Article

Prediction of Harvest Time of Apple Trees: An RNN-Based Approach

Tiago Boechel [1,*] , **Lucas Micol Policarpo** [1] , **Gabriel de Oliveira Ramos** [1] , **Rodrigo da Rosa Righi** [1] and **Dhananjay Singh** [2]

[1] Applied Computing Graduate Program, Universidade do Vale do Rio dos Sinos, Av. Unisinos, 950, Cristo Rei, São Leopoldo 93022-000, Brazil; lmpolicarpo@unisinos.br (L.M.P.); gdoramos@unisinos.br (G.d.O.R.); rrrighi@unisinos.br (R.d.R.R.)

[2] Department of Electronics Engineering, Hankuk University of Foreign Studies, Yongin 17035, Korea; dan.usn@gmail.com

* Correspondence: tboechel@edu.unisinos.br

Abstract: In the field of agricultural research, Machine Learning (ML) has been used to increase agricultural productivity and minimize its environmental impact, proving to be an essential technique to support decision making. Accurate harvest time prediction is a challenge for fruit production in a sustainable manner, which could eventually reduce food waste. Linear models have been used to estimate period duration; however, they present variability when used to estimate the chronological time of apple tree stages. This study proposes the PredHarv model, which is a machine learning model that uses Recurrent Neural Networks (RNN) to predict the start date of the apple harvest, given the weather conditions related to the temperature expected for the period. Predictions are made from the phenological phase of the beginning of flowering, using a multivariate approach, based on the time series of phenology and meteorological data. The computational model contributes to anticipating information about the harvest date, enabling the grower to better plan activities, avoiding costs, and consequently improving productivity. We developed a prototype of the model and performed experiments with real datasets from agricultural institutions. We evaluated the metrics, and the results obtained in evaluation scenarios demonstrate that the model is efficient, has good generalizability, and is capable of improving the accuracy of the prediction results.

Keywords: harvest date prediction; multivariate model; time series; recurrent neural network

1. Introduction

The science that studies the events that occur in the life cycle of plants is known as phenology. Phenology is a tool to interpret the interactions of culture with local climatic conditions. This allows us to characterize the occurrence of different stages and the duration of crop development periods, relating them to seasonal variations [1]. Figure 1 shows the sequence of the phenological stages of the apple tree. Phenological phases can be seen as organ transformations in plants, such as germination, budding, flowering, defoliation, and maturation. Phenological stages are specific phases or subdivisions that involve significant changes or characterize any plant condition of the plant [2]. Phenological events vary between years due to the variety of climatic elements, especially temperature [2]. Thermal availability has a direct influence on plant phenology. Detecting changes in Phenology has become a recurrent theme, and precision agriculture has increasingly attracted the attention of farmers, governments, and researchers, since plant monitoring is one of the essential tools for the management and optimization of agricultural resources [3]. The complex interactions between plant development and the environment represent a significant source of uncertainty for growers. The climate can change very quickly and have significant implications for fruit production, representing costs for growers. Predicting the harvest time is a challenge to develop sustainable fruit production and reduce food waste [4].

Apples are perishable, high-value, and seasonal, and selling prices are often time-sensitive, making harvest characteristics extremely valuable to growers [5].
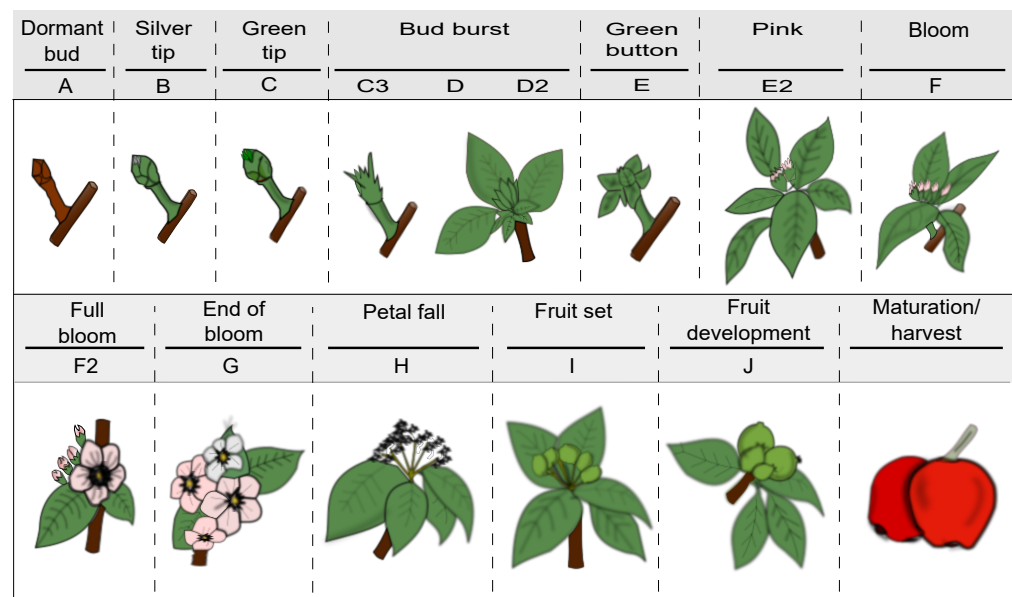
Figure 1. Apple phenologial stages.

Machine Learning (ML) is an area of Artificial Intelligence (AI) that, according to Mitchell et al. [6], is concerned with the question of how to construct computer programs that automatically improve with experience. ML approaches have been used in several areas including, for example, applications in the construction industry [7], in the medical field [8,9], in meteorology [10], and in biochemistry [11]. In the research field of agriculture, ML has been used in a variety of applications [12–15]. Agri-technology and precision agriculture have emerged as new scientific fields that use data-intensive approaches to boost agricultural productivity and minimize environmental impact. The data generated in modern agricultural operations allows a better understanding of the operating environment (an interaction of dynamic crop, soil, and climatic conditions) and the operation, leading to faster decision making [12]. In this context, ML is an important tool to support decision making, assisting in planning, handling and management, forecasting, disease detection, and the quality of agricultural production.

Linear models based on thermal sum have been used to estimate the occurrence of phenological stages of different crops. However, these models, when adjusted to the apple crop, show high variability for certain stages, failing to adequately represent the occurrence of the phenomenon. Solutions proposed for other cultures need adaptation due to the particularities of the apple tree culture. In this work, we developed PredHarv, a machine learning model, based on thermal sum, but using a multivariate approach with an RRN-based supervised learning algorithm. The model was developed to predict the start date of the apple harvest, given the climatic conditions related to the expected temperature for the period. A conventional pipeline of an LSTM network is used, which is a type of network designed specifically for sequence prediction problems that can deal with the temporal structure of inputs and capture the behavior dependent on the sequence of environmental stimuli, which is implicit in the time series. We assumed as basic hypotheses, which guided the development of this work, that the apple harvest date can be estimated through ML methods using a multivariate approach based on historical data of phenology and climatic parameters related to temperature, and also that multivariate models based on thermal summation can improve the accuracy of predictions in days, better explaining the relationship between temperature variables and the amount of heat needed for the plant to complete its cycle. The model contributes with a methodology that makes the predictive

capacity more effective, enabling the prospection of future scenarios. The development of this work contributes to the fruit growing area, making it possible to anticipate information about the harvest date, improving the accuracy of predictions. This information can be helpful to the fruit grower to perform planned activities, avoiding unnecessary costs with treatments, handling, and management as well as reflecting the quality of agricultural production. It can be used as an essential tool to support decision making, generating financial savings for the fruit grower.

This work is organized as follows. Section 2 presents an analysis of the works related to the theme proposed in this article. Section 3 presents the model for predicting the harvest date. Section 3.3 presents the evaluation methodology, details about prototype, input data and parameters, evaluation metrics, and evaluation scenarios. Then, in Section 4, the results obtained with the experiments carried out are presented. Section 5 includes a discussion of the results obtained. Finally, Section 6 features the conclusion.

2. Related Works

The prediction of phenological stages has been extensively investigated in the literature [1,16–21]. Many initiatives are found, which use different methods to offer solutions for the most varied cultures. Among the works related to apple phenology, in Petri et al. [1], the authors present studies on the phenology of apple trees in subtropical climatic conditions. The study uses phenological and climatic data, using a statistical approach, and presents important conclusions about the impacts on fruit production under subtropical climate conditions. In Putti et al. [16], the authors evaluated the phenological development of different fruit structures of apple trees with the objective of characterizing the behavior of the phenological period from the complete flowering until the beginning of the ripening of the apple. In Blazek and Pistekova [18], the fruit growth of apple cultivars was evaluated. The authors recommend using the relationship between the diameters of fruits at the T stage and fruits reached at harvest maturity to predict the harvest time and variety yield of four apple cultivars. In general, these works use statistical approaches, based on experiments, observing the evolution of the stages to verify the effect of the amount of cold or heat and to characterize the behavior of the phenological period of the apple. The authors conclude that the greater the number of cold units in dormancy, the shorter the time and need for heat for sprouting. Due to the peculiarities of fruit growing, the approaches generally developed for commodities need adaptations. In particular, there is a need to use unique methods with specific cultivation characteristics, which makes it difficult to apply techniques from other cultures without adaptation.

Thermal sum models, in the form of degree-days, have been adjusted to estimate heat accumulation [21]. These models represent the integration of effective temperatures for plant growth, which is fixed between the lower and upper limits. In Boechel et al. [19], the authors investigated the use of methods Fuzzy-based Time Series (FTS) to predict phenological stages of apple trees. The authors noted that the quality of results is improved by combining variables and using multivariate FTS methods. Studies, such as those by Rivero et al. [22], Darbyshire et al. [23], and Chitu and Paltineanu [21], evaluate the impacts of rising temperatures and climate change on the flowering season of apple trees due to global warming. These works produce statistics and relate meteorological and phenological data due to global warming and how much the initial phenological stages have changed due to climate change.

Approaches using ML methods are present in Chen et al. [20], Dai et al. [24], Czernecki et al. [25], and Haider et al. [26]. Some authors have applied artificial neural networks (ANN) to predict phenological stages, as in Yazdanpanah et al. [17] and Safa et al. [27]. In Yazdanpanah et al. [17], the authors used an ANN to anticipate different phonological phases of the apple. The authors identified that most of the error was related to the anticipation of the fruit development stage and concluded that it is possible to anticipate the phenological stages of the apple with acceptable accuracy using climatic parameters. In Safa et al. [27], the authors present a study where an Artificial Neural Network was applied

to predict the production of dry wheat. The results demonstrate precision and efficiency with a maximum error between 45 and 60 kg/ha. McCormick et al. [28] proposed a hybrid model using a data-driven model using knowledge-based predictions to predict soybean phenology. According to McCormick et al. [28], the potential that LSTM networks have to capture the impact depends on the sequence of environmental stimuli, which in the context of this work is analogous to, for example, a period of cold weather in plant development. In Chen et al. [20], a hybrid method is proposed combining neural networks with integrated learning to predict the flowering period of "Red Fuji" apple trees. The method uses LSTM networks and the Random Forest and Adaboost classification functions. The proposed model has high applicability and accuracy for flowering prediction. Haider et al. [26] developed a wheat yield production forecasting model using LSTM networks. The objective focuses on the development of a wheat yield prediction model using Robust-LOWESS as a smoothing function in conjunction with an LSTM neural network model. The results show that the proposed model achieves a good performance in terms of productivity prediction.

Aiming at a comparative analysis between the analyzed works, Table 1 presents the main characteristics of each approach. All works present some solutions to related problems. Although they do not have the same focus, they have a solid relationship with interest. The Culture column shows the work related to the apple tree culture. The ML column shows works that use ML techniques. The Other approach column shows works that use some other strategy to predict phenological stages and harvest prediction. The ANN and RNN columns present jobs that use these techniques. The literature review resulted in a set of articles that focus on harvest time prediction initiatives and others related to predicting the phenological stages in general of apple trees and other crops. Among the works, we can emphasize some characteristics. We see a solid tendency to assess the impacts of rising global temperature and climate change on plant phenology. We observe works with approaches in the sense of analysis or even making the function of phenology as a function of climate or other factors that alter phenomena. Many jobs are related to the cultivation of wheat, soybeans, and corn. We observe that a large part of the work related to harvest forecasting is focused on yield forecasting. Although there are many solutions related to the prediction of phenological stages, we observed a gap regarding the use of computational intelligence for a prediction start date of the harvest of apple trees. We did not find an approach similar to the one proposed in this work.

Table 1. Comparative table of approaches in the related work.

Reference(s)	Culture (Apple)	ML	ANN	RNN	Other Approach	Predict Harvest Date
[18]	✓				✓	✓
[19]	✓				✓	
[20]	✓	✓		✓		
[21]	✓				✓	
[25]		✓				
[24]		✓	✓			
[23]	✓				✓	
[26]		✓		✓		
[28]		✓		✓		
[1]	✓				✓	
[16]	✓				✓	
[22]	✓				✓	
[29]	✓	✓	✓			
[27]		✓	✓			
[17]	✓	✓	✓			

3. Materials and Methods

3.1. PredHarv Model Architecture

Figure 2 presents the architecture of the PredHarv model and its main elements. Initially, the model must be trained before making the predictions, with data from the phenology and temperature historical series. The blue arrows in Figure 2 represent the training step flow. The model only needs to be trained once unless there is a need to incorporate new data into the training dataset. Briefly, the phenological data identify the beginning and the end of the periods, and the meteorological data are the daily values of the variables in this period. The data consistency is checked in the pre-processing module, ensuring that all data are in the proper format to be presented to the model. In this step, lost data imputation operations are also carried out, including data normalization, and transformation in input and output patterns when necessary. After training, it is assumed that the model is trained, minimizing the error. This step is represented in Figure 2 by the Trained model module. Based on what is learned from the training stage, inferences are made from the new inputs informed by the user. The red arrows in Figure 2 represent the prediction step flow after training. Scenarios A, B, and C in the output step of Figure 2 represent the output of the model, that is, the duration of the period. These perspectives are possible due to the flexibility and predictive capacity of the model based on input data and allow prospecting for future scenarios.
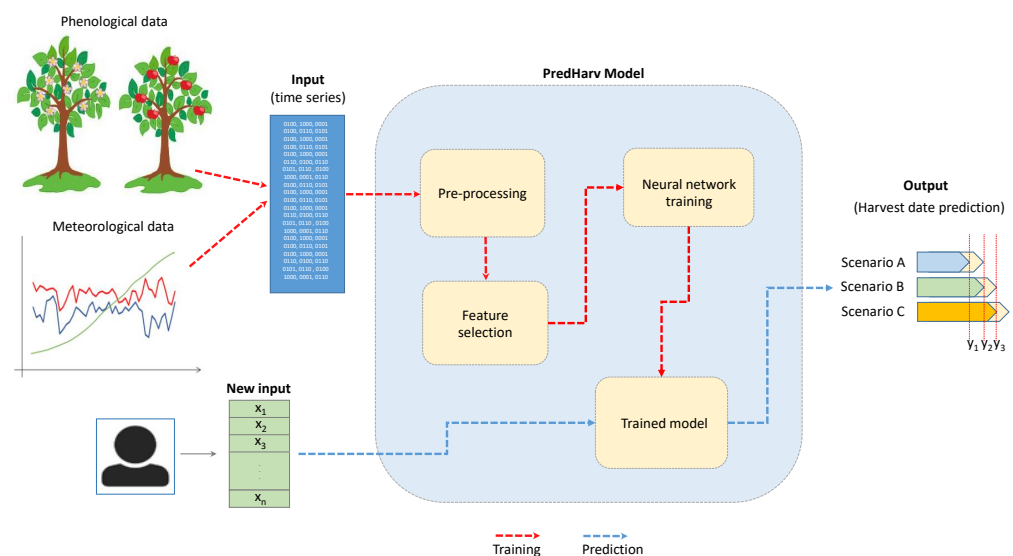

Figure 2. PredHarv model architecture.

The model inputs are time series of meteorological data, which are ordered in time, related, and contribute to the phenomenon's occurrence. We used six variables that serve as input for the model: daily minimum temperature (t_{min}), maximum daily temperature (t_{max}), daily average temperature (t_{med}), average daily minimum temperature (mt_{min}), the average temperature of the period (mt_{med}), and cumulative degree days (CDD). The input variables use the values of previous observations ($t - n, \ldots, t - 2, t - 1$). These previous observations are called lags, which correspond to the value immediately preceding the series. This technique is called a sliding window [30]. We use seven lags for each input variable, that is, the values corresponding to seven lags immediately previous to the current time (t). We performed tests with window sizes of three, five, and seven lags to observe the variability of these models with windows of different sizes. The best results are obtained with seven lags. The output variable is a specific numerical value, which corresponds to the duration of the period in the prediction horizon one step ahead ($t + 1$).

3.2. Prediction Strategy

The PredHarv model makes predictions from information provided by the user. The model does not require any basic configuration by the user. The user's role is extremely restricted but of paramount importance in the prediction result, since the network responds to the context being fed. They are considering that the information fed at the beginning of the period has a longer prediction horizon due to the size of the window. With it, there is a higher level of uncertainty of the future scenario. There is no rigidity regarding the number of predictions. Only the first is considered mandatory. The PredHarv model was designed to allow flexibility and other predictions to be made at any time within the period between full bloom and the start of harvest. These predictions allow the estimate to be adjusted as the input data are updated with information on days that have already occurred and a smaller future window. Figure 3 illustrates the process of updating the prediction as new data inputs are fed. In Figure 3, new entries are fed by the user represented by x, and the predictions are represents by y. As the predictions are made, there is a tendency to converge toward a common point, which is represented in the figure by the convergence area.
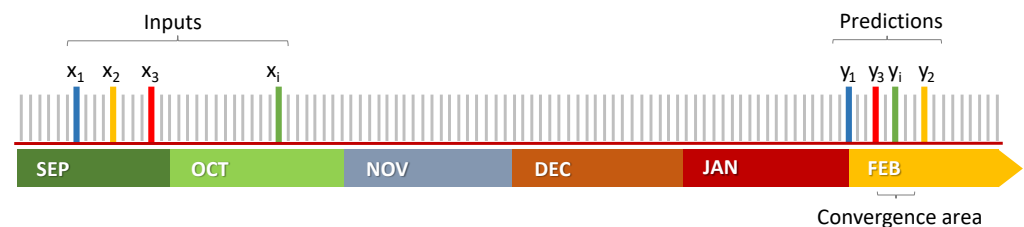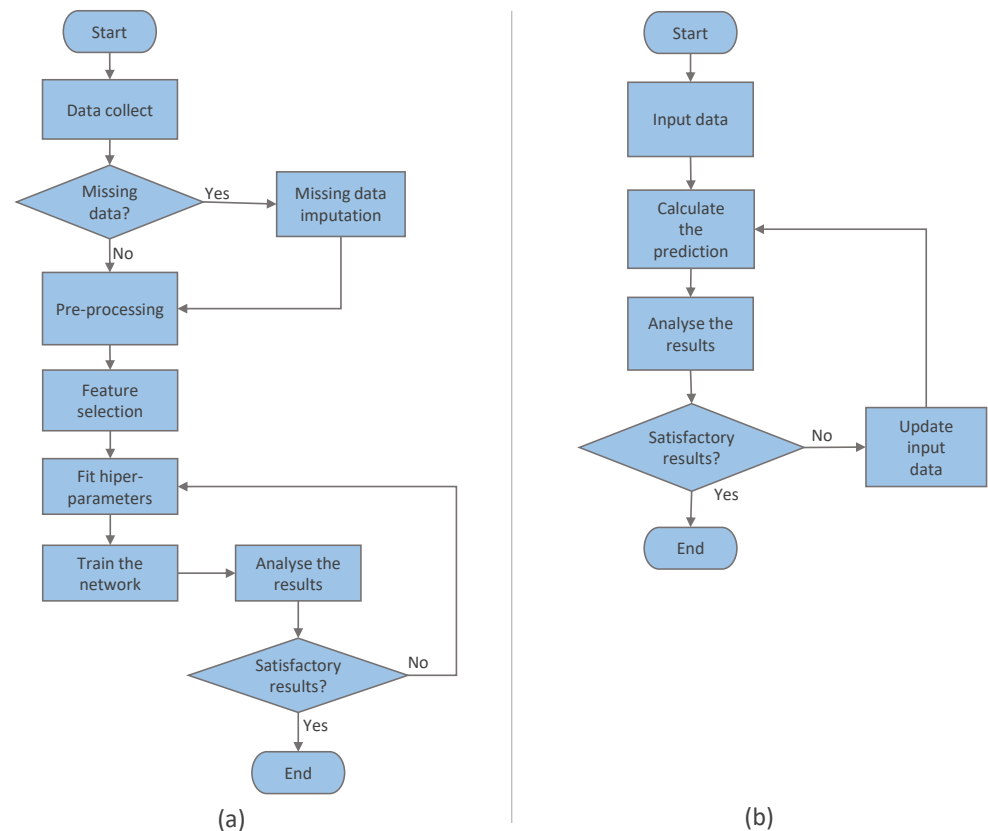

Figure 3. Timeline of the prediction process.

Figure 4 presents the processing flow of the PredHarv model in the training steps in Figure 4a and prediction in Figure 4b. In the training step in Figure 4a after data collection, the missing data are analyzed. A common problem when working with time series involving real scenarios is the occurrence of missing data. Ideally, these series would be complete. However, in real scenarios, this situation is quite rare. Several problems can cause the loss of important information, such as occasional interruptions of automatic stations and malfunction of measuring instruments. To get around this problem, statistical techniques involve replacing missing data with value estimates. One strategy is to impute the missing values, that is, to infer them from the known part of the data. These techniques aim to complete the dataset, making it possible to work with all the data under study. We use the SimpleImputer class from the Scikit-learn library that works with the mean, the median, and the most frequent attributes. The default value of the SimpleImputer class will work with the average. This technique proved to be efficient in filling the gaps; the method applied in imputation preserves the structure, and it was observed that it does not compromise with the characteristics of the original series. In the data transformation and normalization module, operations were performed to ensure that the data are all in the same format and the same order of magnitude. Data transformation is a practice to prevent the algorithm from being biased toward variables with a higher order of magnitude, and normalization aims to place the variables within the range between 0 and 1.

We use the MinMaxScaler class from the Scikit-learn library. In the feature selection module, the input attributes of the model were selected. Details on feature selection can be seen in Section 3.3.3. In the Fit parameters module, the main parameters used in the configuration of the LSTM network were established. These parameters are described in Section 3.3.1. Then, in the training module, network training is carried out. If the training results are not considered satisfactory, as evaluated from the training residues of the training and test series, the values of the hyperparameters are adjusted, and the model is retrained. If considered satisfactory, the model is trained to perform the predictions. In the prediction step in Figure 4b, in the Input data module, the user informs the values of the input attributes that are used to perform the prediction. Then, these input data are

submitted to the trained model that performs the prediction. In the next step, the user analyzes the prediction result. This step can be repeated as many times as the user deems necessary. The input data can be updated at each repetition, and consequently, this change will reflect the prediction value.

Figure 4. Processing flow. (**a**) Training step. (**b**) Prediction step.

In this work, we use a conventional pipeline of the variant of RNN called the LSTM, which was first introduced by Hochreiter and Schmidhuber [31]. LSTM networks were designed to avoid long-term dependency, work very well on a wide variety of problems, and are widely used today. The central idea behind the LSTM architecture is a continuously updated memory. It consists of an elementary unit called a memory cell and gates. Through the gates, the network acquires the ability to remove or add information to the state of the cell. These gates are the forget gate, input gate, and output gate. The forget gate decides which information will be kept and which will be discarded from previous states. The input gate decides what information is to be updated, and the output port is based on the current state of the cell and decides what information is to be produced.

The neural network structure uses a Keras sequential model, providing a series structure where the output of one layer serves as the input for the next. In the implementation of the LSTM network, we used an LSTM layer with 200 neurons. The input_shape parameter was defined according to the dimension of the input layer: that is, the number of columns of features selected in the dataset and the number of lags. The activation function used in this layer was Rectified Linear Unit (ReLU). Then, another layer of the Dense type was included, with 1024 neurons, with activation function Rectified Linear Unit (ReLU). Finally, a Dense layer with one neuron and Linear Activation function is used to return a single continuous value. In the stage of compiling the model, the Adaptive Moment Estimation (Adam) function was used in the optimizer parameter. This function defines how the weights of the neural network are updated. The loss parameter was defined as Mean Absolute Error (MAE). The training of the network in this stage was carried out with 300 epochs.

3.3. Evaluation Methodology

3.3.1. Prototype

We implemented a model prototype using the Python programming language, version 3.8, using the open-source tool Jupyter Notebook and the Keras library. The Keras library (available at: https://keras.io/ accessed on 17 February 2022) is an open-source neural network library written in Python that is capable of running on TensorFlow (available at: https://www.tensorflow.org/accessed on 13 February 2022), which is also an open-source library for numerical computing and machine learning made available by the Google Brain team. We use real data from an agricultural institution to train and test the model. We performed tests with real data of the data portion that the model does not know, that is, a portion of data that was not presented to the model during the training phase of the network. These data are used to evaluate the behavior of predictions in relation to real data.

3.3.2. Dataset

The datasets used in this work are composed of data collected in apple orchards of agricultural institutions located in southern Brazil. Weather data come from weather stations located near the orchards. Dataset1 (DS1) comes from the municipality of Caçador located in western Santa Catarina (26°49′2.884″ S, 50°59′45.044″ W) with an altitude of 960 m comprising the period from 2000 to 2018. Dataset2 (DS2) comes from the municipality of Vacaria located in the northeast region of Rio Grande do Sul (29°32′30″ S, 50°54′54″ W) with an altitude of 962 m covering the period from 2016 to 2020.

The meteorological data were organized in a CSV file. A small excerpt is presented in Table 2. These data were used as input to the implementation. The phenological stages data consist of visual observations of the plant. In order to contribute to the reproducibility of the results of this article, the data set is available in a publicly available repository (available: https://github.com/tboechel/dataset/ accessed on 17 February 2022).

Table 2. Input data used in the implementation.

Date	*Tmin*	*Tmax*	*Tmed*	*Tmin_acc*	*Tmed_acc*	*DD*	*CDD*
22 October 2006	12.0	25.0	18.5	12.0	18.5	15.5	15.5
23 October 2006	13.0	23.8	18.4	25.0	36.9	15.4	30.9
24 October 2006	12.5	25.0	18.7	37.5	55.6	15.8	46.7
25 October 2006	14.0	25.6	19.8	51.5	75.4	16.8	63.5
26 October 2006	13.0	29.0	21.0	64.5	96.4	18.0	81.5

3.3.3. Input Variables

We selected the variables considered empirically eligible to serve as input to the model. The following variables were selected: daily minimum temperature (t_{min}), maximum daily temperature (t_{max}), daily average temperature (t_{med}), amplitude (*amplit*), average daily minimum temperature (mt_{min}), average of the maximum daily temperature (mt_{max}), average temperature of the period (mt_{med}), degree-day (*DD*), and cumulative degree days (*CDD*). Figure 5 shows a boxplot diagram with the variables used.

Thermal accumulation, in the form of the sum of degree-days (*DDs*), is an index used to establish a relationship between the amount of heat the plant needs for its development and the average daily temperature. According to [1], the physiological process and the functions of the plant occur under thermal limits in its environmental development. To complete all the physiological sub-periods of the life cycle, some cultures require an accumulation of a certain amount of heat. This accumulation is usually expressed by the *DD* index and represents the thermal sum above the minimum base temperature (T_{base}).
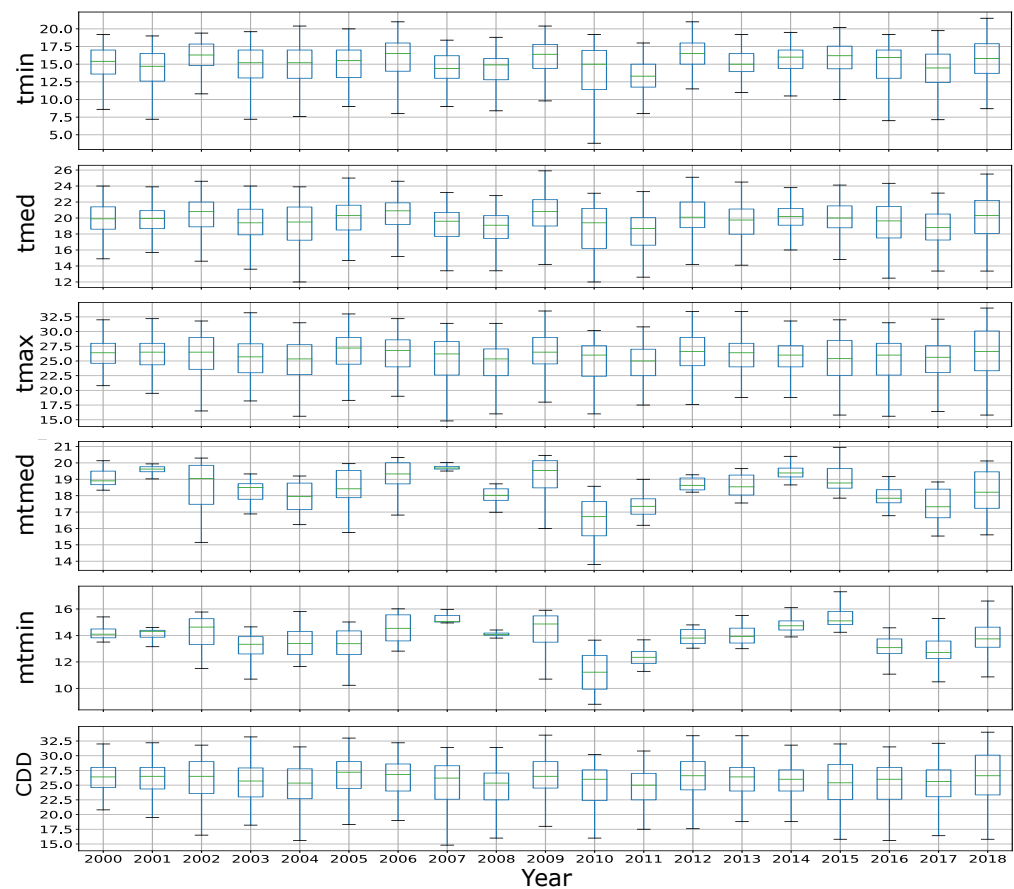
Figure 5. DS1 input variables boxplot.

The concept of *DD* admits that there is a base temperature and that below such limit, the growth and development of the plant are interrupted or extremely reduced. In addition, a linear relationship between temperature and plant development is assumed since there are no limitations on other factors [32]. The *DD* heat unit index can be calculated using Equation (1), where T_{max_i} and T_{min_i} are the maximum and minimum daily temperature, and T_{base} is the base temperature. All temperatures are measures in degrees Celsius (°C). Based on the *DD* value, the thermal accumulation of a given period called the cumulative degree-day *CDD* could be calculated using Equation (2). The base temperature was determined by the method indicated by Arnold [33], which uses the smallest standard deviation. Different values attributed to the base temperature can be found in the literature. We carry out tests with temperatures of 2 °C, 4 °C, and 10 °C. The best result was obtained with a temperature of 2 °C with a coefficient of variation of 10.46, being 6.6% lower than the results obtained with 10 °C and 1.9% lower than the results obtained with 4 °C.

$$DD_i = \frac{T_{max_i} + T_{min_i}}{2} - T_{base} \tag{1}$$

$$CDD = \sum_{i=1}^{n} DD_i \tag{2}$$

We investigate the importance of variables to serve as input to the model. We used Pearson's correlation coefficient to assess the degree of relationship between pairs of variables to assess the relevance of each variable alone as predictors of period length. This coefficient is a measure of dependence between two quantitative variables, and the correlation analysis shows how much one variable is somehow related to the other. Figure 6 graphically shows the correlation heatmap of the initially selected resources.

The variable with the highest degree of correlation with the duration variable alone was the *CDD* variable. We fit a simple linear regression model using the variable *CDD* as a function of period length. The results showed a strong positive correlation. The coefficient of determination R^2 indicated that 83.8% of the variability in period length (number of days) could be explained by the thermal accumulation *CDD*. Next, we observe the influence of the inclusion of other variables in the model. A multivariate linear regression model was adjusted considering the length of the period but with the inclusion of other variables. The variables that had the best response were *CDD*, minimum temperature T_{min}, and average temperature T_{avg}. The coefficient of determination R^2 indicated that 98% of the variability in period length could be explained by combining the variables *CDD*, T_{min}, and T_{avg}. We considered the correlation analysis, but we performed tests to assess the importance of attributes as predictors, intending to observe whether or not it contributes to improving the model's accuracy. The best results were obtained with the combination of six variables (see Figure 5), the others having been discarded.
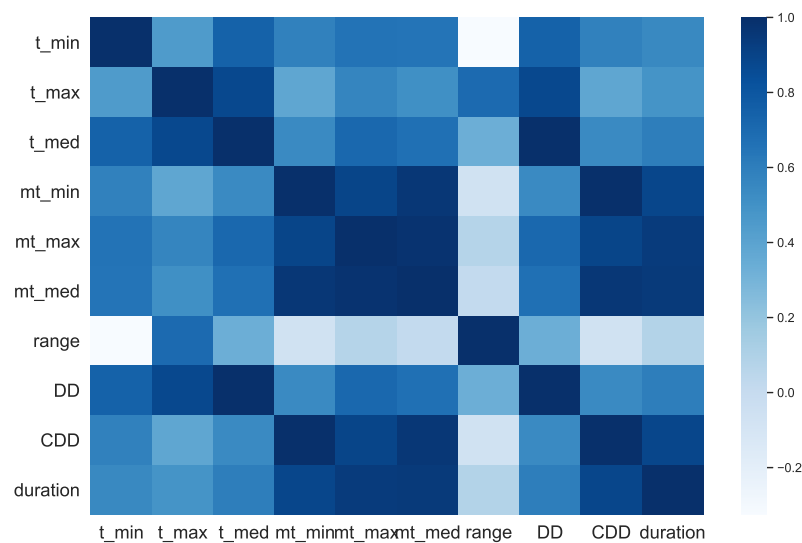
Figure 6. Feature correlation heatmap.

3.3.4. Evaluation Scenarios

To evaluate the PredHarv model, tests were performed in four evaluation scenarios. The use of these scenarios aims to perform a performance estimate and provide information about the operation of the model and how it can be improved. The evaluation of the model was carried out considering the following scenarios:

- Scenario 1: The first evaluation scenario is defined using the validation methodology called K-fold cross-validation (CV). CV is a basic form of cross-validation to assess a model's generalizability from a dataset [34]. The CV divides the systematic measures of data into k groups and performs a cross-validation that produces a more robust evaluation set than just the model once. Figure 7 illustrates how the method works. For each sample, performance is evaluated according to the metrics described in Section 3.3.5.

- Scenario 2: In the second evaluation scenario, we run tests with real data from dataset1 and dataset2. We use samples from the datasets to analyze the actual values against the predicted values returned by the model. The samples were selected according to the validation scheme used in the evaluation scenario 1. The datasets were divided into subsets with part of the data for training and another for testing, respecting the temporal order of the data. For each sample, the RMSE and MAE metrics were calculated.

Figure 7. K-fold cross-validation with k = 5.

- Scenario 3: In the third evaluation scenario, we use synthetic data. Variations were made in some time windows in certain periods, with an increase or decrease in the values of the temperature variables. The objective is to visualize the behavior of the method with guided changes in the input data.
- Scenario 4: In the fourth evaluation scenario, we conducted tests with other ML algorithms applied to the problem. This evaluation scenario focuses on the selection of the best supervised ML algorithm and its hyperparameters tuning. We use Automated Machine Learning (AutoML) methods to select the best models from a set of models and optimize their hyperparameters. In order to produce a just comparison between the models, the assessments were divided into categories: General ML (GML) algorithms selection and Deep Leaning (DL). In the GML category, we use Auto-Sklearn (available at: https://www.automl.org/automl/auto-sklearn/ accessed on 13 February 2022), which is an AutoML library built on top of the Scikit-Learn ML framework. The choice of algorithms and hyperparameters implemented by Auto-Sklearn takes advantage of recent advances in Bayesian optimization, meta-learning, and Ensemble Learning [35]. We use Auto-SkLearn version 0.14.4. In the DL category, we use Auto-Keras (available at: https://autokeras.com/ accessed on 13 February 2022), which is a Python library based on the Keras module and that is focused on an automatic DL Neural Architecture Search (NAS) [36]. The search is performed by using a Bayesian optimization, with the tool automatically tuning the number of dense layers, units, type of activation functions used, dropout values, and other DL hyperparameters. In this work, we adopt Auto-Keras version 1.0.18. Each tool is meansured in terms of its predictive performance using K-fold cross-validation with k = 10. The best model result of each AutoML tool is presented.

3.3.5. Evaluation Metrics

We use the metrics Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to measure and evaluate the methods' accuracy. Both are described in Equations (3) and (4). Here, Y represents the real observations dataset, while $\hat{Y}$ is the prediction value. These indicators are based on operations to quantify the difference between the actual values observed and the values predicted by the method.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (4)$$

4. Results

We performed tests with the implemented prototype of the PredHarv model in order to observe its functioning. We use real data from two agricultural institutions and synthetic data. The results for each scenario are presented below:

- Results of the First Evaluation Scenario:
  Table 3 shows the results of the implementation of the first evaluation scenario. The data used come from the DS1 dataset. The objective of implementing this evaluation scenario was to quantify how well this model behaves when making predictions on never-before-seen data and to observe the generalization capacity of the model in different subsets. In the K-fold cross-validation approach used, the dataset was divided 10 into training and test subsets. The RMSE metric values remain at the mean of 0.90 with a standard deviation of 0.17, and the MAE metric values remain at the mean of 0.80 with a standard deviation of 0.15. It was observed in the results obtained in this scenario that the model has a good generalization capacity. No evidence of overfitting and underfitting occurs.

Table 3. Results of the metrics with K-fold cross-validation method.

Subset	RMSE	MAE
1	0.99	0.93
2	0.96	0.77
3	1.13	1.01
4	0.71	0.66
5	0.73	0.67
Average	0.90	0.80
Standard deviation	0.17	0.15

- Results of the Second Evaluation Scenario:
  Table 4 shows the results of the tests in the second evaluation scenario, which were obtained with the DS1 dataset. Table 5 shows the results of the tests in the second evaluation scenario, which were obtained with the DS2 dataset. The Real column shows the duration of observed data from samples in the actual dataset. These values refer to the duration of the period in days referring to the sample. The Predict column shows the data predicted by the PredHarv model. These values in the Predict column and the RMSE and MAE metrics are calculated after ten runs of each sample. In Table 4, the RMSE metric varied between 0.35 and 2.37 and MAE varied between 0.34 and 1.97 for the selected samples. In Table 5, it is observed that the results are very close, presenting RMSE metrics between 0.41 and 1.32 and MAE between 0.33 and 1.08.

Table 4. Prediction results with the second evaluation scenario—dataset DS1.

Sample	Real	Predict	RMSE	MAE
1	104	104.3	0.35	0.34
2	108	108.5	0.99	0.94
3	110	108.7	2.37	1.97
4	115	114.3	1.02	0.89
5	122	121.2	1.32	1.08
Average			**1.41**	**1.23**

Table 5. Prediction results with the second evaluation scenario—dataset DS2.

Sample	Real	Predict	RMSE	MAE
1	128	127.4	0.45	0.43
2	124	123.3	0.50	0.39
3	135	135.8	0.41	0.33
4	132	132.7	1.32	1.08
5	134	133.3	1.31	1.02
Average			**0.79**	**0.65**

- Results of the Third Evaluation Scenario:

  In this scenario, synthetic data with temperature variations were used in specific periods. Figure 8 shows the dynamics of the predictions obtained with the PredHarv model. The initial prediction was made on 8 October, which is the date that marks the beginning of the period of full flowering. Based on the data fed, the estimated duration was 121 days. The blue line, in Figure 8, identifies the predicted value in the initial stage. Subsequently, new predictions were made, which are called adjustment predictions. The first adjustment prediction was made on 8 November, 31 days after the beginning of the period. A period of two weeks was simulated, between the dates of 20 October and 2 November, with temperatures 20% higher than initially predicted. The second adjustment prediction was made on 20 November, 43 days after the beginning of the period. Within this period, temperatures between 12 November and 17 November were simulated 10% below those initial predicted. The third adjustment prediction was made on 30 December, 83 days after the beginning of the period. Between 1 January and 19 January, temperatures were simulated 10% below those initially predicted. In the fourth and last adjustment prediction, carried out on 20 January, 107 days after the beginning of the period, temperatures were simulated 20% below those initially predicted. As can be seen in Figure 8, the red line shows the duration of the period corresponding to the date on which the prediction was made. These values vary, and it can be seen that the model responds to the stimulus of changes in the input data, adjusting the prediction as the data are being introduced. Variations in temperature in specific periods result in a more significant accumulation of heat, shortening the duration of the period. At other times, less accumulation occurs, increasing the duration as expected. It was observed in the tests with this scenario that the answers obtained corroborate the initial hypothesis that the apple harvest date can be estimated through ML methods using a multivariate approach based on historical data of phenology and climatic parameters related to temperature.
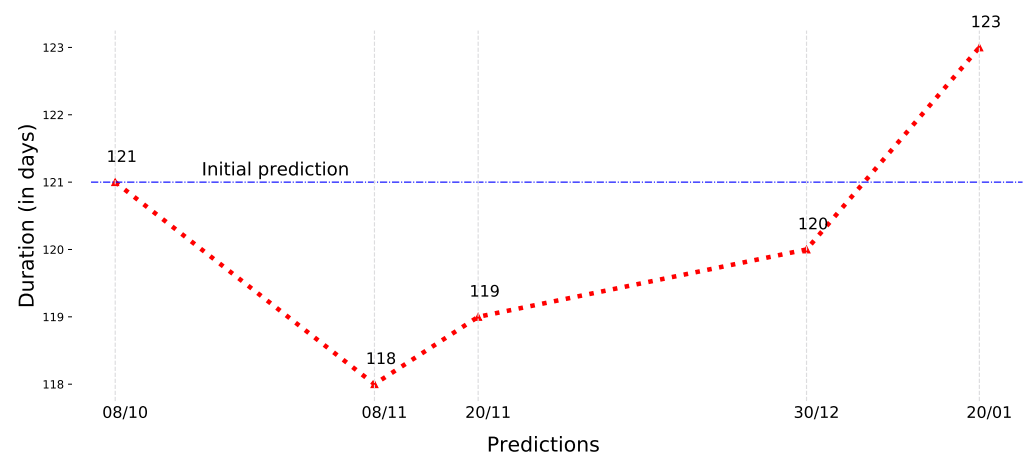
Figure 8. Dynamics of the predictions obtained with the PredHarv model.

- • Results of the Fourth Evaluation Scenario:
  In the fourth evaluation scenario, we used AutoML methods to select the best algorithm and tune its hyperparameters applied to the DS1 dataset. We use samples according to the K-fold cross-validation scheme, with k = 10, as described in Section 3.3.4. We divided the solutions into two categories. In the GML category, we use Auto-Sklearn. The best result in the ranking of the models tested by AutoSklearn was with Multiple Linear Regression (MR). Table 6 presents the results obtained from the model.

Table 6. Best results using AutoSklearn with MR.

Sample	RMSE	MAE
1	3.68	3.47
2	1.54	1.16
3	1.23	0.95
4	3.92	3.41
5	2.56	2.16
Average	2.58	2.23
Standard deviation	1.21	1.19

In the DP category, we use AutoKeras. The library is focused on automatic searching of DL neural architecture (NAS), automatically adjusting the number of dense layers, units, activation functions used, dropout values, and other DL hyperparameters. This task is often called architecture research, and it is often called neural architecture research. The MLP neural network structure returned by Autokeras is composed of a sequential model composed of a Dense layer with 128 neurons and ReLU activation function. Then, there is another layer of Dense type with 1024 neurons and ReLU activation function. In the output step, there is a Dense layer with only one neuron, and 134 parameters were tested. The network training was performed with 300 epochs. Table 7 presents the results of the metrics obtained from samples submitted to the model.

Table 7. Best results using AutoKeras with MLP.

Sample	RMSE	MAE
1	4.13	3.78
2	2.94	3.31
3	1.97	1.69
4	4.46	3.67
5	1.96	1.55
Average	3.09	2.80
Standard deviation	1.17	1.09

5. Discussion

In the results obtained with the MR method (Table 6), the RMSE metric obtained an average value of 2.58 in the selected samples, and the metric of the MLP method (Table 7) obtained an average value of the RMSE metric of 3.09. Comparing these results, the MR method had a better result with a difference of 0.51. However, it was observed that the standard deviation measure, which is a measure of dispersion around the mean, shows that the results obtained with MLP are more uniform. The values obtained with PredHarv (Table 3) in the RMSE metric are 0.90 with a standard deviation of 0.17 for the selected samples. These values are lower than those obtained with the MR and MLP method. The difference between the results obtained with PredHarv is 2.19 lower compared to MLP, and it is 1.68 lower than with MR. The standard deviation measurement of the PredHarv method also shows uniformity in the results obtained.

Auto-sklearn and Autokeras are fully automatic AutoML methods. They are black box models, and their internal structure is unknown, being limited to measures of input and output relationships. However, these methods search a vast space of models and build complex sets of high precision, building models based on the data, which collaborates with the purpose for which they were submitted in this work. Investigations of other methods that can be applied to the problem demonstrate that the choice of the LSTM method is adequate to the nature of the problem. The ML method used exploits the potential of LSTM networks, which are a special type of RNN to deal with problems involving time series. PredHarv model predictions are made from the phenological stage of full bloom. These time series are a set of observations ordered in time, and they have a temporal dependence between the observations. Recurrent networks applied to this type of problem have the ability to capture part of the information implicit in the series, reflecting the impact of a sequence of environmental stimuli on plant development.

Linear models that use the sum of degree-days have been used to estimate the duration of phenological periods of apple trees and other crops [37]. These models, when adjusted to estimate phenological sub-periods of apple trees, present results with a lower variability in the accumulation of degree-days for most phenological stages between seasons, compared to the number of days, responding better to the thermal sum than to the chronological time (calendar days). According to [38], linear models have the obvious defect that the capability of the model is limited to a linear function, and therefore, the model may not explain the interaction between any two input variables. The PredHarv model uses thermal sum models as a baseline, but it uses a multivariate approach. We use the thermal sum relating the accumulation of degree-days to the duration of the period and other temperature-related variables in order to estimate the chronological time. The inclusion of other variables, with the multivariate strategy associated with the use of an ML method as proposed in this work, significantly improves the response, as can be seen in the comparative graphs of predictions with the PredHarv model and the linear model, in Figures 9 and 10, managing to better explain the occurrence of the phenomenon.

The results observed in the experiments carried out confirm the basic hypotheses that guided the development of this work: that the apple harvest date can be estimated through ML methods using a multivariate approach based on historical data of phenology and climatic parameters related to temperature, and also that multivariate models based on thermal summation can improve the accuracy of predictions in days, better explaining the relationship between temperature variables and the amount of heat needed for the plant to complete its cycle.
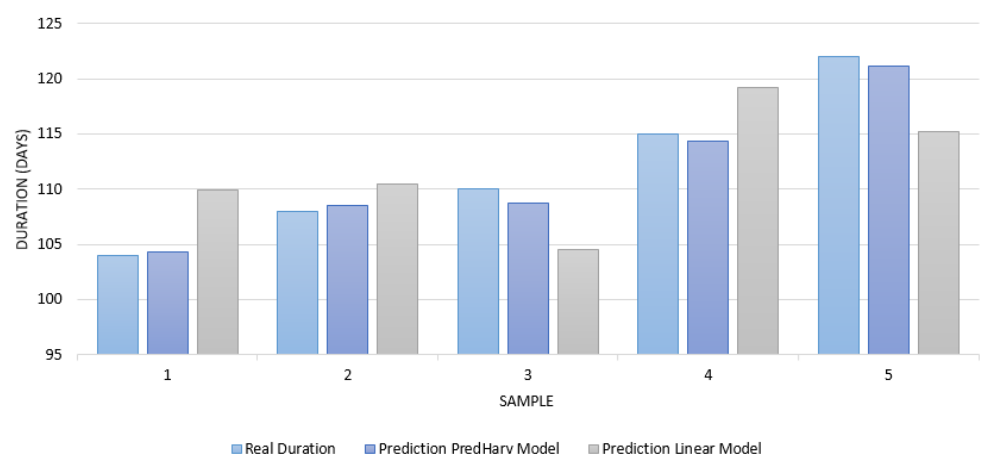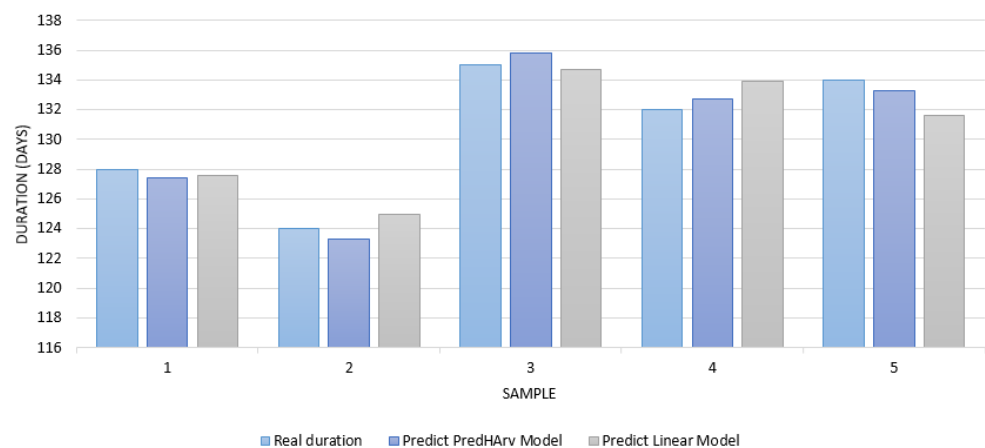

Figure 9. PredHarv model and linear model prediction results—DS1 dataset.

Figure 10. PredHarv model and linear model prediction results—DS2 dataset.

6. Conclusions

This work proposed a new approach to predict the harvest date in apple trees. The computational model PredHarv uses a multivariable approach, using historical and meteorological data with a strategy that uses RNN. The proposed evaluation scenarios and the comparative tests between the models show that the PredHarv model is a useful alternative that is capable of improving the accuracy of the prediction results. This work contributes to fruit growing, enabling anticipating information about the harvest date, generating financial savings for the fruit grower, avoiding unnecessary costs, and improving planning and productivity. The PredHarv model contributes with a methodology that makes the predictive capacity and applicability more effective, enabling the prospecting of future scenarios.

A prototype of the PredHarv model was developed and submitted to evaluation scenarios. The results were evaluated according to the RMSE and MAE metrics. In the tests of the first PredHarv evaluation scenario, it demonstrated a generalization capacity when submitted to new data. In the second scenario, PredHarv showed results very close to the real values in tests with real data. In the third scenario, when submitted to synthetic data, the results showed that the PredHarv model showed a positive response to the stimuli caused by the changes made to the input data, adjusting the prediction as the data are being introduced. We performed tests with other supervised learning methods, and the average of the RMSE metric results obtained with LSTM was superior to the results obtained with other algorithms.

Author Contributions: Conceptualization, T.B.; Investigation, T.B. and L.M.P.; Methodology, T.B. and L.M.P.; Supervision, G.d.O.R. and R.d.R.R.; Validation, T.B.; Writing—original draft, T.B.; Writing—review & editing, L.M.P., G.d.O.R., R.d.R.R. and D.S. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In order to contribute to the reproducibility of the results of this article, the data set is available in a publicly available repository at https://github.com/tboechel/dataset/ (accessed on 3 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Petri, J.L.; Hawerroth, F.J.; Leite, G.B.; Couto, M.; Francescatto, P. Apple phenology in subtropical climate conditions. In *Phenology and Climate Change*; Zhang, X., Ed.; InTech Europe: Rijeka, Croatia, 2012; pp. 195–216.
2. Bergamaschi, H. O clima como fator determinante da fenologia das plantas. In *Fenologia: Ferramenta Para Conservação, Melhoramento e Manejo de Recursos Vegetais Arbóreos*; Embrapa Florestas: Colombo, Sri Lanka, 2007; Volume 1, pp. 291–310.
3. Broich, M.; Huete, A.; Paget, M.; Ma, X.; Tulbure, M.; Coupe, N.R.; Evans, B.; Beringer, J.; Devadas, R.; Davies, K.; et al. A spatially explicit land surface phenology data product for science, monitoring and natural resources management applications. *Environ. Model. Softw.* **2015**, *64*, 191–204. [CrossRef]
4. Lee, M.A.; Monteiro, A.; Barclay, A.; Marcar, J.; Miteva-Neagu, M.; Parker, J. A framework for predicting soft-fruit yields and phenology using embedded, networked microsensors, coupled weather models and machine-learning techniques. *Comput. Electron. Agric.* **2020**, *168*, 105103. [CrossRef]
5. Morris, J.; Else, M.; El Chami, D.; Daccache, A.; Rey, D.; Knox, J.W. Essential irrigation and the economics of strawberries in a temperate climate. *Agric. Water Manag.* **2017**, *194*, 90–99. [CrossRef]
6. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997. Available online: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf (accessed on 13 February 2022).
7. Xu, Y.; Zhou, Y.; Sekula, P.; Ding, L. Machine learning in construction: From shallow to deep learning. *Dev. Built Environ.* **2021**, *6*, 100045. [CrossRef]
8. Garg, A.; Mago, V. Role of machine learning in medical research: A survey. *Comput. Sci. Rev.* **2021**, *40*, 100370. [CrossRef]
9. Darmawahyuni, A.; Nurmaini, S.; Caesarendra, W.; Bhayyu, V.; Rachmatullah, M.N.; Firdaus. Deep learning with a recurrent network structure in the sequence modeling of imbalanced data for ECG-rhythm classifier. *Algorithms* **2019**, *12*, 118. [CrossRef]
10. Schultz, M.; Betancourt, C.; Gong, B.; Kleinert, F.; Langguth, M.; Leufen, L.; Mozaffari, A.; Stadtler, S. Can deep learning beat numerical weather prediction? *Philos. Trans. R. Soc. A* **2021**, *379*, 20200097. [CrossRef]
11. Wang, H.; Wang, H.; Zhang, J.; Li, X.; Sun, C.; Zhang, Y. Using machine learning to develop an autoverification system in a clinical biochemistry laboratory. *Clin. Chem. Lab. Med. (CCLM)* **2020**, *1*, 883–891. [CrossRef]
12. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef]
13. Kamilaris, A.; Prenafeta-Boldú, F.X. Deep learning in agriculture: A survey. *Comput. Electron. Agric.* **2018**, *147*, 70–90. [CrossRef]
14. Elavarasan, D.; Vincent, D.R.; Sharma, V.; Zomaya, A.Y.; Srinivasan, K. Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Comput. Electron. Agric.* **2018**, *155*, 257–282. [CrossRef]
15. Tripathi, M.K.; Maktedar, D.D. A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A survey. *Inf. Process. Agric.* **2020**, *7*, 183–203. [CrossRef]
16. Putti, G.; Mendez, M.E.G.; Petri, J.L. Unidades de frio e de Calor para a Brotação de Macieira (Malus domestica, Borck), Gala e Fuji. *Rev. Bras. Agrociência* **2000**, *6*, 194–196.
17. Yazdanpanah, H.; Ohadi, D.; Soleimani, T.M. Forecasting different phenological phases of apple using artificial neural network. *J. Res. Agric. Sci.* **2010**, *6*, 97–106.
18. Blazek, J.; Pistekova, I. Prediction of the harvesting time for four apple cultivars on the basis of beginning of flowering and attaining of t-stage of fruitlets and dependence of diameter of fruitlets at t-stage and fruits at ripening stage. *J. Hortic. Res.* **2017**, *25*, 55–59. [CrossRef]
19. Boechel, T.; Policarpo, L.M.; Righi, R.; Ramos, G.d.O. Fuzzy Time Series for Predicting Phenological Stages of Apple Trees. In Proceedings of the SAC 2021—36th ACM/SIGAPP Symposium On Applied Computing, Gwangju, Korea, 22–26 March 2021; pp. 934–941.
20. Chen, C.; Zhang, X.; Tian, S. Research on Dynamic Forecast of Flowering Period Based on Multivariable LSTM and Ensemble Learning Classification Task. *Agric. Sci.* **2020**, *11*, 777. [CrossRef]
21. Chitu, E.; Paltineanu, C. Timing of phenological stages for apple and pear trees under climate change in a temperate-continental climate. *Int. J. Biometeorol.* **2020**, *64*, 1263–1271. [CrossRef]
22. Rivero, R.; Sønsteby, A.; Heide, O.; Måge, F.; Remberg, S. Flowering phenology and the interrelations between phenological stages in apple trees (*Malus domestica* Borkh.) as influenced by the Nordic climate. *Acta Agric. Scand. Sect. B—Soil Plant Sci.* **2017**, *67*, 292–302.
23. Darbyshire, R.; Farrera, I.; Martinez-Lüscher, J.; Leite, G.B.; Mathieu, V.; El Yaacoubi, A.; Legave, J.M. A global evaluation of apple flowering phenology models for climate adaptation. *Agric. For. Meteorol.* **2017**, *240*, 67–77. [CrossRef]
24. Dai, W.; Jin, H.; Zhang, Y.; Liu, T.; Zhou, Z. Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: Application of machine learning in phenological model. *Agric. For. Meteorol.* **2019**, *279*, 107702. [CrossRef]
25. Czernecki, B.; Nowosad, J.; Jablonska, K. Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset. *Int. J. Biometeorol.* **2018**, *62*, 1297–1309. [CrossRef] [PubMed]
26. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM neural network based forecasting model for wheat production in Pakistan. *Agronomy* **2019**, *9*, 72. [CrossRef]
27. Safa, B.; Khalili, A.; Teshnehlab, M.; Liaghat, A. Artificial neural networks application to predict wheat yield using climatic data. In Proceedings of the 20th International Conference on IIPS. Iranian Meteorological Organization, Vienna, VA, USA, 10–15 January 2004; pp. 1–39.

28. McCormick, R.F.; Truong, S.K.; Rotundo, J.; Gaspar, A.P.; Kyle, D.; van Eeuwijk, F.; Messina, C.D. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. *bioRxiv* **2020**, *3*, diab004. [CrossRef]

29. Sabzi, S.; Abbaspour-Gilandeh, Y.; García-Mateos, G.; Ruiz-Canales, A.; Molina-Martínez, J.M.; Arribas, J.I. An automatic non-destructive method for the classification of the ripeness stage of red delicious apples in orchards using aerial video. *Agronomy* **2019**, *9*, 84. [CrossRef]

30. Khandelwal, P.; Konar, J.; Brahma, B. Training RNN and it's variants using sliding window technique. In Proceedings of the 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 22–23 February 2020; pp. 1–5.

31. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

32. Brunini, O.; Lisbão, R.; Bernardi, J.; Fornasier, J.; Pedro Júnior, M. Temperatura-base para alface cultivar "White Boston", em um sistema de unidades térmicas. *Bragantia* **1976**, *35*, 213–219. [CrossRef]

33. Arnold, C.Y. The determination and significance of the base temperature in a linear heat unit system. In Proceedings of the American Society for Horticultural Science, Alexandria, VA, USA, 1 January 1959; Volume 74, pp. 430–445.

34. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. 2015. Available online: https://doi.org/10.1007/978-0-387-39940-9_565 (accessed on 12 February 2022).

35. Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F. Efficient and Robust Automated Machine Learning; In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc. 2015; Volume 28. Available online: https://proceedings.neurips.cc/paper/2015/hash/11d0e6287202fced83f799 75ec59a3a6-Abstract.html (accessed on 13 February 2022).

36. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1946–1956.

37. Anzanello, R.; de Christo, M.C. Temperatura base inferior, soma térmica e fenologia de cultivares de videira e quivizeiro. *Rev. Ciências Agroveterinárias* **2019**, *18*, 313–322. [CrossRef]

38. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 13 February 2021).