Proceedings of PAMDAS 2025 International conference on Physical Asset Management and Data Science

Thursday, July 17, 2025 - Friday, July 18, 2025

Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Portugal

Organization

Endorsements

ISBN: 978-989-8331-19-9

Proceedings of PAMDAS 2025 International conference on Physical Asset Management and Data Science

This book is available at `https://pamdas.rcm2.pt`

Edited by:
Mateus Mendes, mmendes@isec.pt
José Torres Farinha, tfarinha@isec.pt
Ana Rita Malta, amalta@rcm2.pt
Hugo Raposo, hugo.raposo@isec.pt

RCM^{2+} — Research Centre for Asset Management and Systems Engineering
`https://www.rcm2.pt`

Coimbra, Portugal, 13 July 2025

ISBN: 978-989-8331-19-9

Organization

Executive Commission

Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Guilherme Ramos Pereira, DSPA - Data Science Portuguese Association, Portugal
Beatriz Santana, DSPA - Data Science Portuguese Association, Portugal

Steering Committee

Ana Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Annemieke Meghoe, University of Twente, Netherlands
António Paulino, Polytechnic University of Coimbra, Portugal
Beatriz Santana, DSPA - Data Science Portuguese Association, Portugal
Bernardo Tormos, Universitat Politècnica de València, Spain
David Baglee, Sunderland University, United Kingdom
Fernanda Brito Correia, Polytechnic University of Coimbra, Portugal
Guilherme Ramos Pereira, DSPA - Data Science Portuguese Association, Portugal
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
João Costa, Polytechnic University of Coimbra, Portugal
Jyoti Sinha, University of Manchester, United Kingdom
José Martinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jorge Almeida, Polytechnic University of Coimbra, Portugal
Luís Margalho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Nuno Lavado, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Pascoal Silva, Polytechnic University of Coimbra, Portugal
Rita Malta, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Wojciech Golebiowski, University of Life Sciences in Lublin, Poland

Scientific Committee

Alexandre Martins, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ana Camanho, University of Porto, Portugal
A. Paulo Coimbra, University of Coimbra, Portugal
Akbayan Bekarystankyzy, Narxoz University, Kazakhstan
Ana Alves, Polytechnic University of Coimbra, Portugal
Ana Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Anabela Gomes, Polytechnic University of Coimbra, Portugal
Annemieke Meghoe, University of Twente, Netherlands
António Paulino, Polytechnic University of Coimbra, Portugal
Bagdat Yagaliyeva, Satbayev University, Kazakhstan
Balduíno Mateus, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Batyrkhan Omarov, Al Farabi University, Kazakhstan
Beatriz Fidalgo, Polytechnic University of Coimbra, Portugal
Bernardo Tormos, Universitat Politècnica de València, Spain
Bruno Oliveira, Railway Competence Center, Portugal
Carlos Pereira, Polytechnic University of Coimbra, Portugal
Dammika Seneviratne, Tecnalia Research and Innovation, Spain
David Baglee, Sunderland University, United Kingdom
Felippe Souza, University of Beira Interior, Portugal
Filipe Amaral, Polytechnic University of Coimbra, Portugal
Felipe de Miguel Diez, University of Freiburg, Germany
Fernanda Brito Correia, Polytechnic University of Coimbra, Portugal
Fernanda Coutinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Gulfarida Tulemissova, Suleyman Demirel University, Kazakhstan
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jânio Monteiro, University of Algarve, Portugal
Janet Lin, Lulea Technical University, Sweden
João Barata, University of Coimbra, Portugal
João Costa, Polytechnic University of Coimbra, Portugal
João Durães, Polytechnic University of Coimbra, Portugal
João Pombo, Huddersfield University, United Kingdom
João Rodrigues, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jorge Almeida, Polytechnic University of Coimbra, Portugal
Jorge Barreiros, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
José Martinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
José Sobral, Lisbon Polytechnic, Portugal
Jyoti Sinha, University of Manchester, United Kingdom
Kiumars Teymourian, Lulea Technical University, Sweden
Leonor Melo, Polytechnic University of Coimbra, Portugal
Luís Andrade Ferreira, Railway Competence Center, Portugal
Luís Margalho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal Luís Oliveira, Tomar Polytechnic University, Portugal
Lyazzat Atimtayeva, Suleyman Demirel University, Kazakhstan

Manuel M. Crisóstomo, Institute of Systems and Robotics, Portugal
Marcin Kaminski, Lodz University of Technology, Poland
Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Miguel Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Moldir Kuatova, Al Farabi University, Kazakhstan
Nuno Cid Martins, Polytechnic University of Coimbra, Portugal
Nuno Lavado, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Pascoal Silva, Polytechnic University of Coimbra, Portugal
Philippe Mathieu, University of Lille, France
Rafael Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ramin Karim, Lulea Technical University, Sweden
Raúl Salas, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ricardo Mateus, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Rita Malta, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Sebastião Pais, University of Beira Interior, Portugal
Simão Paredes, Polytechnic University of Coimbra, Portugal
Telmo Silva, University of Aveiro, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Wojciech Golebiowski, University of Life Sciences in Lublin, Poland

Contents

Preface

As technology evolves, industries become more and more dependent on data for decision-making. Concepts such as the Internet of Things become more prevalent in modern life and industrial settings. Sensors provide data which need to be analysed. Analytics algorithms provide insights not only into the past and present of the systems and processes, but also highlight trends and probable future directions.

PAMDAS 2025 international congress took place in Coimbra, days 17 and 18 of July 2025, under the lemma "When Intelligence Enhances Asset Management". The goal was to bridge the gap between industry and data science. While modern industry needs powerful algorithms and data science for enhanced decision-making, data science researchers also benefit from industrial connections where they can gather real data and find qualified research partners and customers.

This congress attracted speakers and authors from different continents, interested in Physical Asset Management and Data Science. Keynote speakers Prof. Diego Galar, from the Lulea Technical University, Sweden, and Prof. Celso Azevedo, from AssetsMan, France, contributed with their experience and wisdom to the the event.

The proceedings are now made available to the public under open access, for use and proper citation.

The editors,

Mateus Mendes
Torres Farinha
Ana Rita Malta
Hugo Raposo

Chapter 27

Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning

Contribution #52
Cite as:
Filipe Carvalho, Paula Couceiro, Mateus Mendes, Pascoal Silva. Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-52,
  title ={Analysis and Modeling of Phenolic Compounds
     and Antioxidant Activity of two Banana Varieties
     Using Machine Learning},
  author={Filipe Carvalho and  Paula Couceiro and
     Mateus Mendes and  Pascoal Silva},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning

Filipe Carvalho[1,*], Paula Couceiro[1,*], Raquel Guiné[2], Pascoal Silva[1,3], Mateus Mendes[1,4,5]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[2] Polytechnic University of Viseu
[3] CMUC-Center for Mathematics, University of Coimbra
[4] RCM^{2+} Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[5] Institute of Systems and Robotics,
Department of Electrical and Computer Engineering, University of Coimbra,
3030-290 Coimbra, Portugal
[*] Corresponding author
{a2021148720,a2018056493, pascals, mmendes}@isec.pt, raquelguine@esav.ipv.pt

Abstract

Phenolic compounds (PC) and antioxidant activity (AA) are very important for human well-being. Therefore, quantification and prediction of PC and AA are crucial for food production and distribution chains. In the present study, a dataset with experimental measurement results was analysed. Machine learning models MLPRegressor and Random Forest models were used to predict PC and AA based on input variables: variety, drying state, extract type and extract order. The importance of the input variables was evaluated, where for the MLPRegressor the variable with the largest weight was the drying state and for the Random Forest it was the order of the extract, both being strongly correlated with the phenolic compounds and the antioxidant activity. Performance metrics were applied to evaluate the results obtained with both models, where it was concluded that MLPRegressor obtained lower values than Random Forest for RMSE, MAE, MAPE and MPE, and higher values for R^2, which proved that MLPRegressor performs better.

Keywords: Phenolic compounds; Antioxidant activity; Banana properties; Machine learning models; MLPRegressor; Random Forest

1 Introduction

Polyphenols, widely present in foods from the Mediterranean diet, have been recognised as essential bioactive compounds in the prevention of chronic non-communicable diseases such as cardiovascular diseases, type 2 diabetes and some types of cancer [1]. Their importance comes largely from their high antioxidant activity, which neutralises the free radicals responsible for oxidative

stress - one of the main mechanisms associated with cell degeneration and the development of various pathologies [2]. Scalbert *et al.* [3] also show that the beneficial effects of polyphenols go beyond simply neutralising free radicals, acting on molecular pathways related to inflammation and metabolic regulation. However, the concentration and effectiveness of these compounds can be significantly affected by agronomic and technological factors. Tomas-Barberán & Espín [4] and Guiné *et al.* [5] show that conditions such as the degree of ripeness, the type of crop and drying methods directly influence the phenolic content of foods. It is therefore clear that both the regular consumption of foods rich in polyphenols and care in the production and processing processes are fundamental to maximise their benefits in promoting human health.

The present work aims aim to perform exploratory data analysis over a dataset of results of laboratory analysis of two banana varieties. Two machine learning models, MLPRegressor and Random Forest, were also deployed to predict the phenolic compounds and antioxidant activity present in the two banana varieties. The study aims to understand which of the input variables has the most influence on the results obtained and then which model achieves the best performance.

The following sections present the state of the art and explain the methodology used, the models chosen and their respective architectures. The models will then be evaluated using performance metrics, concluding which is the best model and which variable has the greatest influence on phenolic compounds and antioxidant activity. The results will be presented in the respective section and discussed in the following section. Finally, the last section will present some conclusions.

2 State of the art

According to Manach *et al.* [1], polyphenols, in addition to being widely present in the Mediterranean diet, play an important role in the prevention of chronic non-communicable diseases such as cardiovascular disease, type 2 diabetes and certain neoplasms.

Lobo *et al.* [2] emphasise that oxidative stress, caused by an imbalance between the production of free radicals and the body's antioxidant capacity, is directly associated with various chronic diseases such as cancer, diabetes, cardiovascular and neurodegenerative diseases. The authors point out that dietary antioxidants, especially bioactive compounds such as polyphenols, play a fundamental role in neutralising these free radicals, contributing to the prevention of these pathologies. Functional foods rich in natural antioxidants have therefore emerged as important allies in health promotion.

Scalbert *et al.* [3] emphasise that polyphenols exert significant antioxidant activity, but also act on cellular pathways that regulate inflammation, metabolism and gene expression. The authors argue that the beneficial effects of polyphenols go beyond neutralising free radicals and involve complex molecular mechanisms.

Tomas-Barberán and Espín [4] analyse how the phenolic compounds present in fruit and vegetables are influenced by factors such as ripeness, type of cultivation, climatic conditions and agricultural practices. The authors emphasise that these elements directly affect the quantity and profile of polyphenols in foods, with relevant implications for their nutritional and functional quality.

Guiné *et al.* [5] studied the impact of different drying conditions on the content of total phenolic compounds and antioxidant activity in bananas from different cultivars. A model using classical feed-forward artificial neural networks was developed. They conclude that air drying had the greatest impact on the final results. Another conclusion was that total phenolic compounds and antioxidant activity could be predicted with great precision from the model built.

Mongi *et al.* [6] investigated the effect of solar drying on vegetables and its relationship with the amount of phenolic compounds and antioxidants present in them. They concluded that drying does significantly affect the results obtained.

Sarpong *et al.* [7] also carried out a study on bananas, but on the influence of convention drying with air and controlled humidity on the production of bioactive compounds and antioxidant degradation in dehydrated slices. These authors concluded that higher drying temperatures allow for greater nutrient retention.

3 Materials and methods

This section presents the methodology used to carry out the study and the models that will be used. It also presents the dataset and an exoloratory analysis of it, as well as the data processing that had to be carried out.

3.1 Methodology

The present study followed the CRISP-DM methodology. CRISP-DM (Cross-Industry Standard Process for Data Mining) is a standard methodology for conducting data mining projects. It organises the process into six main phases: business understanding, data understanding, prepare data, model data, results and deployment, promoting an iterative and structured approach. It is widely used because it can be adapted to different sectors and types of data.

For the modeling phase, two machine learning models were used: MLPRegressor (Multi-Layer Perceptron Regressor) and Random Forest Regressor. Both are widely used in regression tasks, where they normally achieve good performance predicting continuous variables. MLPRegressor is based on artificial neural networks and is capable of modelling complex, non-linear relationships between variables, although it is less interpretable. Random Forest, on the other hand, is an ensemble learning model that uses multiple decision trees to obtain more stable and robust predictions, and is especially effective with noisy and high-dimensional data. Two main differences between the two are the type of structure (neural network vs. trees) and the level of interpretability, with Random Forest generally being more explainable. Both models are available in the scikit-learn library, one of the most popular Python libraries for classical machine learning[1].

3.2 Dataset

The dataset consisted of 288 results of laboratory analysis of bananas of two different cultivars. Table 1 describe the study's input variables—four variables, all categorical. The first variable is the variety of banana, and the study contained two varieties of banana. The second variable is the drying state, as the bananas could have four different drying states. The third variable is the type of extract used, of which there were two types, and the fourth variable is the order of the extract, of which there were three possible orders.

Table 1: List of initial input variables and their possible values.

Variety	Madeira, Costa Rica
Drying state	Fresh, Dried at 50 ^{o}C, Dried at 70 ^{o}C, Freeze-dried
Extract type	Methanol, Acetone
Extract order	1, 2, 3

[1] https://scikit-learn.org/stable/index.html (last checked on 2025-06-02).

The dataset under study therefore has the variables mentioned above as inputs and total phenols and antioxidant activity as outputs. For each set of input conditions, 6 repetitions were made, and in some cases no results were obtained.

Figure 1 shows all the data from the dataset. The phenolic compounds obtained with methanol and acetone as a function of drying state are shown on the left, and the antioxidant activity with methanol and acetone as a function of drying state is shown on the right. The graph also shows a numbering from 1 to 3, which corresponds to the order of the extract for each result obtained. It can be seen that there are anomalies in the results, most likely due to the variability inherent to the laboratorial nature of the experiments.

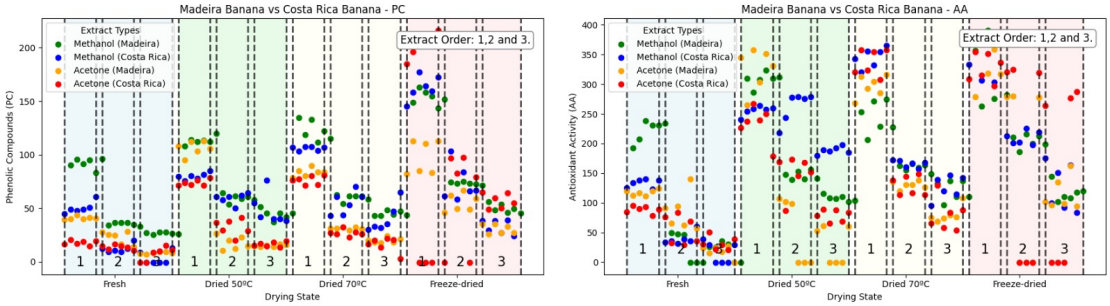

Figure 1: Graphical representation of total phenols (left) and antioxidant activity (right) of bananas from Madeira and Costa Rican bananas with methanol and acetone as a function of drying state.

Following on from the data analysis, it was clear from the previous graphs that there was a lack of data that was assumed to be '-1' and that there were areas that were difficult to interpret, i.e. there may or may not be outliers in these areas. To check this, a more detailed analysis was carried out using boxplot graphs.

When carrying out the study for the Madeira banana, two boxplots were constructed, one for methanol and the other for acetone in relation to total phenols, on the left. It was possible to see that there is a dispersion of data for both extracts, although slightly less for acetone and there were no outliers, as can be seen in figure 2. Figure 2 also shows the boxplot of antioxidant activity as a function of the type of extract used on the Madeira banana, on the right. The conclusions were the same: there is a great deal of dispersion in the data, but there are no outliers. It can also be seen that the average results with methanol, whether for total phenolics or antioxidant activity, are always higher.

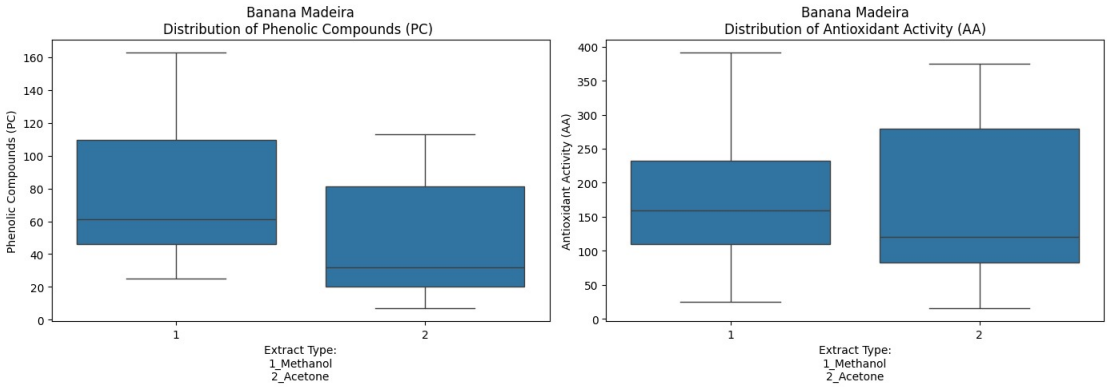

Figure 2: Boxplot of total phenols and antioxidant activity of Madeira bananas as a function of methanol and acetone.

In the case of Costa Rica bananas, the same analysis, in figure 3, distribution of phenolic compounds on the left, immediately revealed the presence of anomalous values, as can be seen in figure 3. These outliers are present in the bloxplot of total phenols, they exist for both methanol and acetone and take on much higher values than the rest. With regard to antioxidant activity, there was a greater dispersion of data, but no outliers. It was also found that the average results were higher in all cases with methanol.

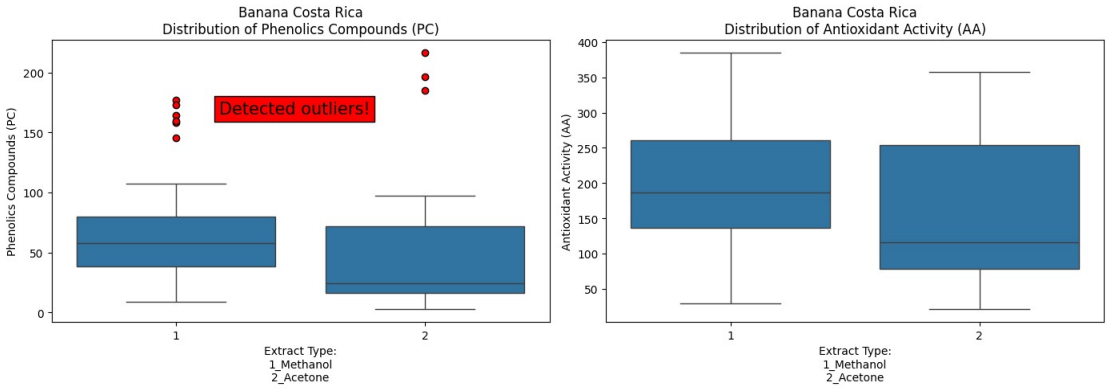

Figure 3: Boxplot of total phenols and antioxidant activity of Costa Rican bananas as a function of methanol and acetone

.

So, after analysing the dataset, all the lines with null or negative values were eliminated because they made no sense in the dataset. The initial dataset, which contained 288 samples, was reduced to a total of 257 viable samples for the study. Tables 2 and 3 summarise the dataset after data cleaning.

Tables 2 and 3 present a description of the data relating to the samples that will enter the study for PC and AA, respectively, for two sample varieties: Madeira and Costa Rica. Both tables include the number of valid samples, the average, the minimum and maximum values, and

the standard deviation. In terms of the number of samples, both varieties are balanced, with 129 samples from Madeira and 128 from Costa Rica.

The Madeira variety has an average PC of 50.51, slightly higher than Costa Rica, which has an average of 44.22. However, the Costa Rican variety has a wider range of values, with a minimum of 2.77 and a maximum of 216.41, while Madeira varies between 6.99 and 162.78. In addition, the standard deviation for Costa Rica (43.28) is higher than for Madeira (38.60), suggesting greater variability in the data for this variety.

For the AA variable, the Costa Rica variety has a higher average (165) compared to Madeira (140), which indicates that, on average, Costa Rica has more antioxidant activity than Madeira. However, both varieties have very similar maximum values (391.18 for Madeira and 384.88 for Costa Rica) and relatively close minimum values (15.59 for Madeira and 21.10 for Costa Rica). The standard deviation is high for both varieties, with 99.94 for Madeira and 102.68 for Costa Rica, reflecting a high dispersion of the data, with slightly greater variability in Costa Rica.

Table 2: Samples valid for output variable PC.

Variety	Nº of samples	Average	MIN	MAX	Standard deviation
Madeira	129	50.51	6.99	162.78	38.60
Costa Rica	128	44.22	2.77	216.41	43.28

Table 3: Samples valid for output variable AA.

Variety	Nº of samples	Average	MIN	MAX	Standard deviation
Madeira	129	140	15.59	391.18	99.94
Costa Rica	128	165	21.10	384.88	102.68

Thus, in figure 4, the data has been represented in a similar way to what was done when analysing the dataset. On the left, the figure shows the phenolic compounds obtained with methanol and acetone as a function of drying state and extract order, and on the right, the same for antioxidant activity. These graphs only show the data after processing and it is clear to see the changes when compared to the graphs in figure 1. With regard to both phenolic compounds and antioxidant activity, for the first drying state, fresh, all the samples for the third order of extract were excluded, and there was a reduction in the number of samples from the second order. For the second drying state, there was a reduction in the number of samples for the second and third order extracts, and for the last drying state, there was a reduction in the number of samples for the first and second order extracts. For the 'dried at 70º' drying state, no samples were removed.

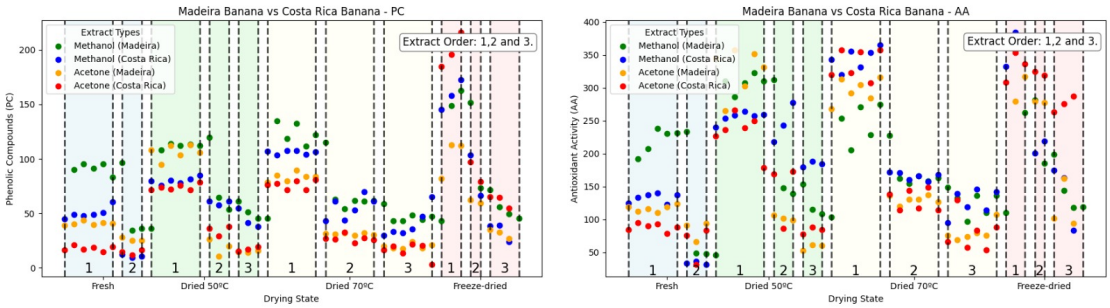

Figure 4: Graphical representation of total phenols (left) and antioxidant activity (right) of bananas from Madeira and Costa Rican bananas with methanol and acetone as a function of drying state, after data processing.

To begin the study, 70% of the data for Madeira bananas and 70% of the data for Costa Rica bananas were selected, with the remaining 30% of data for each variety being kept for testing the models. This division in terms of variety was made to ensure that the model was trained with the same percentage for each variety, thus avoiding the risk of unbalanced distribution of the data.

With the data selected, two forecasting models were built using MLPRegressor and Random Forest, the main objective being to understand which is best applied to the study in question.

4 Results

This section provides an in-depth analysis of the data after it has been processed, with the aim of gaining a better understanding of the dataset. Initially, an exploratory analysis was carried out using a correlation heatmap, which made it possible to assess the strength of the relationships between the numerical variables. Next came the modelling stage, where we began by checking that the models were learning well so that we could then check the importance of each variable for each model. The models were then optimised to see which architecture was best. Once the architecture had been chosen, the models could finally be trained in order to obtain results.

4.1 Exploratory data analysis

Tables 2 and 3 show that Madeira bananas, on average, contain more PC, while Costa Rica bananas are richer in AA. Bearing in mind that the experiment and the comparison between results were carried out keeping the inputs the same and varying only the variety of banana, if for the same inputs the banana from Madeira contains more PC and the banana from Costa Rica contains more AA, the only reason for this is the quality of the banana.

The heatmap in figure 5 shows the linear relationship between the independent and dependent variables in the data set. The values vary between -1 and 1, where positive values indicate a direct correlation and negative values indicate an inverse correlation.

The dependent variables PC and AA show a strong positive correlation with each other (0.85), which suggests that these two variables are strongly related: as one increases, the other tends to increase as well. This strong correlation is justified, having in mind that the phenolic compounds are responsible for a great deal of the antioxodantn in foods where they are present. Among the predictor variables, the `Extract_Order` variable stands out, with a significant negative correlation

with PC (-0.65) and with AA (-0.66). This indicates that as the extraction order increases, the PC and AA values tend to decrease. This negative relationship suggests that extracts obtained at more advanced stages of the process perform less well in terms of response variables, which can be explained by a decrease in the concentration of active compounds over the course of consecutive extractions.

The `State` variable also shows a moderate positive correlation with PC (0.44) and with AA (0.48). This indicates that the physical state of the material (e.g. fresh or dry) positively influences both output variables, and it is likely that less processed products, as those in the fresh state, associated with better results.

On the other hand, the `Variety` and `Extract_Type` variables show weak or practically zero correlations with PC and AA, suggesting that, in isolation, they do not have much of a direct linear impact on the output variables.

In short, analysis of the correlation matrix shows that `Extract_Order` and `State` are the predictor variables with the greatest relevance in explaining the behaviour of the dependent variables, and are therefore natural candidates to be given greater weight during predictive modelling. The strong correlation between PC and AA also suggests that multivariate models can benefit from this relationship to improve predictive capacity.

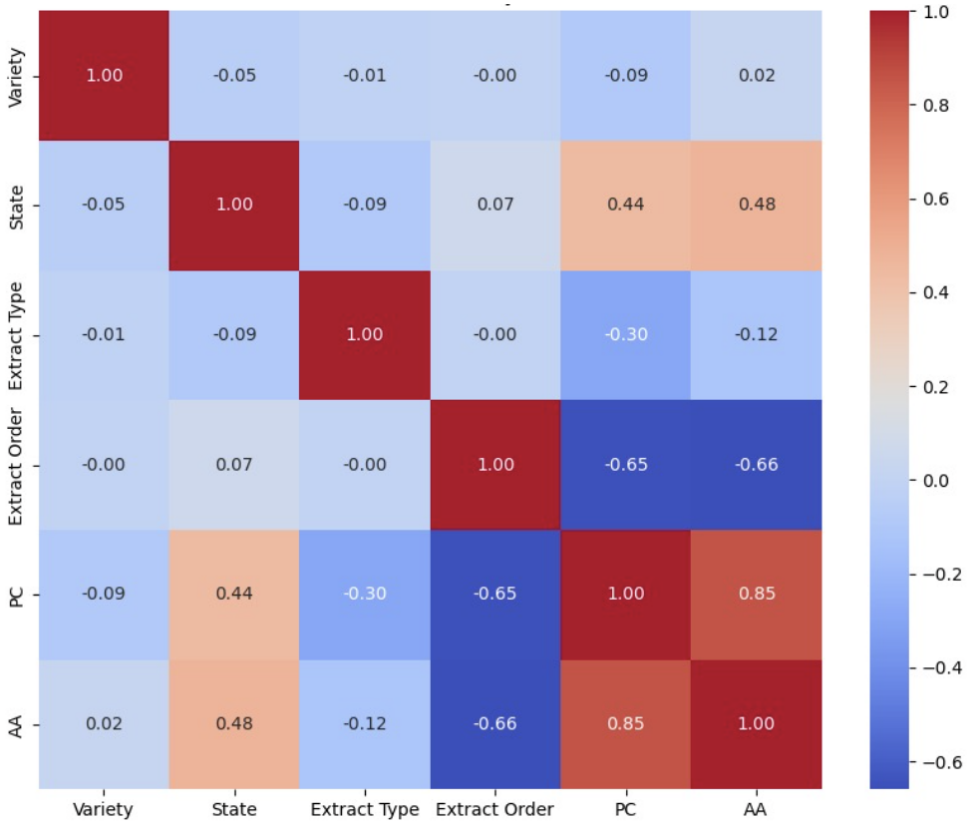

Figure 5: Heatmap of the correlations between inputs and outputs.

4.2 Modeling

This sub-section explains how the learning of the models was verified, the architecture chosen and the results obtained.

4.2.1 Input Weight analysis

MLPRegressor (Multi-Layer Perceptron Regressor) and Random Forest are two types of regression models used in machine learning, but with different approaches. MLPRegressor is an artificial neural network made up of layers of neurons connected to each other, capable of learning complex, non-linear relationships between variables, being sensitive to the choice of hyperparameters and requiring data normalisation. Random Forest is an ensemble of several decision trees trained with random subsets of the data and variables, combining their results to improve robustness and reduce the risk of overfitting; it is easier to interpret and usually works well with fewer adjustments.

Both models have different sets of parameters that must be adjusted to optimise their performance. In MLPRegressor, the main parameters include `hidden_layer_sizes`, which defines the structure of the network's hidden layers, activation, which specifies the activation function (such as 'relu' or 'tanh'), solver, which determines the optimisation algorithm (such as 'adam' or 'sgd'), as well as alpha, `learning_rate_init`, `max_iter` and `early_stopping`, which control regularisation, learning and convergence. In RandomForestRegressor, parameters such as `n_estimators` (number of trees), `max_depth` (maximum tree depth), `min_samples_split` and `min_samples_leaf`, which control tree growth, as well as `max_features`, which determines how many variables to consider in each split, and bootstrap and `oob_score`, which affect how samples are used. These parameters directly influence the balance between bias and variance in the models.

It was then decided to use MLPRegressor with 2 hidden layers with 200 neurons in the first and 100 in the second and Randon Forest with 500 trees.

To see if the models were learning well during training, two different evaluation methods were applied, one for each model. In the case of MLPRegressor, the loss curve was applied. The loss curve obtained, figure 6, during the training of the MLPRegressor model shows a sharp reduction in loss in the first few epochs, followed by a more gradual decrease until it stabilises. This behaviour is typical of an effective learning process, indicating that the model is managing to adjust its parameters to minimise the error between predictions and actual values. The absence of sudden oscillations or increases in loss suggests that the training was stable and there was no obvious overfitting at this stage. Thus, the loss curve confirms that the model is learning progressively and consistently over the epochs.

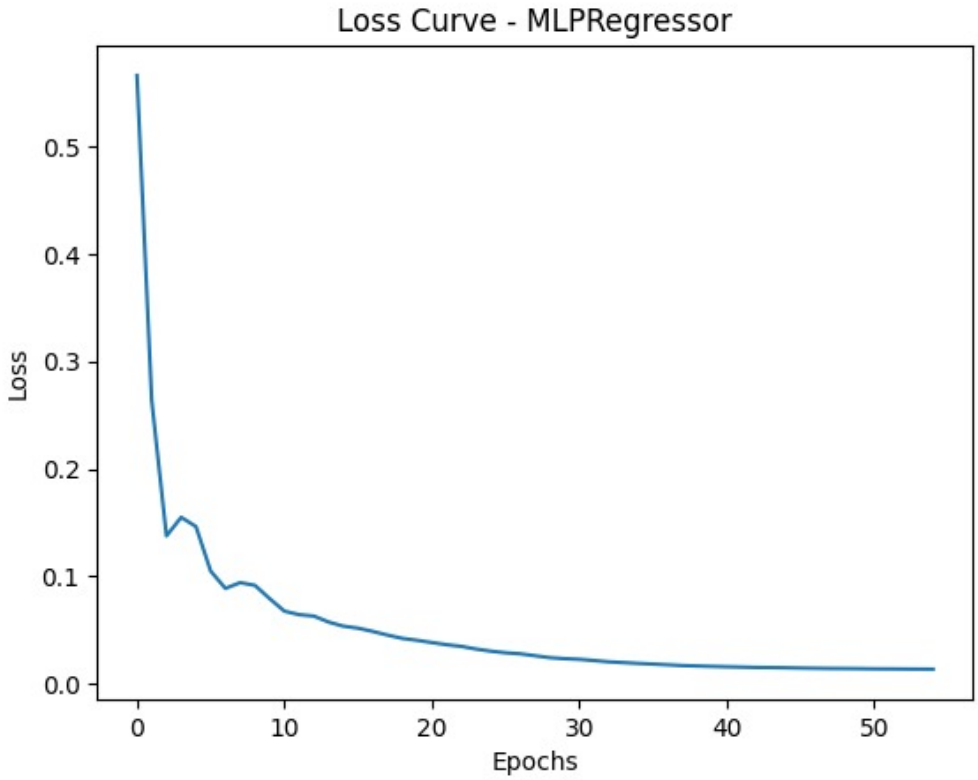

Figure 6: Loss curve.

Although the Random Forest model is not trained by epochs like neural networks, it is possible to generate a curve analogous to the loss curve by observing how performance varies with the increase in the number of trees in the forest. To do this, the model is trained by progressively increasing the `n_estimators` parameter and the desired metric (such as root mean square error or R^2) is evaluated on a validation or test set at each step. The result is a graph showing whether the model continues to improve as more trees are added or whether it reaches saturation point. This approach makes it possible to understand the stability of learning and identify the ideal number of trees for a good balance between performance and computational cost. Figure 7 shows this stabilisation from 100 trees onwards.

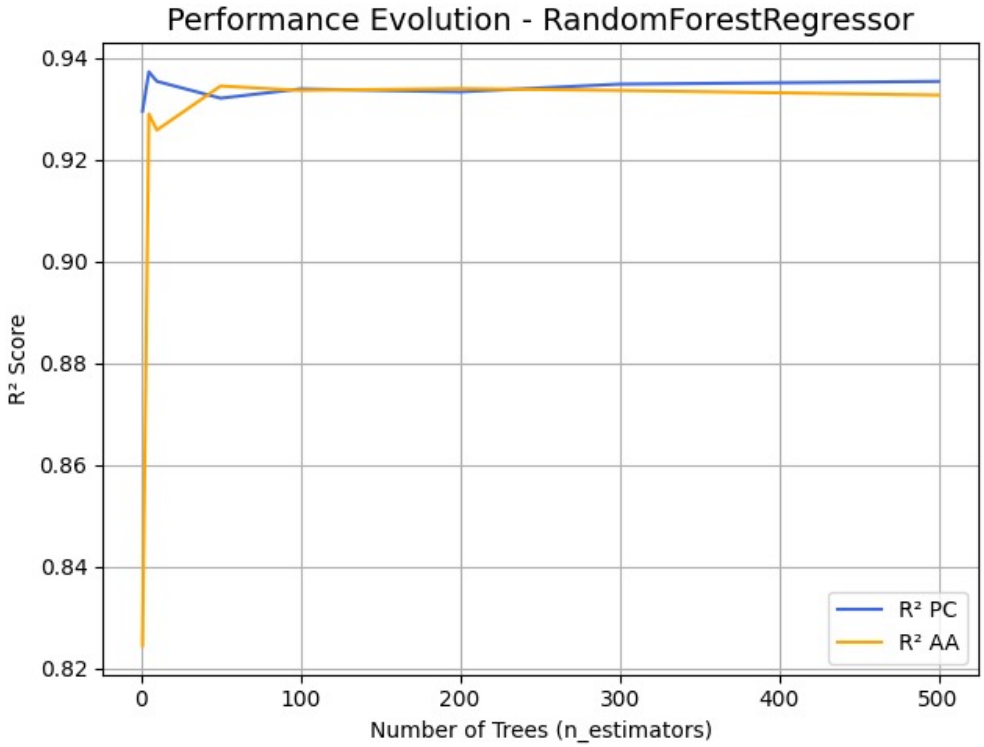

Figure 7: Random Forest performance evaluation.

After applying the models, the importance of each variable was checked. Table 4 shows the weights of each variable for each model. It was concluded that for MLPRegressor the state of drying is more important, while for Random Forest the order of the extract has more influence.

Table 4: Weights of each variable for each model.

Model	Variety	Drying state	Extract type	Extract order
MLPRegressor	0.80	1.00	0.75	0.72
Random Forest	0.05	0.38	0.07	0.49

4.2.2 Model optimization

With the two models chosen, an analysis was carried out to find the best architecture for each one. Figure 8 shows the analysis carried out with several different architectures for MLPRegressor. The analysis relates the R^2 to the number of neurons present in each hidden layer.

The activation function used was the ReLU (Rectified Linear Unit), which is an activation function widely used in artificial neural networks. It works in a simple way: for any input x, the output is x if $x > 0$ and 0 if $x \leq 0$. In other words, it 'zeroes out' all negative values and leaves positive values unchanged. This introduces non-linearity into the model, allowing the network to learn complex patterns, while avoiding problems such as gradient fading that can occur with other functions, such as the sigmoid. Because of its simplicity and computational efficiency, ReLU is a standard choice in many modern deep learning architectures.

Figure 9 shows the same analysis, but this time for Random Forest, where the R^2 is related to the number of decision trees. It was found that the best architecture for MLPRegressor has two hidden layers: the first with 200 neurons and the second with 100. For Random Forest, it was found that up to 50 decision trees there was a difference in the result, but from 100 decision trees onwards there was stability in the R^2, so with 100 or 500 decision trees the model's performance would be the same.

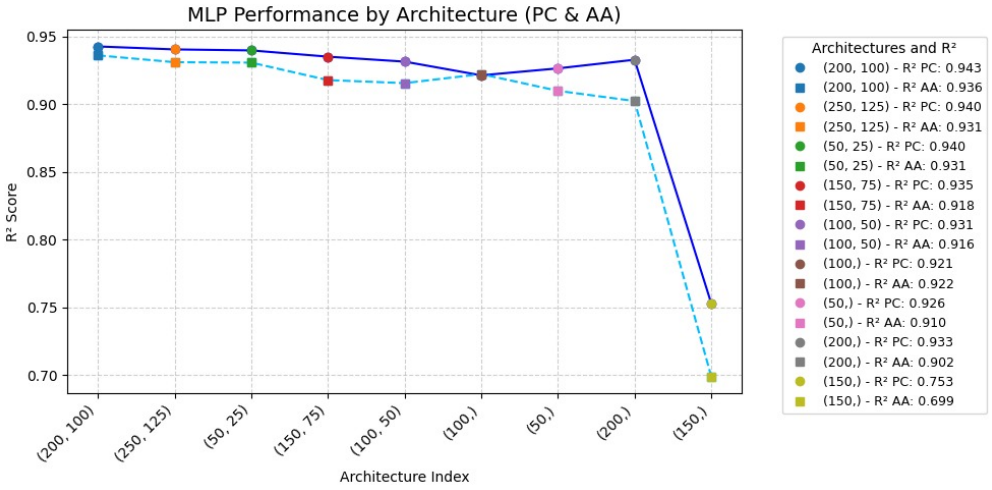

Figure 8: Impact of the MLPRegressor architecture on the value of R^2.

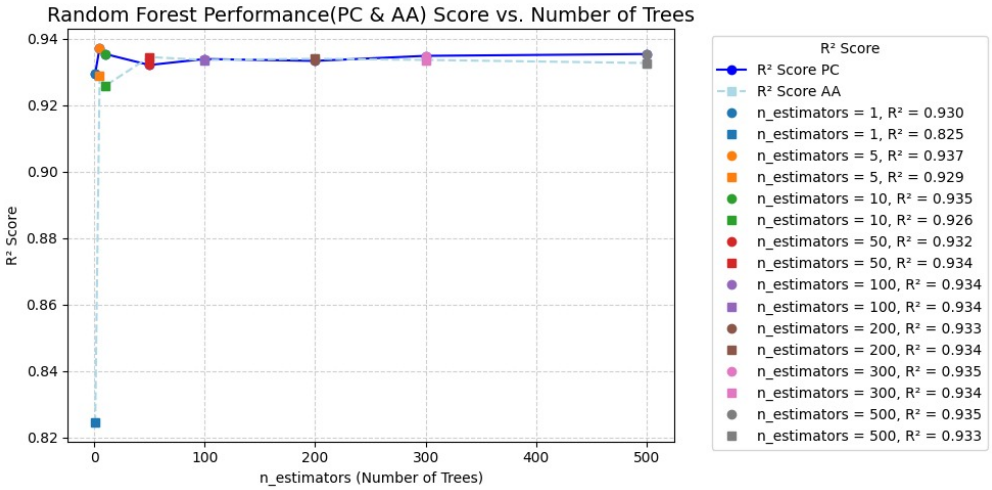

Figure 9: Impact of the Random Forest architecture on the value of R^2.

Although the outputs were predicted simultaneously, the results were obtained individually. Figures 10 and 11 show the results obtained with MLPRegressor and Random Forest, respectively.

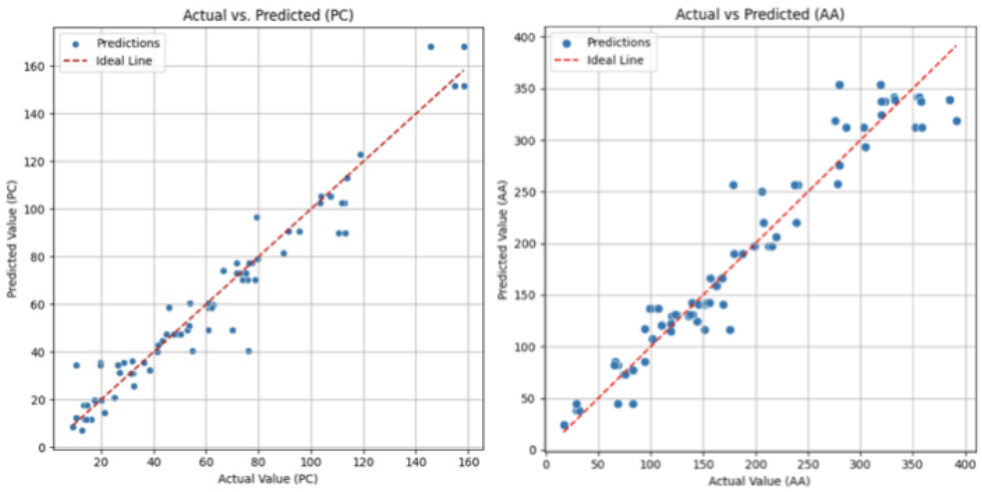

Figure 10: Results obtained with MLPRegressor.

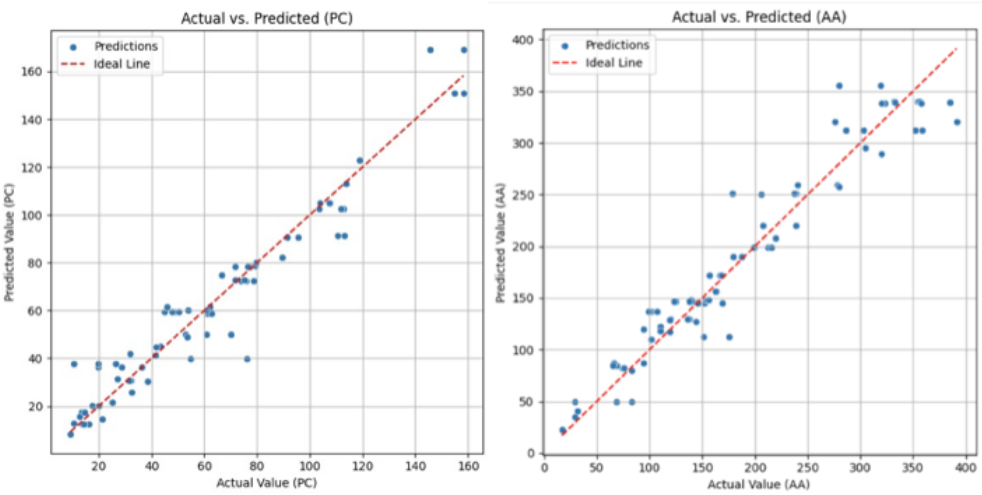

Figure 11: Results obtained with Random Forest.

5 Discussion

In the previous section, studying the weight of the variables, it was found that with MLPRegressor the variable that most influences the outputs is the state of drying, while with Random Forest the variable that has the most weight is the order of the extract.

To effectively compare the results obtained in order to understand which model is the most efficient, various performance metrics (PM) were applied to evaluate the results obtained. R^2, RMSE, MAE, MAPE and MPE were the performance metrics used to assess the performance of regression models. The R^2 (coefficient of determination) measures the proportion of the variance in the data explained by the model, ranging from 0 to 1 (the closer to 1, the better).

RMSE (Root Mean Square Error) is the root of the mean square error, penalising larger errors and being useful for understanding the average magnitude of errors. MAE (Mean Absolute Error) calculates the average of the absolute errors and is more robust to outliers than RMSE. MAPE (Mean Absolute Percentage Error) expresses the errors in percentage terms in relation to the real values, facilitating relative interpretation. The MPE (Mean Percentage Error) is also a percentage, but can indicate whether the model tends to overestimate (negative values) or underestimate (positive values), as it maintains the sign of the error.

Tables 5 and 6 show the values obtained for each performance metric applied.

Table 5: Results of the performance metrics applied to both models that predicted PC.

Model	R^2	RMSE	MAE	MAPE	MPE
MLPRegressor	94.26%	8.9477	5.8358	15.14%	229.20%
Random Forest	93.54%	9.4984	6.3836	16.86%	258.98%

Table 6: Results of the performance metrics applied to both models that predicted AA.

Model	R^2	RMSE	MAE	MAPE	MPE
MLPRegressor	93.61%	24.6275	17.7588	12.12%	54.46%
Random Forest	93.27%	25.2730	19.0156	12.94%	70.89%

Although the results are very similar, it can be seen that MLPRegressor always shows lower values than Random Forest, except for R^2, which is higher, showing that MLPRegressor explains more of the variance in the data.

6 Conclusion

The study carried out aimed to study the phenolic compounds and antioxidant activity, taking into account their importance for humans, present in two varieties of bananas.

An analysis of the dataset provided was carried out, where it was verified for both models used, MLPRegressor and Random Forest, which input was most important. The drying state and the order of the extract are the variables that have larger influence on both AA and PC. The order of the extract being a lab variable is only important for experimental results, but the drying state shows that food preservation may have a significant impact on the PC and AA.

After this analysis, through performance metrics, it was verified that MLPRegressor showed better efficiency than Random Forest, although by a small margin, to predict PC and AA based on the input variables. Thus, both models can be used as predictors with a good degree of confidence.

References

[1] Claudine Manach; Augustin Scalbert; Christine Morand; Christian Rémésy and Liliana Jiménez. Polyphenols: food sources and bioavailability. *The American Journal of Clinical Nutrition*, 79(5):727–747, 2004.

[2] V. Lobo; A. Patil; A. Phatak and N. Chandra. Free radicals, antioxidants and functional foods: Impact on human health. *Pharmacognosy Reviews*, 4(8):118–126, 2010.

[3] Augustin Scalbert; Ian T Johnson; and Mike Saltmarsh. Polyphenols: antioxidants and beyond. *The American Journal of Clinical Nutrition*, 121(5):215–217, 2005.

[4] Francisco A Tomás-Barberán and Juan Carlos Espín. Phenolic compounds and related enzymes as determinants of quality in fruits and vegetables. *Journal of the Science of Food and Agriculture*, 81(9):853–876, 2001.

[5] Raquel Guiné; Maria João Barroca; Fernando Gonçalves; Mariana Alves; Solange Oliveira and Mateus Mendes. Aplicação da modelização por redes neuronais ao teor de compostos fenólicos e atividade antioxidante em bananas de diferentes cultivares secadas sob condições distintas. *Lisboa*, 1(1):296–299, 2014.

[6] Richard J. Mongi1; Bernadette K. Ndabikunze1; Trude Wicklund; Lucy M. Chove and Bernard E. Chove. Effect of solar drying methods on total phenolic contents and antioxidant activity of commonly consumed fruits and vegetable (mango, banana, pineapple and tomato) in tanzania. *African Journal of Food Science*, 9(5):291–300, 2015.

[7] Frederick Sarpong; Xiaojie Yu; Cunshan Zhou; Leticia Peace Amenorfe; Junwen Bai1; Bengang Wu and Haile Ma. The kinetics and thermodynamics study of bioactive compounds and antioxidant degradation of dried banana (musa ssp.) slices using controlled humidity convective air drying. *Journal of Food Measurement and Characterization*, 12(3):1935–1946, 2018.