Workshops at 18th International Conference on Intelligent Environments (IE2022)
H.H. Alvarez Valera and M. Luštrek (Eds.)

doi:10.3233/AISE220028

Machine Learning Forecast of Soybean Yields on South Brazil

Lilian HOLLARD [a], Angelica DURIGON [b] and Luiz Angelo STEFFENEL [a,1]

[a] *Université de Reims Champagne Ardenne, LICIIS - LRC CEA DIGIT,*
51687 Reims Cedex 2, France

[b] *Universidade Federal de Santa Maria, Centre de Ciências Rurais,*
97105-900 Santa Maria - RS, Brazil

Abstract. This article investigates the problem of agricultural yield prediction, including that of soybeans. Soya is a very nutritious legume and one of the species producing the most protein per hectare. It is used in particular as a source of protein in animal feed and as an oilseed. However, agricultural systems are dependent on climatic variability, and farmers must deal with this factor to optimize their activities. This article describes a method for predicting soybean yields based on machine learning. The comparative study shows that one can obtain forecasts with less than 2% margin of error using the Random Forest algorithm. In addition, the results obtained in this study can be extended to many other crops such as maize or rice.

Keywords. Machine Learning, Regression, Random Forest, Soybean, Yield forecast

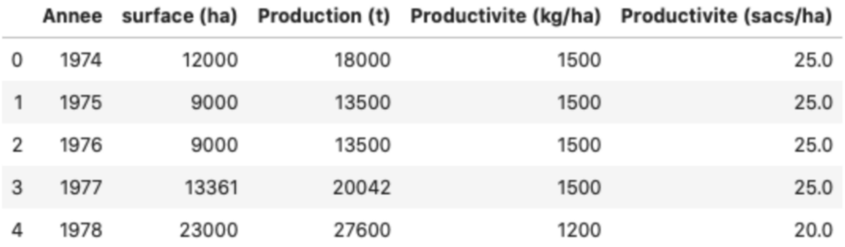
1. Introduction

In the field of artificial intelligence for agriculture, several applications of Data Mining and Machine Learning can be used to help farmers, bringing information about crop estimations and weather conditions. Indeed, [1] have implemented the K-Means algorithm to predict pollution in the atmosphere, while the K Nearest Neighbor method is applied [2] to simulate daily precipitation as well as other meteorological variables. Crop prediction is also a popular subject [3], and several machine learning algorithms have been studied to support crop yield prediction research [4].

This work is oriented toward the prediction of agricultural yields, focusing on soybean. Soybean is a very nutritious legume that counts among the species producing the most protein per hectare. Therefore, it is used today as a source of protein for animal feeding, which drives massive exports from producing countries to meet an increasing demand worldwide.

With the world population increasing from one billion at the beginning of the 20th century to more than seven billion today, food needs have not stopped rising. A considerable increase in world production is thus necessary in the context of climate change. Agricultural systems are thus dependent on climate variability, and farmers must deal with this factor to make their activities as profitable as possible.

[1] Corresponding Author. E-mail: angelo.steffenel@univ-reims.fr



	Annee	surface (ha)	Production (t)	Productivite (kg/ha)	Productivite (sacs/ha)
0	1974	12000	18000	1500	25.0
1	1975	9000	13500	1500	25.0
2	1976	9000	13500	1500	25.0
3	1977	13361	20042	1500	25.0
4	1978	23000	27600	1200	20.0

Figure 1. Crop Statistics for Santa Maria, RS, Brazil

In order to forecast the production of a given year, we rely on historical data and machine learning models to identify patterns and tendencies according to relevant features. Our algorithms and machine learning methods are all based on a combination of historical yield records and weather data. We will then focus on soybean as a model for our research, using relevant data from two producing regions in the South of Brazil.

This paper is organized as follows. Section 2 presents some definitions and introduces the datasets used in this work. Section 3 describes the data preparation steps, while Section 4 presents the machine learning models retained for this work. In Section 5, we present the first results for the case of soybean in the county of Santa Maria, while Section 6 demonstrates the generalization of the model to other crops and areas, proving the interest of the approach. Finally, Section 7 presents our conclusions on this preliminary work and discusses the future works.

2. Definitions and Datasets

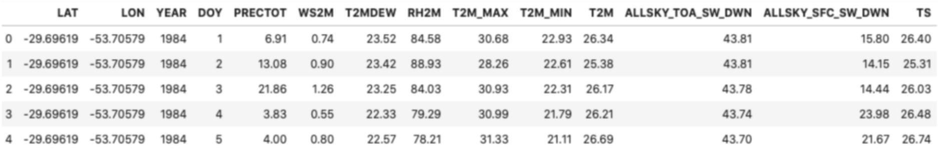
Soybean is a species of legume widely grown for its edible bean. Plant sowing takes place from late spring to early summer, and soybeans must have warm soil to germinate and grow. Indeed, to reach maturity, soybeans must accumulate 1400°C in the ground; from this point, farmers usually add a week to get the harvest date. Soil temperature is, therefore, the main feature for the forecast, but other data such as pluviometry and temperature ranges can help to obtain a precise forecast [4].

Helping producers estimate their harvest becomes a key advantage in today's agriculture: from harvest and storage planning to the price of products and commodities, all the logistic and economic chain depends on these values.

Our prediction analysis of soybean yield depends on two datasets. The first one includes crop statistics from 1974 to 2019 for the county of Santa Maria, RS, Brazil (53.7°S, 29.8°W, 150m from the sea level), providing information such as the average yield in kg/ha and the cultivated surface. An excerpt of this dataset is presented in Figure 1. This dataset provides, therefore, a base of explained variables.

The second dataset concerns meteorological and solar-related parameters from NASA's POWER project[2], which provides a list of explanatory variables. We speak of explanatory variables when the variable explains the explained variable. In other words, our weather data must explain soybean production.

[2] `https://power.larc.nasa.gov/data-access-viewer/`



	LAT	LON	YEAR	DOY	PRECTOT	WS2M	T2MDEW	RH2M	T2M_MAX	T2M_MIN	T2M	ALLSKY_TOA_SW_DWN	ALLSKY_SFC_SW_DWN	TS
0	-29.69619	-53.70579	1984	1	6.91	0.74	23.52	84.58	30.68	22.93	26.34	43.81	15.80	26.40
1	-29.69619	-53.70579	1984	2	13.08	0.90	23.42	88.93	28.26	22.61	25.38	43.81	14.15	25.31
2	-29.69619	-53.70579	1984	3	21.86	1.26	23.25	84.03	30.93	22.31	26.17	43.78	14.44	26.03
3	-29.69619	-53.70579	1984	4	3.83	0.55	22.33	79.29	30.99	21.79	26.21	43.74	23.98	26.48
4	-29.69619	-53.70579	1984	5	4.00	0.80	22.57	78.21	31.33	21.11	26.69	43.70	21.67	26.74

Figure 2. POWER dataset for Santa Maria, RS, Brazil

Daily values covering the entire period from sow to harvest are used in the model, accounting for climatic variations and weather conditions over the years. Figure 2 shows an example of the dataset, whose variables are the following:

- LAT - Latitude
- LON - Longitude
- YEAR - Year (format YYYY)
- DOY - Day of the year (1 to 365)
- PRECTOT - Total precipitation (mm)
- WS2M - Wind speed at 2 Meters (m/s)
- T2MDEW - Dew temperature at 2 Meters (Celsius)
- RH2M Relative Humidity at 2 Meters (%)
- T2M_MAX - Maximum temperature of the day at 2 Meters (Celsius)
- T2M_MIN - Minimum temperature of the day at 2 Meters (Celsius)
- T2M) Temperature at 2 Meters (Celsius)
- ALLSKY_TOA_SW_DWN - Top-of-atmosphere Insolation (MJ/m^2/day)
- ALLSKY_SFC_SW_DWN - All-Sky Insolation Incident on a Horizontal Surface (MJ/m^2/day)
- TS Soil temperature (Celsius)

Please note that weather information can be improved with on-site sensors. Indeed, several initiatives such as [5] aim at deploying sensors to characterize better both weather and soil conditions, which can lead at term to more precise models.

3. Data Preparation

Before applying different machine learning algorithms, we had to prepare the dataset. The first decision was on the covered period. While the production dataset covers crops since 1974, the POWER dataset only has data since 1984. For this reason, training and test subsets were selected only in the period from 1984 to 2018. Similarly, the validation subset comprises harvest and climate data from the 2019 season.

3.1. Data Correlation

Hence, we applied a correlation matrix to assess the dependence between several variables at the same time, as seen in Figure 3. The result then contains a correlation coefficient calculated by different methods of correlation tests: Pearson, Kendall, and Spearman correlation. This method makes it possible to establish a link between several random variables, which contradicts their independence. The interest in our case is to select

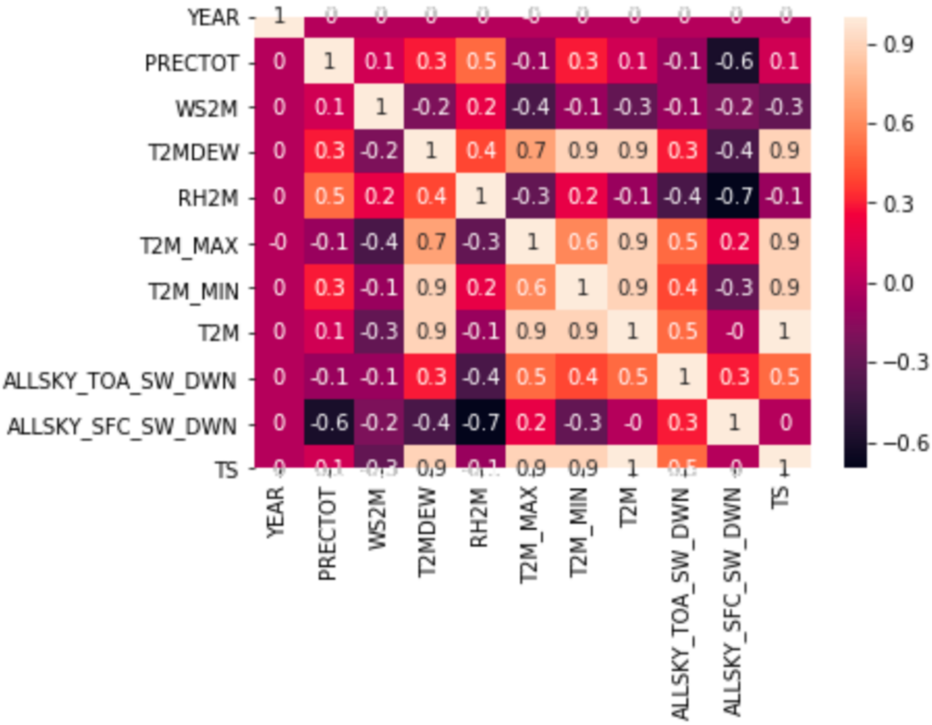


Figure 3. Correlation Matrix on POWER variables

explanatory variables which would improve the accuracy of our models and eliminate certain variables having no influence.

Thanks to this research, we observed that the year variable has a significant weight on the decision-making of our algorithms. Indeed, agricultural production has evolved since the 1980s due to the evolution of agricultural techniques and genetic improvement of the plants. However, as this impact is not directly related to the year itself, we propose using an inflation rate coefficient as presented in Equation 1.

$$Inflation\ rate = \frac{current\ year\ production}{previous\ year\ production} \tag{1}$$

The results being more interesting on the three algorithms with this coefficient, we decided to keep it. In addition, this coefficient follows a logic over time, which allows us to determine a coefficient for an unknown year using the average of the 5 or 10 previous years.

3.2. Scaling

Some algorithms are also sensitive to the variations in the scale of values between each characteristic. Indeed, several linear regression algorithms regularize the linear regression by imposing a penalty on the size of the coefficients. Thus, the coefficients are narrowed towards zero and each other. If the independent variables do not have the same scale, the shrinkage does not behave well. On the contrary, algorithms based on decision trees are less sensitive to data scaling and do not require scaling.

Using scikit-learn [6], we opted for the scaling technique called the Standard Scaler. This technique recalculates each feature so that the data is centered around 0 and scaled to obtain a variance of 1.

3.3. Metrics

Finally, to determine the robustness of the models, we compare their R^2 and RMSE scores. In statistics, Pearson's coefficient of determination denoted denoted R^2 is a measure of the quality of the prediction of linear regression. This coefficient, between 0 and 1, increases with the adequacy of the regression to the model:

- If the R^2 is close to zero, then the regression line sticks to almost none of the points.
- If the R^2 of a model is 0.50, then the points can explain half of the variation observed in the calculated model.
- If the R^2 is 1, then the regression determines 100% of the distribution points.

The other relevant indicator is the RMSE. This index provides information with respect to the dispersion or variability of the quality of the prediction. However, RMSE values are often difficult to interpret since we may not be able to tell if a variance value is low or high. This indicator can vary a lot, depending on the context:

- An RMSE of 10 is relatively low if the average of the observations is 500.
- On the contrary, a model will have a high variance if it leads to an RMSE of 10 while the average of the observations is 15.

Indeed, in the first case, the variance of the model corresponds to only 5% of the average of the observations, whereas the variance reaches more than 65% in the second case.

4. Machine Learning Models

This section compares different machine learning algorithms, looking for an accurate forecast model. Indeed, this is a regression problem as we must determine the soybean production in kg/ha as precisely as possible.

In order to determine the different algorithms to use in our work, we observed the characteristics of the dataset. Indeed, most machine learning algorithms used in crop yield estimation [7,8,4] require large amounts of data to perform well but, in our case, we require algorithms that support less than 100,000 data entries. Two groups of algorithms stand out: linear regression and random forest regression. For the same reason, we decided against more complex algorithms such as Neural Networks.

4.1. Linear Regression

Linear Regression is a family of statistic methods that aim to create a linear model with coefficients $w = (w_1, \dots, w_n)$ by minimizing the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation.

Table 1. Test scores for different ML Models

Model	R^2	% RMSE
ElasticNet	0.62	450
Ridge	0.62	400
Random Forest	0.9	300

Linear regression methods require a minimum of entries, making them good candidates here.

Among the Linear Regression methods, we selected Elastic Net and Ridge Regressor. The first one can react with little data input among the entire dataset, while the second one is known as having good performances with a complete dataset [9]. These two regularized regression methods will distort the space of solutions to prevent the appearance of too high values. This involves modifying the cost function of the linear regression problem by supplementing it with a penalty term [10].

4.2. Random Forests

The Random Forests is an algorithm proposed by Leo Breiman in 2001 [11]. This algorithm is based on a collective decision over an ensemble of decision trees. Random Forests are quick to train and produce results that are generalizable and intuitive to understand.

As stated above, the operation of the Random Forest trains a set of independent decision trees. Each tree has a fragmented vision of the problem (features and/or data entries) and, in the end, all these independent decision trees are grouped.

Indeed, for each independent tree, the training algorithm will look among all the possible decisions and choose the one that seems the best to serve as the root (first decision). The algorithm then adds branches to the tree using the Gini score. The Gini score is a value between 0 and 1, indicating how likely the tree is to make a decision error. Each time it is necessary to add a branch, this score will be calculated to assess the relevance of the decision that the algorithm wishes to add. Finally, the decision that allows the tree to guess or calculate the result most correctly is selected and grafted to the tree.

While there is no guarantee that a single tree is optimal, Random Forests explores different alternatives by varying the subsets of features and data entries for each tree and makes its decisions by voting on all the decision trees that compose it. As a result of this ensemble approach, a more robust model arises.

5. Fitting Harvest Production

During the training phase, we seek to improve scores through Hyperparameter Optimization research. This step involves choosing a set of optimal hyperparameters for a learning algorithm. Indeed, machine learning algorithms propose parameters whose value is used to control the learning process, called hyperparameters. As a result, Table 1 presents the results obtained with the models identified in Section 4.

When applied to 2019 data, these models deliver yield forecasts that are compatible with the R^2 and RMSE from Table 1. Indeed, the predictions listed in Table 2 show a clear advantage for the Random Forest model, which was able to forecast the average soybean production in 2019 with a 20kg error/ha (0.6% error margin).

Table 2. Soybean yield forecast for 2019 in Santa Maria

Model	Prediction	Real production performance
ElasticNet	2104kg/ha	3300 kg/ha
Ridge	2234 kg/ha	3300 kg/ha
Random Forest	3280 kg/ha	3300 kg/ha

Table 3. Random Forest yield forecast for different crops

Location	Crop	Real production	Error margin
Santa Maria	Soybean	3300 kg/ha	20 kg/ha (0.6%)
	Rice	7060 kg/ha	39.77 kg/ha (0.56%)
Campos Novos	Soybean	4080 kg/ha	66.6 kg/ha (1.63%)
	Maize	11400 kg/ha	200 kg/ha (1.75%)

6. Generalization to other Crops and Locations

To conclude our study, we expanded the scenarios to assess the robustness of the Random Forest model. Hence, we looked for other crops in Santa Maria as well as crops in the county of Campos Novos (27°24'06"S,51°13'30"W, 946m elevation), located 450km away from Santa Maria. This choice was driven by the climatic differences between these towns and the variable impact of weather conditions on different crops.

Table 3 summarizes these tests. It demonstrates that the choice of meteorological features and a Random Forest model allow for a robust forecast with limited error margins. Indeed, we obtain less than 2% error, a value that can still be improved with further optimizations and more data.

7. Conclusion and Future Works

This paper presents a preliminary study on how machine learning techniques can help cereal growers to forecast their future crops. In this first work, we concentrated on identifying which features contribute most to the accuracy of the models and how different machine learning models perform.

The results suggest a clear advantage of Random Forests models over more traditional linear regression models, the latter being more sensitive to data scaling and variations in the dataset.

In addition, we applied the Random Forest model to other crops and locations, obtaining excellent results when predicting the average productivity in kg/ha. In all the studied scenarios, we obtained less than 2% of error margins, proving the interest of the approach.

The present model is being integrated into a web bulletin board application that provides updated estimations as the season advances, in accordance with new weather information. We believe, however, that fine-grain models can be obtained by combining a county-level model with farmers' own data, including their historical production, soil details, weather data from IoT sensors deployed on the fields [5] as well as some other crop specificities (plant variety, sowing date, organic agriculture...).

References

[1] Singh G. Classification and Clustering in Yield Prediction based on Soil Properties. International Journal of Advanced Research in Computer Science. 2017 08;8:253-8.

[2] Amonkar Y, Farnham DJ, Lall U. A k-nearest neighbor space-time simulator with applications to large-scale wind and solar power modeling. Patterns. 2022;3(3):100454. Available from: https://www.sciencedirect.com/science/article/pii/S2666389922000277.

[3] Chemchem A, Alin F, Krajecki M. Combining SMOTE Sampling and Machine Learning for Forecasting Wheat Yields in France. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE); 2019. p. 9-14.

[4] van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture. 2020;177:105709. Available from: https://www.sciencedirect.com/science/article/pii/S0168169920302301.

[5] Coppola M, Noaille L, de Oliveira RO, Pierlot C, Gaveau N, Rondeau M, et al. Innovative Vineyards Environmental Monitoring System Using Deep Edge AI. In: Vermesan O, John R, Lucca CD, Coppola M, editors. Artificial Intelligence for the Digitizing Industry. Rivers Publisher; 2021. p. 261-78.

[6] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.

[7] Sun J, Di L, Sun Z, Shen Y, Lai Z. County-level soybean yield prediction using deep CNN-LSTM model. Sensors. 2019;19(20):4363.

[8] Terliksiz AS, Altÿlar DT. Use of deep neural networks for crop yield prediction: A case study of soybean yield in lauderdale county, alabama, usa. In: 2019 8th international conference on Agro-Geoinformatics (Agro-Geoinformatics). IEEE; 2019. p. 1-4.

[9] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning; 2013. p. 108-22.

[10] Geron A. Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly Media; 2017.

[11] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. Available from: http://dx.doi.org/10.1023/A%3A1010933404324.