-1})								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2013	26	197.50	(167.69; 227.30)	73.78	(57.86; 101.85)	16.28	230.17	269.75
TB 2013	27	171.73	(150.74; 192.71)	53.06	(41.78; 72.71)	50.97	188.56	247.76
TN 2013	19	231.49	(189.07; 237.92)	50.68	(38.29; 74.94)	123.68	216.91	319.90

pH Index								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2013	26	3.76	(3.59; 3.93)	0.42	(0.33; 0.58)	3.05	3.81	4.44
TF 2014	37	3.46	(3.37; 3.55)	0.27	(0.22; 0.35)	2.93	3.47	3.97
TB 2013	27	3.50	(3.35; 3.66)	0.39	(0.31; 0.53)	2.90	3.45	4.48
TN 2013	19	3.59	(3.43; 3.74)	0.33	(0.25; 0.49)	3.00	3.64	4.13

Sugar Content (ºBrix)								
Variety	**N**	**Mean**	**95% CI**	**St. Dev.**	**95% CI**	**Min**	**Median**	**Max**
TF 2013	26	18.59	(16.83; 20.35)	4.37	(3.42; 6.03)	8.10	19.50	25.00
TF 2014	37	13.93	(12.45; 15.41)	4.45	(3.62; 5.78)	7.87	12.87	25.66
TB 2013	27	22.65	(20.65; 24.64)	5.04	(3.97; 6.90)	11.40	23.51	30.85
TN 2013	19	23.34	(22.12; 24.57)	2.54	(1.92; 3.76)	17.20	23.92	27.20

APPENDIX M – Descriptive statistics of the laboratory results of the samples used on the generalization sets for the SVR model

Anthocyanin Concentration (mg.L^{-1})

Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	192.80	(160.80; 224.81)	79.25	(62.15; 109.39)	16.28	221.38	269.75
TB 2013	27	179.74	(158.71; 200.76)	53.14	(41.85; 72.83)	50.97	193.73	247.76
TN 2013	19	219.71	(196.52; 242.90)	48.11	(36.35; 71.15)	123.68	237.21	319.90

pH Index

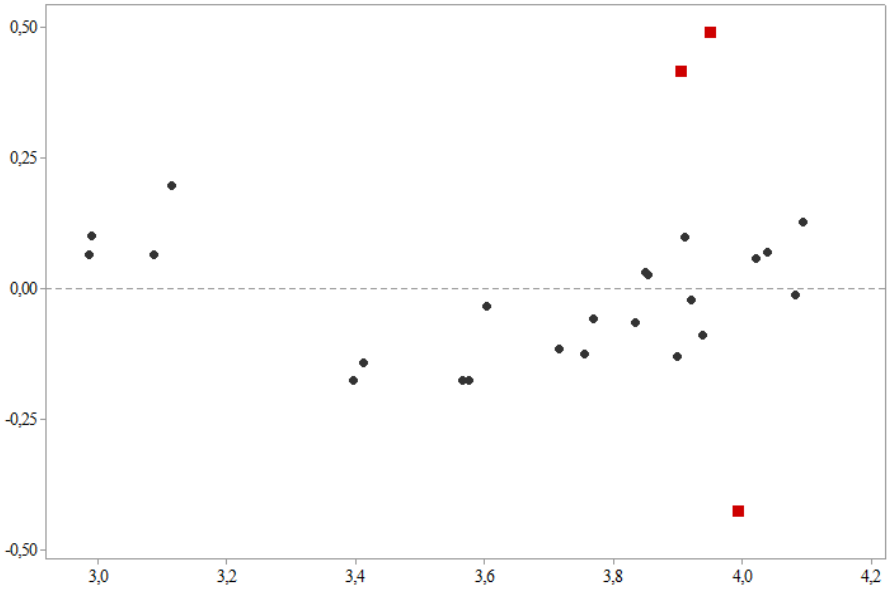
Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	3.66	(3.49; 3.84)	0.43	(0.34; 0.59)	3.05	3.69	4.44
TF 2014	37	3.51	(3.41; 3.61)	0.30	(0.24; 0.39)	2.93	3.52	3.97
TB 2013	27	3.57	(3.42; 3.73)	0.39	(0.30; 0.53)	2.90	3.60	4.48
TN 2013	19	3.59	(3.43; 3.75)	0.33	(0.25; 0.49)	3.00	3.64	4.13

Sugar Content (ºBrix)

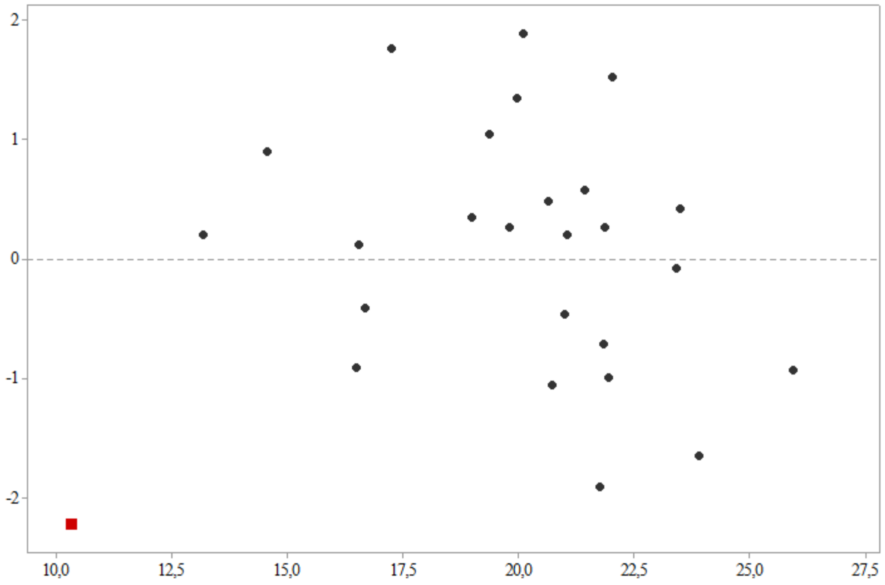
Variety	N	Mean	95% CI	St. Dev.	95% CI	Min	Median	Max
TF 2013	26	19.51	(17.88; 21.13)	4.02	(3.15; 5.55)	8.10	20.81	25.00
TF 2014	37	14.30	(12.63; 15.97)	5.01	(4.07; 6.51)	7.87	13.14	25.66
TB 2013	27	21.07	(19.01; 23.13)	5.21	(4.10; 7.13)	11.40	20.60	30.85
TN 2013	19	22.88	(21.32; 24.44)	3.24	(2.45; 4.80)	17.20	23.80	27.20

APPENDIX N – Residuals vs fit values plots for the prediction of pH index and sugar content, respectively, on the TF 2013 generalization set by the NN model

Residuals vs fit values plot for the prediction of pH index on the TF 2013 generalization set by the NN model

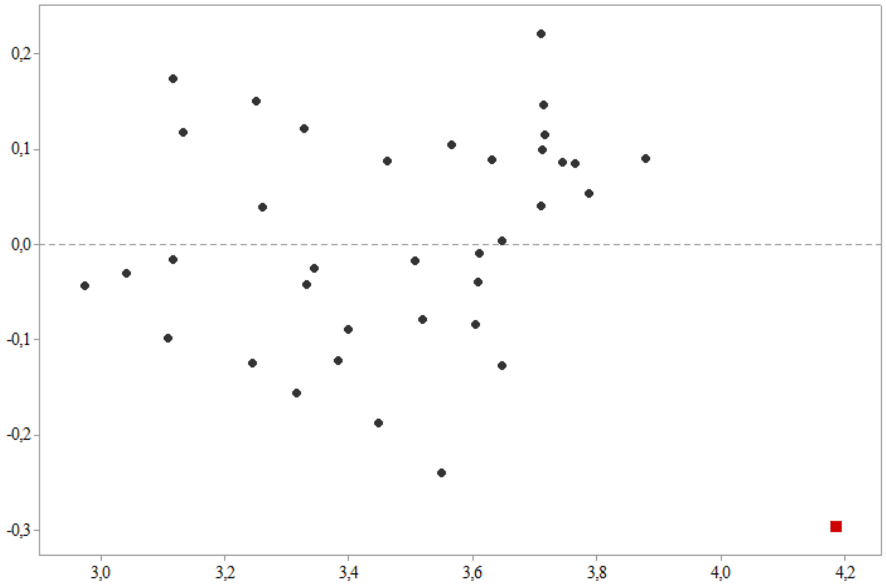


Residuals vs fit values plot for the prediction of sugar content on the TF 2013 generalization set by the NN model

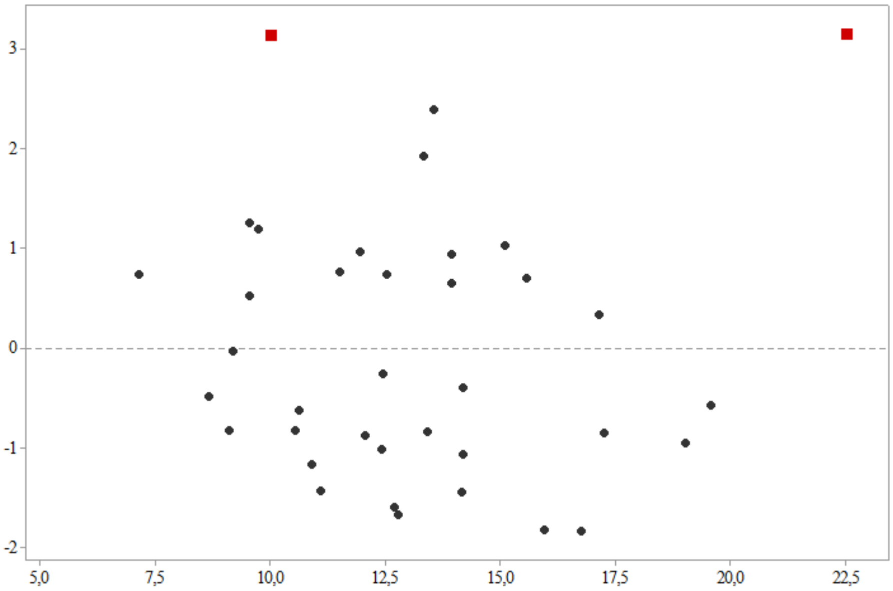


APPENDIX O – Residuals vs fit values plot for the prediction of pH index and sugar content, respectively, on the TF 2014 generalization set by the NN model

Residuals vs fit values plot for the prediction of pH index on the TF 2014 generalization set by the NN model

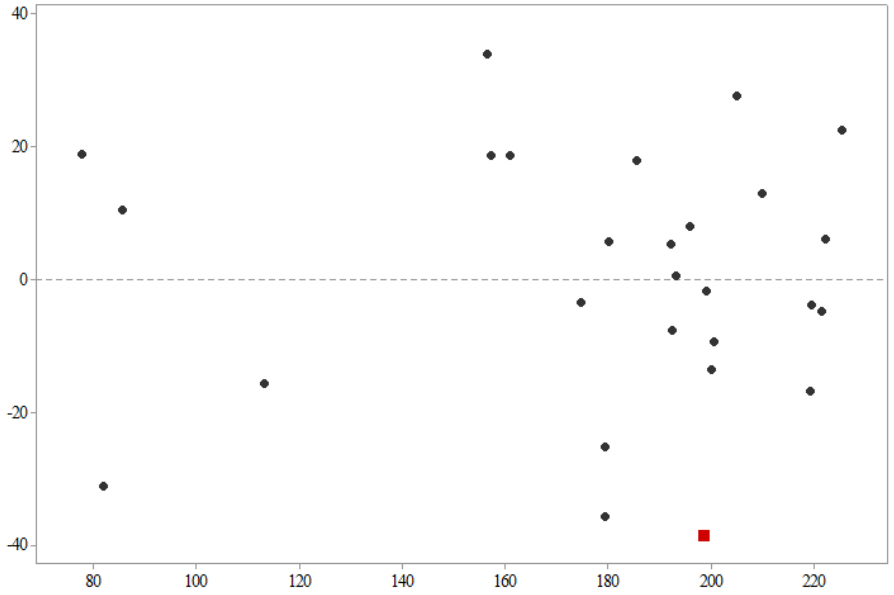


Residuals vs fit values plot for the prediction of sugar content on the TF 2014 generalization set by the NN model

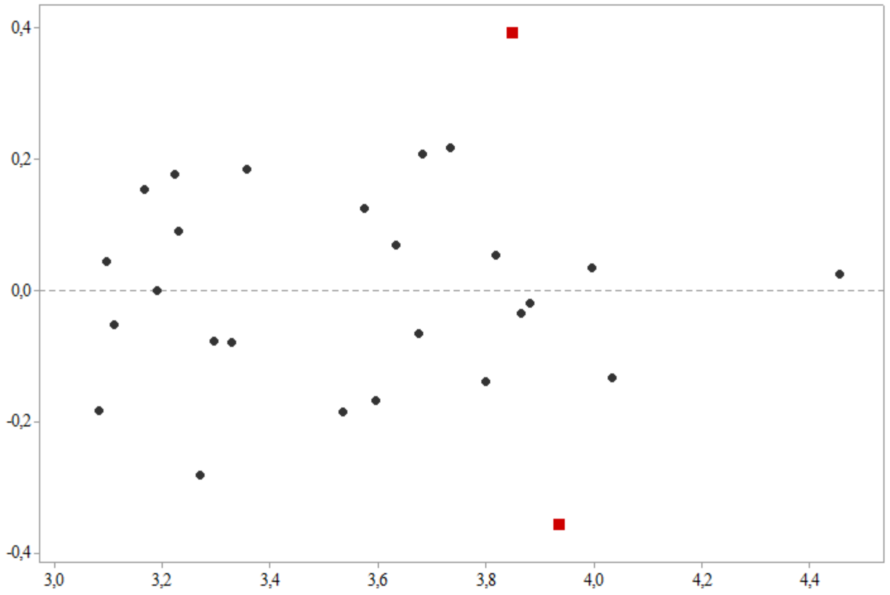


APPENDIX P – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TB 2013 generalization set by the NN model

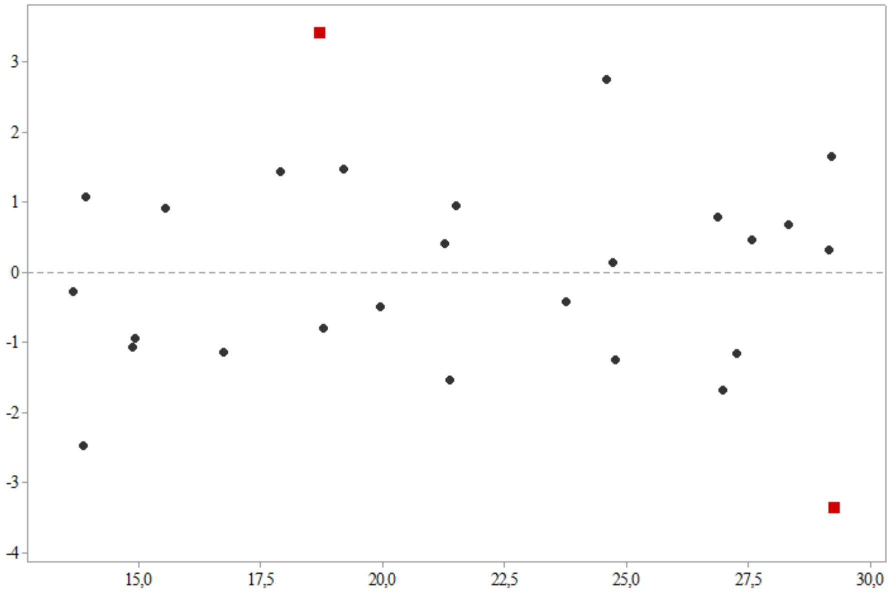
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TB 2013 generalization set by the NN model



Residuals vs fit values plot for the prediction of pH index on the TB 2013 generalization set by the NN model

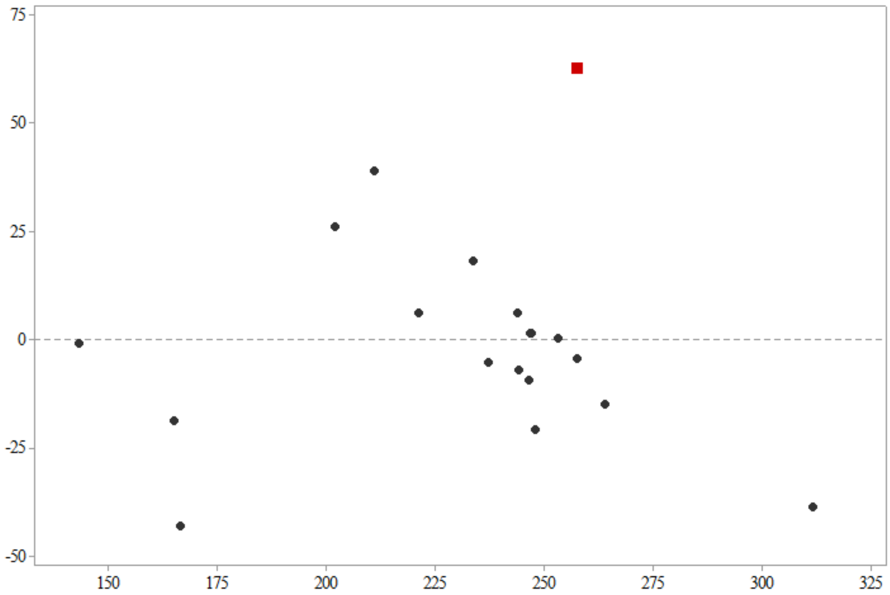


Residuals vs fit values plot for the prediction of sugar content on the TB 2013 generalization set by the NN model

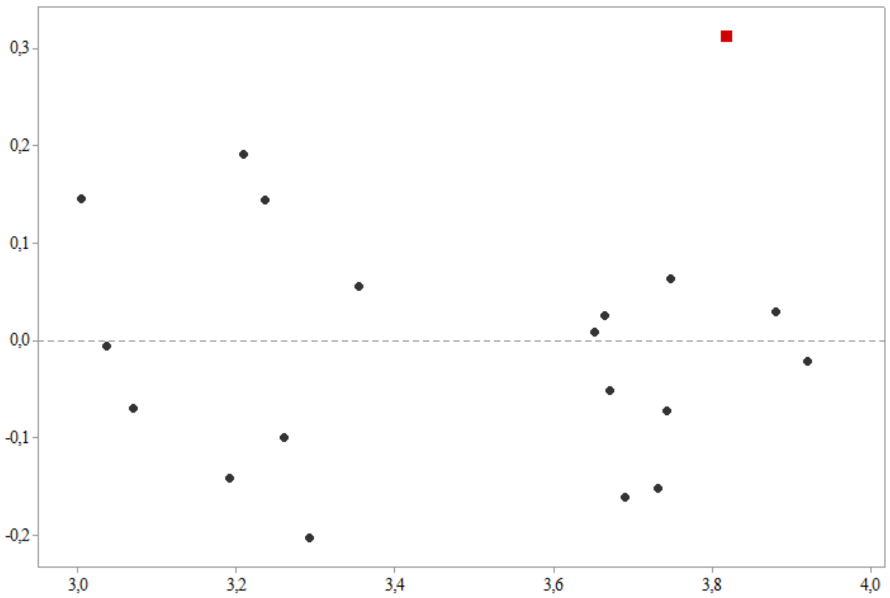


APPENDIX Q – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TN 2013 generalization set by the NN model

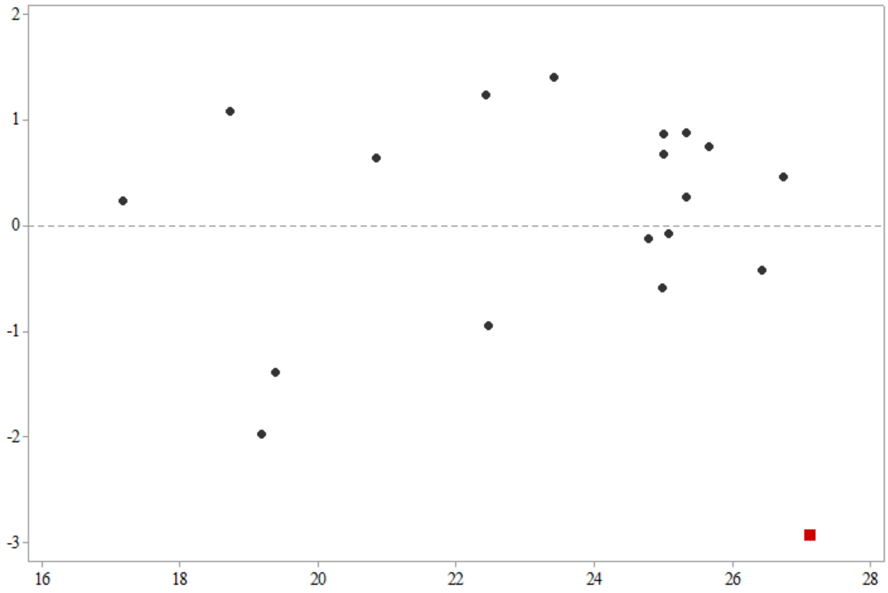
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TN 2013 generalization set by the NN model



Residuals vs fit values plot for the prediction of pH index on the TN 2013 generalization set by the NN model

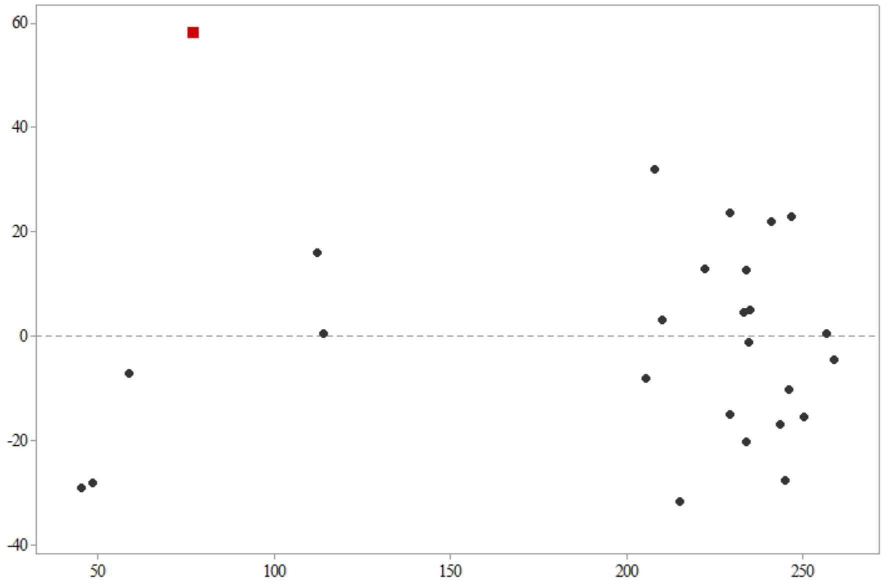


Residuals vs fit values plot for the prediction of sugar content on the TN 2013 generalization set by the NN model

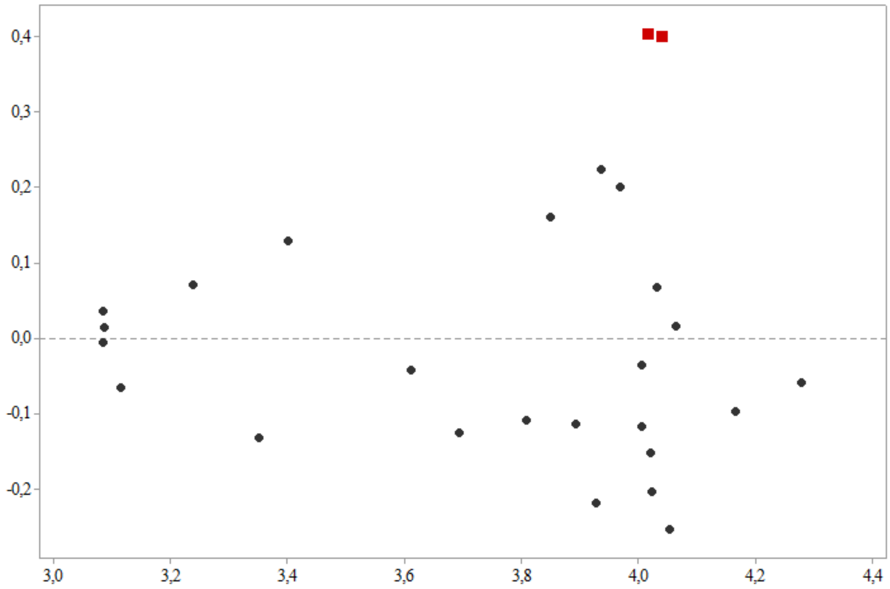


APPENDIX R – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TF 2013 generalization set by the DT model

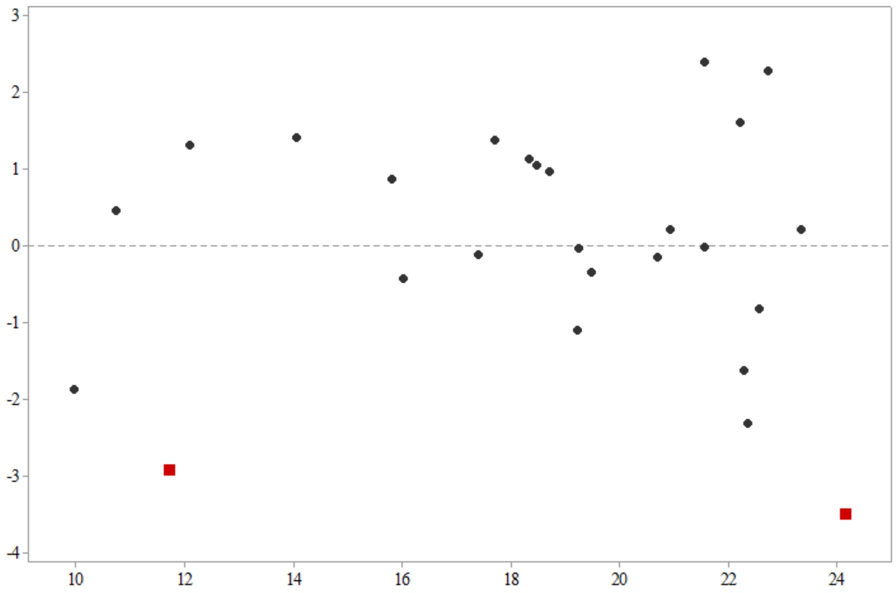
Residual vs fit values plot for the prediction of anthocyanin concentration on the TF 2013 generalization set by the DT model



Residual vs fit values plot for the prediction of pH index on the TF 2013 generalization set by the DT model

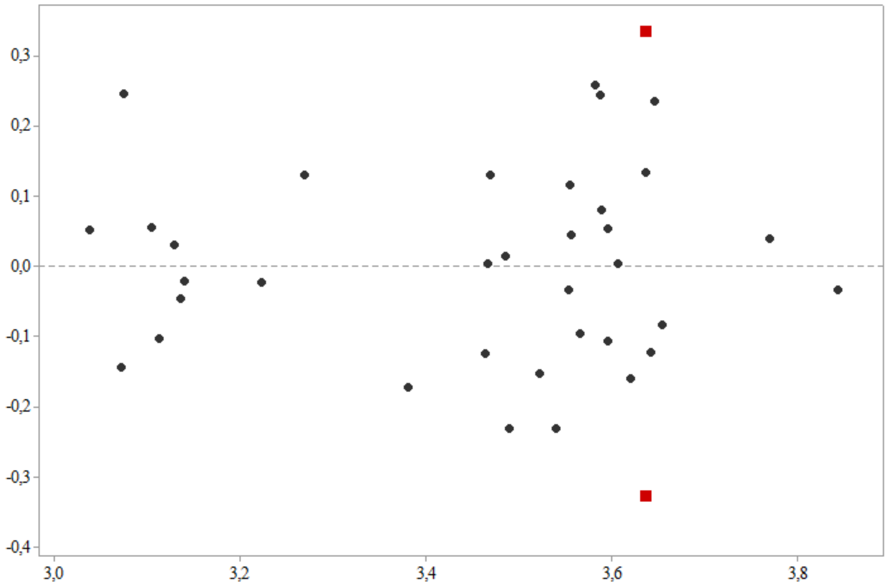


Residual vs fit values plot for the prediction of sugar content on the TF 2013 generalization set by the DT model

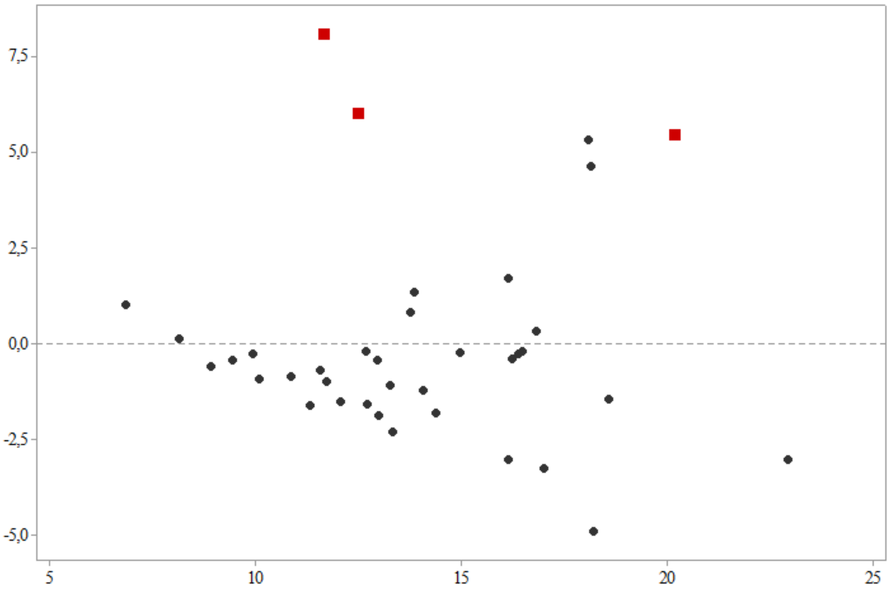

3
2
1
0
-1
-2
-3
-4
10
12
14
16
18
20
22
24

APPENDIX S – Residuals vs fit values plot for the prediction of pH index and sugar content, respectively, on the TF 2014 generalization set by the DT model

Residuals vs fit values plot for the prediction of pH index on the TF 2014 generalization set by the DT model

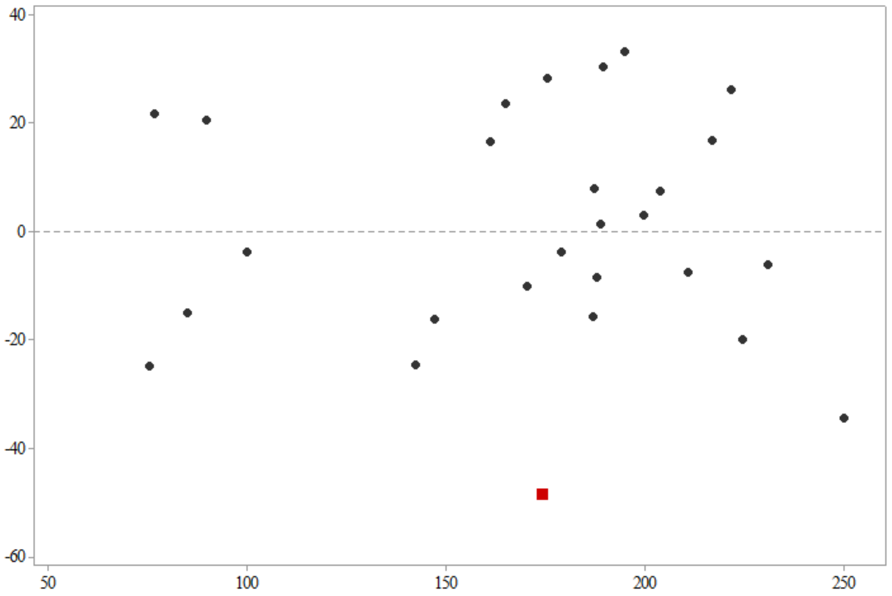


Residuals vs fit values plot for the prediction of sugar content on the TF 2014 generalization set by the DT model

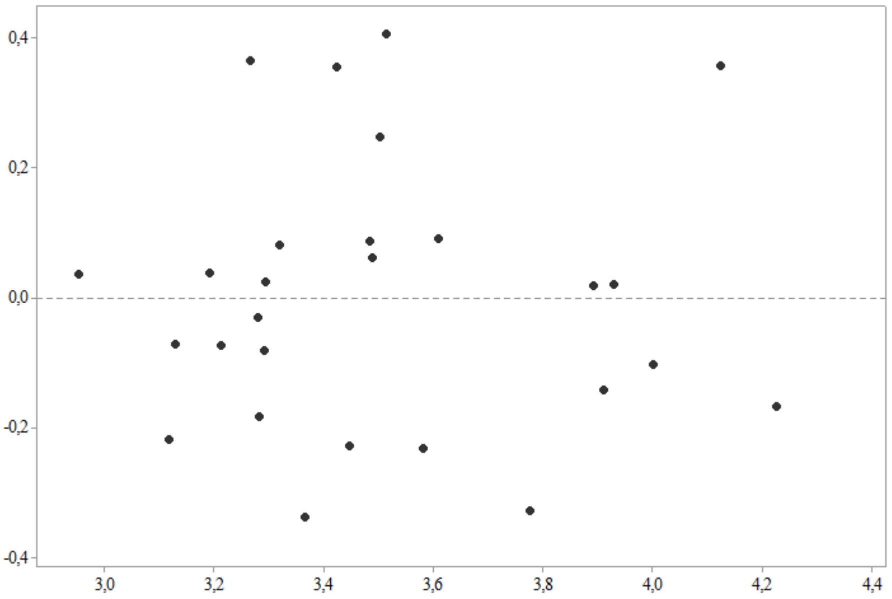


APPENDIX T – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TB 2013 generalization set by the DT model

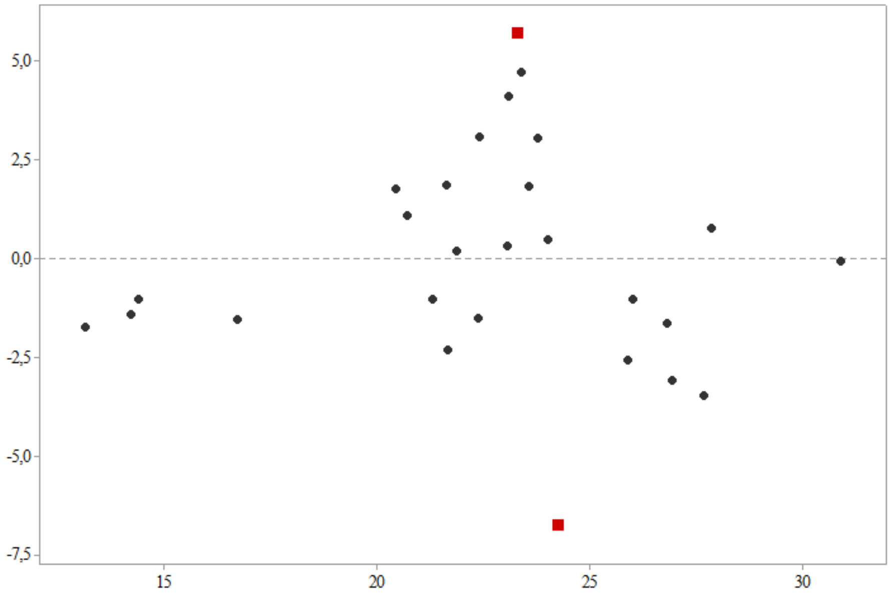
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TB 2013 generalization set by the DT model



Residuals vs fit values plot for the prediction of pH index on the TB 2013 generalization set by the DT model

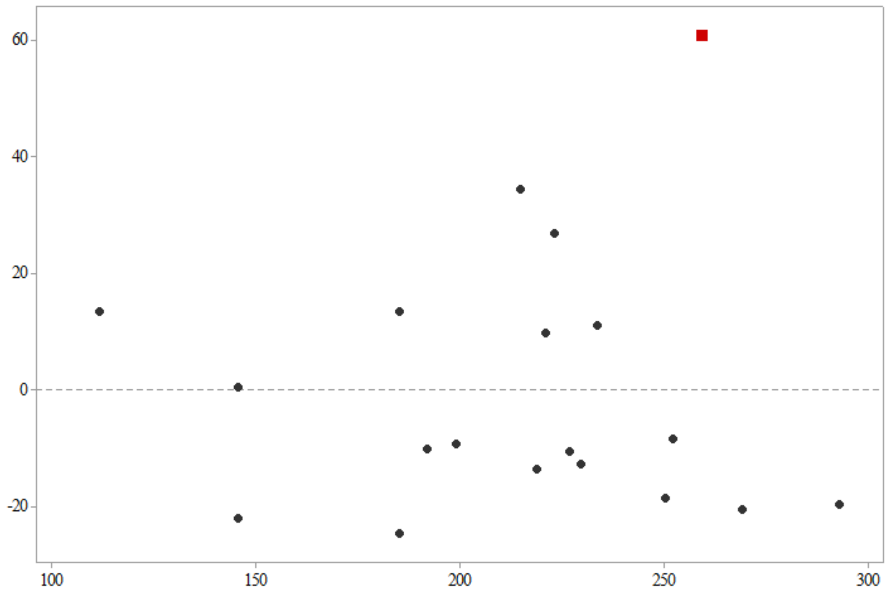


Residuals vs fit values plot for the prediction of sugar content on the TB 2013 generalization set by the DT model

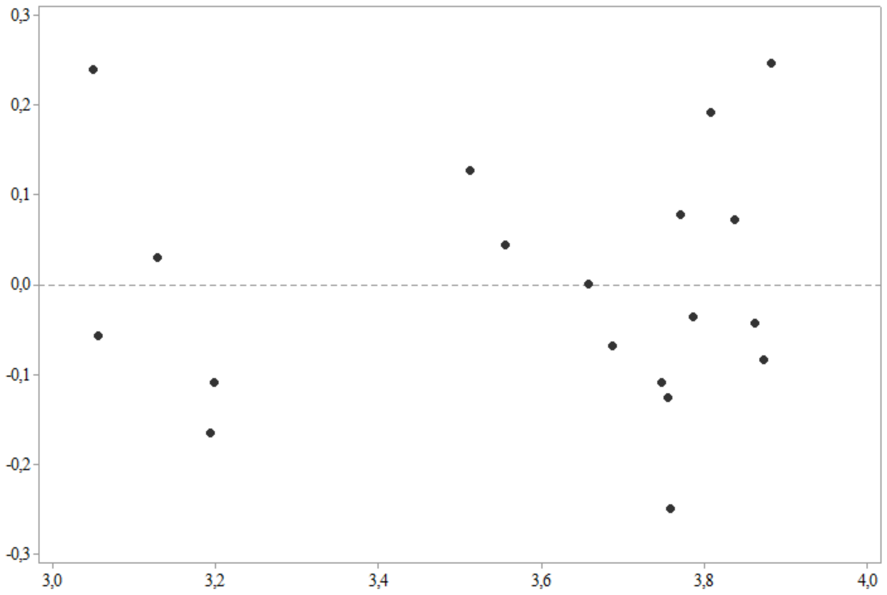


APPENDIX U – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TN 2013 generalization set by the DT model

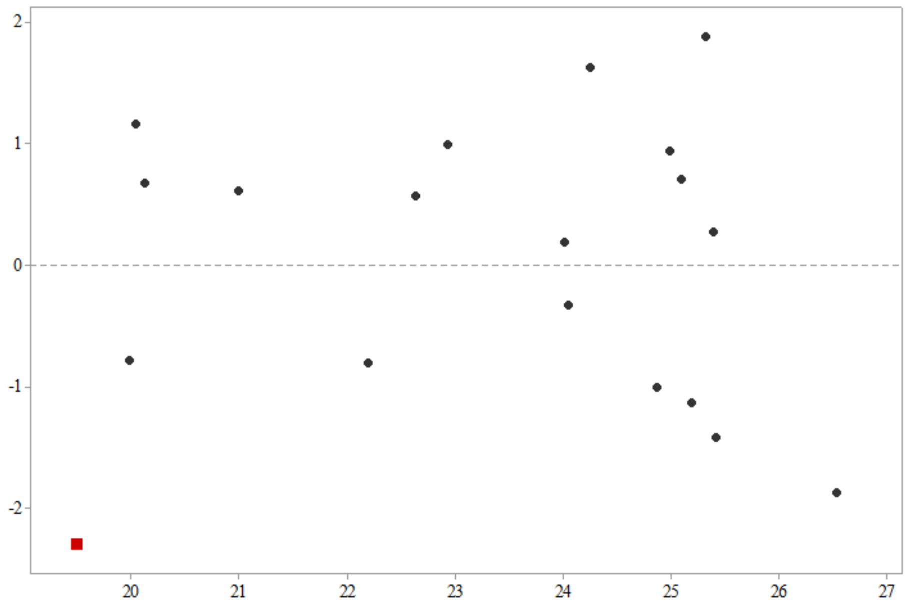
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TN 2013 generalization set by the DT model



Residuals vs fit values plot for the prediction of pH index on the TN 2013 generalization set by the DT model

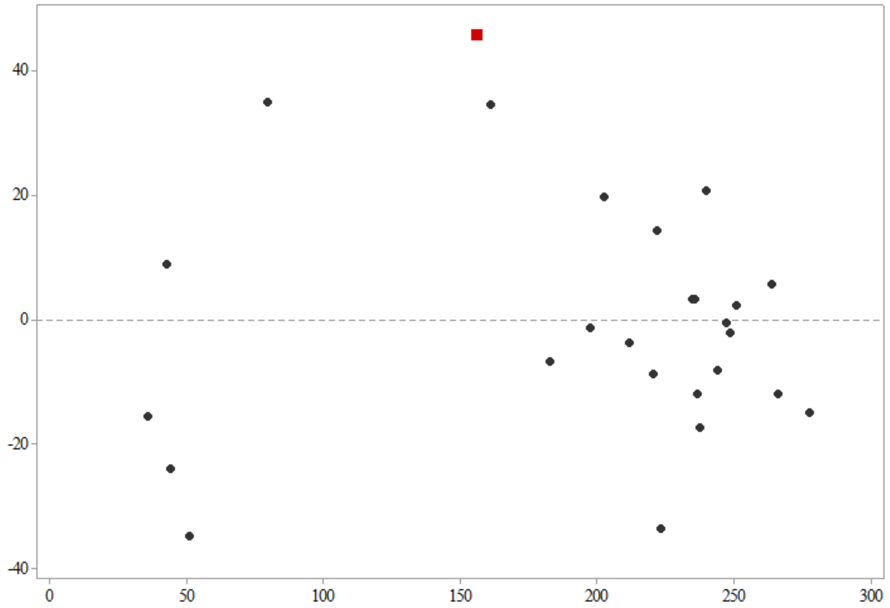


Residuals vs fit values plot for the prediction of sugar content on the TN 2013 generalization set by the DT model

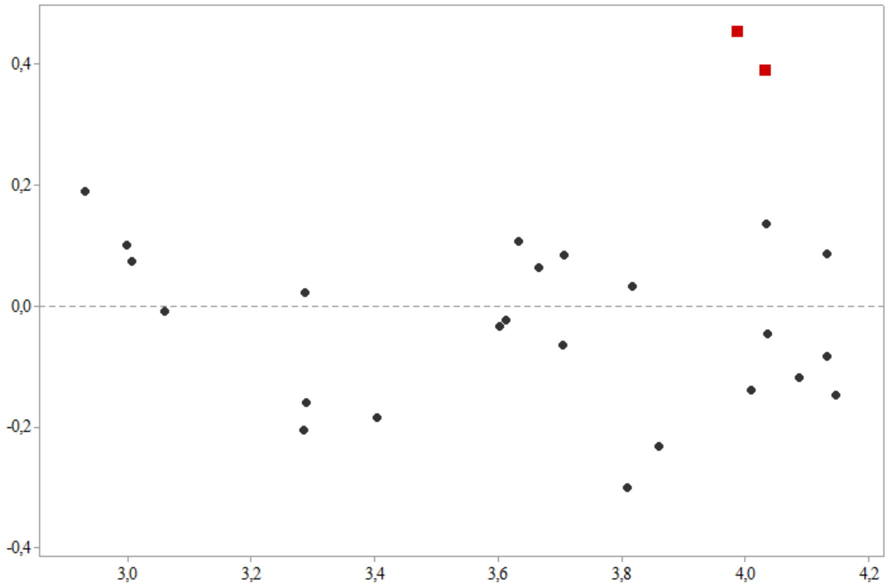


APPENDIX V – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TF 2013 generalization set by the SVR model

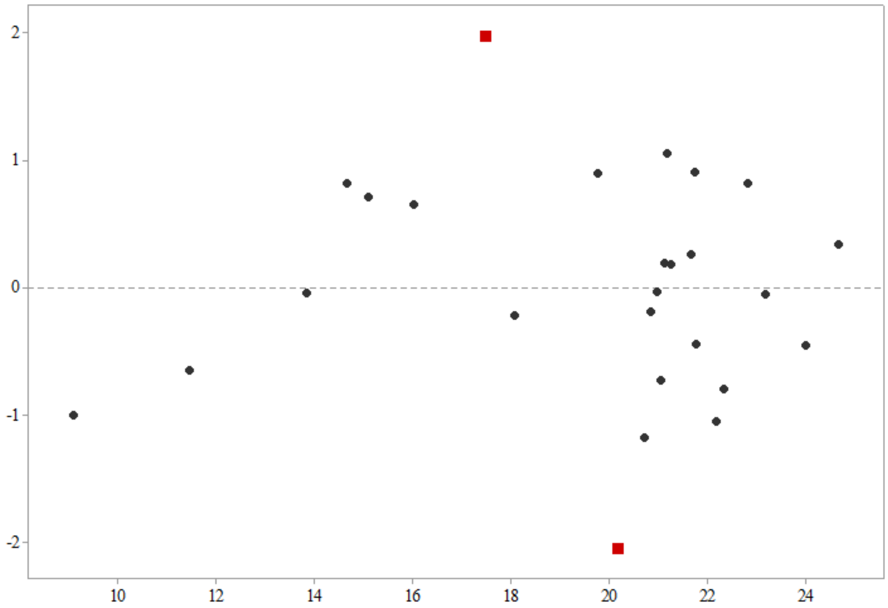
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TF 2013 generalization set by the SVR model



Residuals vs fit values plot for the prediction of pH index on the TF 2013 generalization set by the SVR model

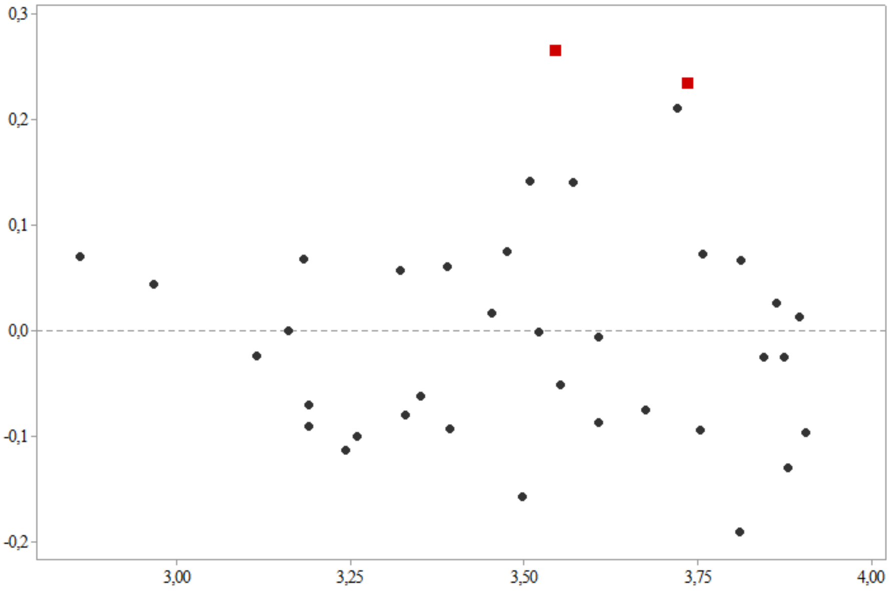


Residuals vs fit values plot for the prediction of sugar content on the TF 2013 generalization set by the SVR model

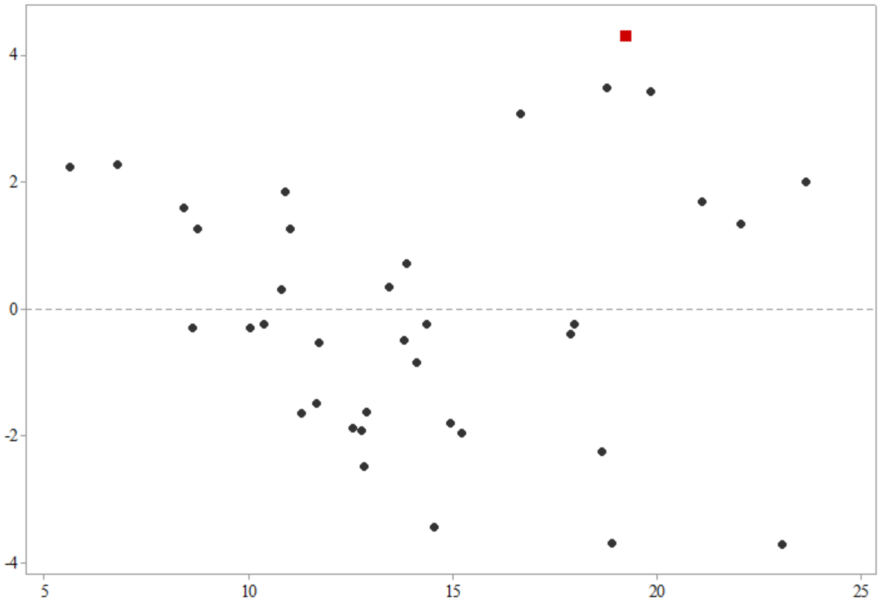


APPENDIX W – Residuals vs fit values plot for the prediction of pH index and sugar content, respectively, on the TF 2014 generalization set by the SVR model

Residuals vs fit values plot for the prediction of pH index on the TF 2014 generalization set by the SVR model

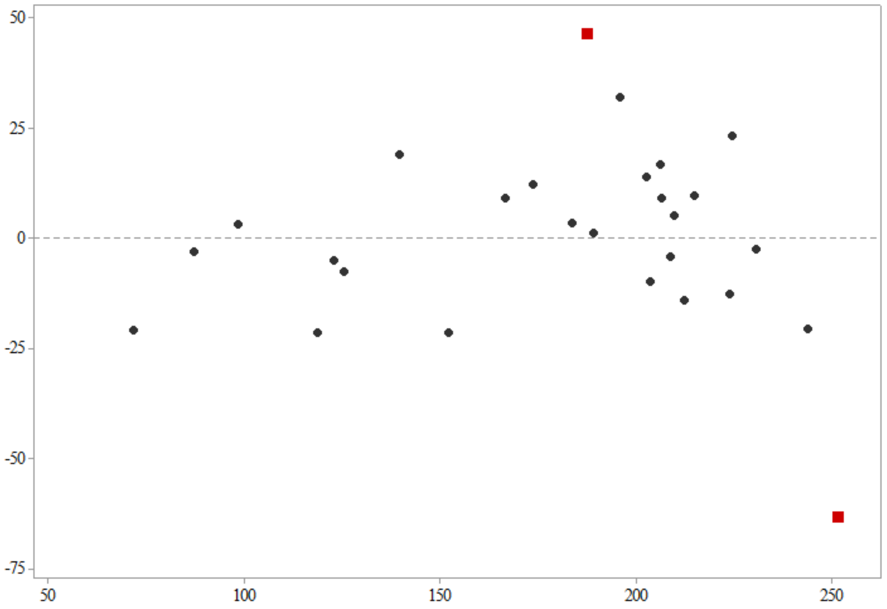


Residuals vs fit values plot for the prediction of sugar content on the TF 2014 generalization set by the SVR model

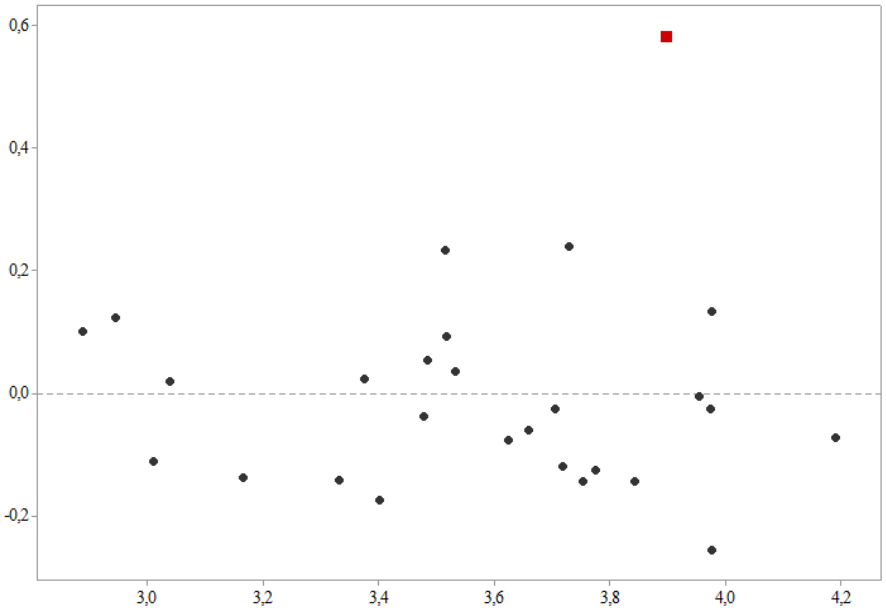


APPENDIX X – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TB 2013 generalization set by the SVR model

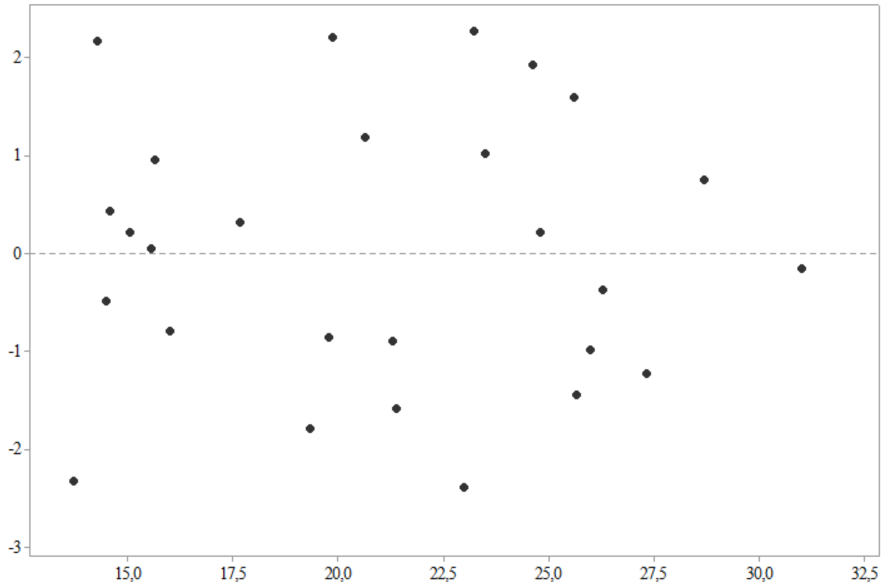
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TB 2013 generalization set by the SVR model



Residuals vs fit values plot for the prediction of pH index on the TB 2013 generalization set by the SVR model

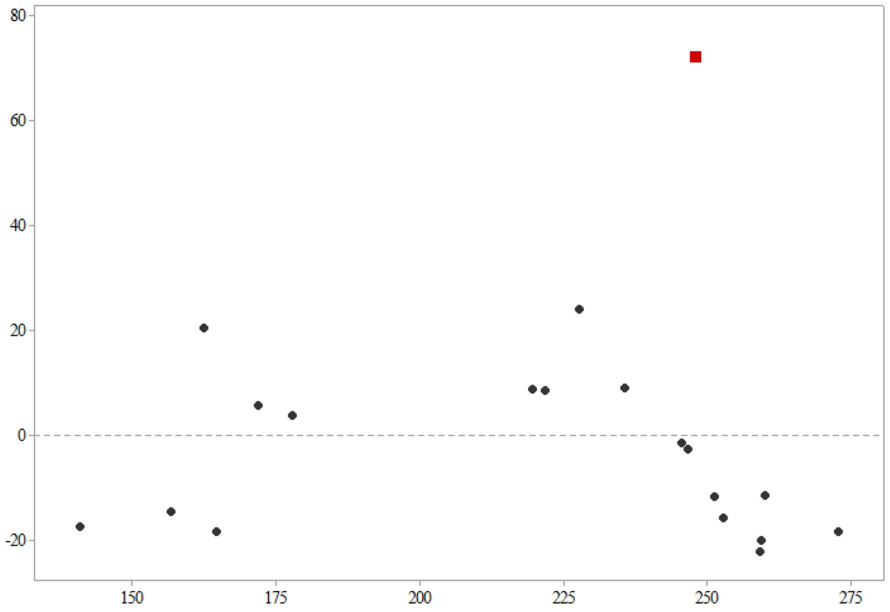


Residuals vs fit values plot for the prediction of sugar content on the TB 2013 generalization set by the SVR model

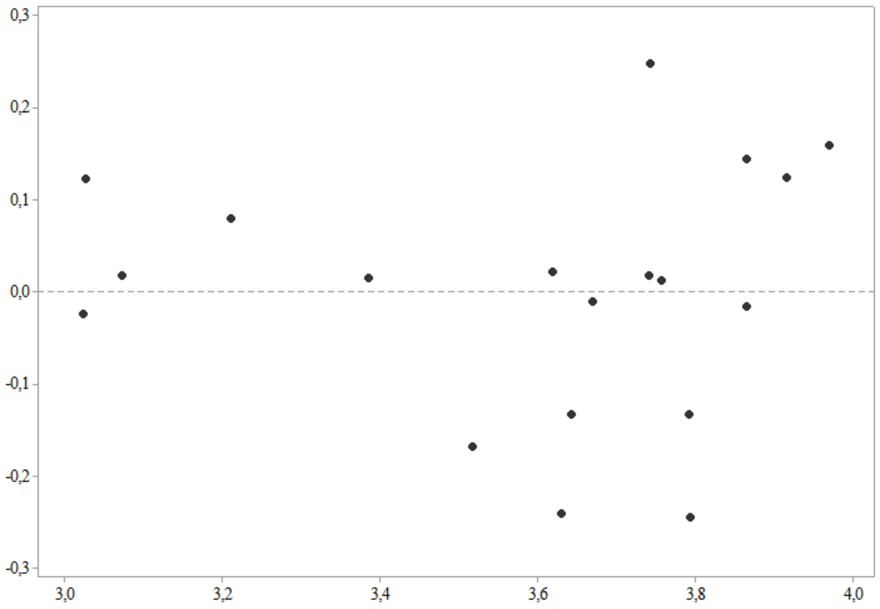


APPENDIX Y – Residuals vs fit values plot for the prediction of anthocyanin concentration, pH index and sugar content, respectively, on the TN 2013 generalization set by the SVR model

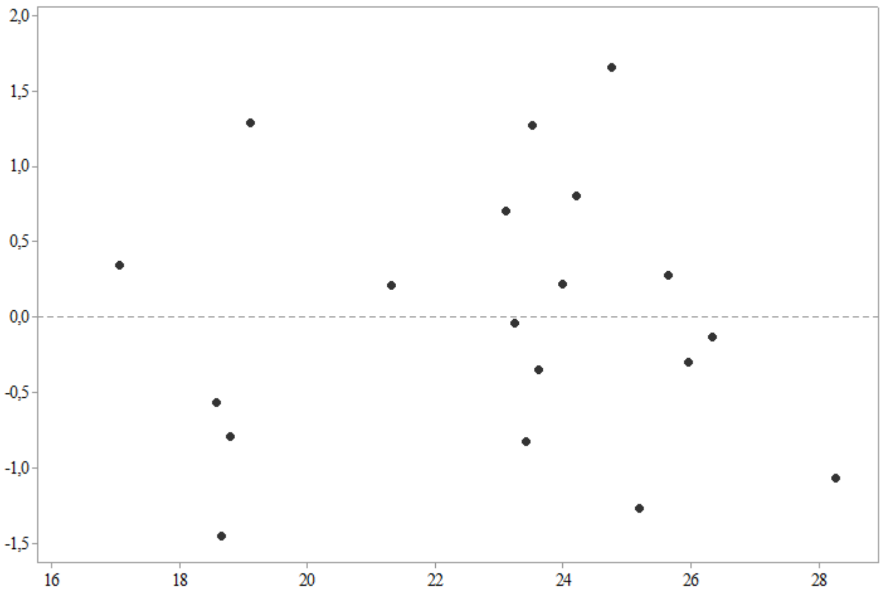
Residuals vs fit values plot for the prediction of anthocyanin concentration on the TN 2013 generalization set by the SVR model



Residuals vs fit values plot for the prediction of pH index on the TN 2013 generalization set by the SVR model



Residuals vs fit values plot for the prediction of sugar content on the TN 2013 generalization set by the SVR model


2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
16
18
20
22
24
26
28