Final Report

Week 10

[redacted]

–

Team 34

[redacted]

Content

Context

Since the seventeenth-century, man has ceased to depend on the annual dynamics of nature for the capture of food; in fact, it has begun to recreate most natural scenarios, so that it has a controlled environment where it can manipulate food production [1-3]. However, although the ability to control certain environmental variables allows human beings to have partial independence from natural external conditions, some factors have an irremediable impact on food production, such as changing weather conditions, geopolitical and economic conditions, among others [4].

In recent decades, Latin America and the Caribbean have made significant progress in reducing rural hunger and poverty [5]. However, it is estimated that almost half of the rural population is poor and one third live in extreme poverty.

Most of the rural inhabitants dedicate themselves to agriculture as the main means of subsistence and, in general, they work on a small scale, based on family labor, both for men and women. This productive sector is key to regional food security but faces significant limitations in terms of access to productive resources, social services, basic infrastructure, services, financing, and extension.

Conversely, according to National Geographic [6], Colombia is the most biodiverse country in the world by area, however, the agriculture that is carried out in the country is substantially affected by thermal floors, meteorological, economic and political aspects, among others. In addition, farmers in Colombia do not have a systematized method that allows them to identify which is the most suitable product to farm and when to farm it, in accordance with these meteorological variables (such as natural phenomena of "El Niño" and "La Niña") and economic factors (such as the variation of the currency, of the product price, among others). Other variables that influence this identification include crop time, production time, changes in soil composition and variations in the market. Nowadays, the Colombian peasant follows a series of popular cultural beliefs in a sometimes improvised exercise of high financial risk, since the price of the product can be significantly low when obtaining the crop. Additionally, this causes the following issues:

- A waste of the food that is grown, due to the loss of current crop production.
- A waste of the natural resources used for the unsuccessful cultivation of the product.
- An increase in hunger and poverty of the population in the sowing regions.
- The bankruptcy of many of the farmers.

Project Definition

This section details the scope of the business solution developed and the workbench defined for its development.

●.1. Business Solution

Our solution offers to the farmers a web application able to suggest which is the more suitable product and when it should start to farm. Our solution have this capability based on the analysis of years of market behaviour and crop classification, obtained from governmental entities and historical information about climate phenomenon and market trends.

●.2. Solution Scope

The solution is framed in the use of the techniques developed in the department of Cundinamarca, one of the departments with the greatest agricultural projection in the country. Three main products were considered that are grown in the thermal floors of this department and whose production, among others, depends on its economy: potatoes, onions and corn.

If more information about the project definition and prototype versioning is needed, please refer to Appendix A - Project Workbench and Prototyping document.

● Dataset Description

This section describes the technical summary of each of the datasets used for the development of the proposed solution. If more information about where to find the original data is needed, please refer to Appendix B - Dataset Description document

●.1. Datasets Summary and Tables Design

The following table summarizes the main characteristics of the datasets that we will take into account throughout the development of our project:

Table 1. Summary of the main characteristics of the datasets.

Name	From	To	Daily	Monthly	Yearly	Source	Table Name
Wholesale Prices	2010	today	X			link	commodities_prices
Commodities Stock	2013	today	X			link	commodities_stock
Name	**From**	**To**	**Daily**	**Monthly**	**Yearly**	**Source**	**Table Name**
Inputs and Labor Costs	2013	today		X		link	inputs_cost
Climatic Events	2012	today		X		link	climatic_events
Total Monthly Precipitation	1938*	May 2018		X		link	precipitation
Mean, Maximum, Minimum Temperature	1938*	May 2018, 2016, 2016		X		link link link	temperature_stats
Precipitation, temperature, humidity and wind speed	2018	2019	X			sent by mail	ideam_weather
Production Chain	2015	2018			X	link link	production_chain
Land aptitude	2018	2018				scraped from ws	upra_products, upra_aptitudes
Dark Sky	2000	today	X	X	X	link	

The datasets were uploaded to Postgres tables in AWS RDS. The equivalent relational tables are mapped in the summary that was included before and are shown below:

Figure 1. Entity-Relation diagram.

● Solution Architecture

The architecture diagram and the instruments used for the development of the proposed solution are as follows:

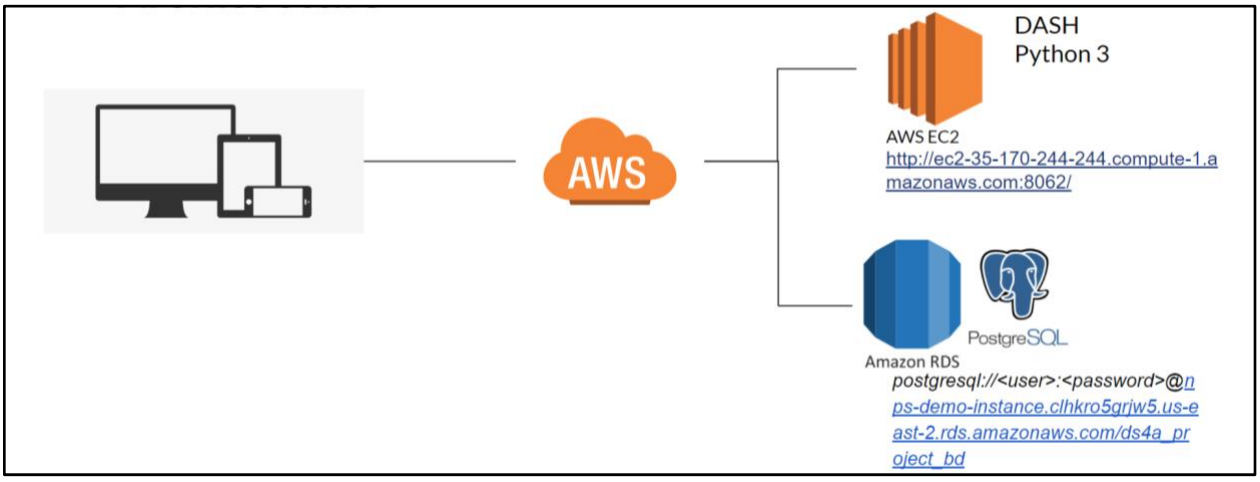


Figure 2. Architecture diagram for the proposed solution.

●.1. Tools Used:

- Spyder
- Google Collaborative
- PgAdmin
- Anaconda
- Dash
- PostGis

● Results

The results obtained after the follow-up of the stages and prototypes described in the previous sections are detailed below.

●.1. Datasets wrangling

The dataset wranging process was divided into three phases: structuring, cleaning and enriching.

●.1.1. Datasets Structuring

As data is given have some aggregated columns, the first step for the dataset wrangling process is to separate the relevant data into independent columns, so that the analyzes can be executed by grouping by common values in a separate way. In this step the following modifications were made to the dataset:

- The time registration column of each commodity, price, aptitude and weather tables was separated into columns that allowed identifying the day, month and weekday, independently.
- The time at which the service was taken was extracted, so that it was in a format of hours, minutes and seconds.

●.1.2. Datasets Cleaning

In this step, the data is cleaned for high-quality analysis. The existence of null values values in the data was verified, corrections of the download code of the identified data were made from the analysis of null data and the data is re-verified by identifying null values in the imports_exports, inputs_cost, temperature_stats tables, upra_aptitude_variables, and upra_product_aptitudes.

Table 2 shows an example of this analysis in where it is detailed the number of null values present in the temperature dataset, corresponding to each of the records of the entities shown in the Entity-Relationship diagram of Figure 1. It can be seen that there were 15 over 30,336 null values in the temperature_stats dataset. These values were directly deleted because they were not representative in the sample and were values regarding very old registers of temperature.

Table 2. Dataset corresponding to Temperature.

Feature name	Number of null values
station_name	0
station_code	0
location	0
municipality_name	0

year	0
month	0
temp	15

The following considerations were taken into account with respect to the data:

- The null data points found in the geographics dataset of Cundinamarca region were due to the non-existence of records. This data points were deleted.
- The null values corresponding to exports in the CIF (Cost, Insurance and Freight) are due to the characteristics of exports, which don't have this kind of overcosts.
- The null values corresponding to the minimum and maximum temperatures in the temperature dataset were included, if and only if the mean temperature value was recorded. If this value was not measured, the entire row was deleted.

●.1.3. Datasets Enriching

After cleaning, the data is enriched, by augmenting some variables using some additional data in order to make it better for the processing stages. The following processes were executed:

- Product names were standardized in all tables and special characters were removed from the data. Product names were also transformed to uppercase names and corn crop was renamed from *mazorca* to *maiz*.
- The municipality codes were standardized according to the geographical information provided by the Colombian Government's Geoportal.
- The different types of data in the dataset were homogenized to numerical, dates and strings.
- A semester column was created on each dataset that recorded dates. This action was made to be able to compare the data across all data sources.
- The latitude and longitude columns were created as Geometric Points from Shapely, in order to standardize the geographical references and be able to join points into polygon regions.
- From the temperature table it was decided to filter the information from 1938 to 1991 because the average, min and max temperature values for month of the year by municipality were not found. Additionally, temperatures where the average were 0 or below were deleted because they go against the natural Colombian weather conditions.

●.2. Exploratory Data Analysis (EDA)

In the present section, a description of the Exploratory Data Analysis executed and the results obtained, are presented.

●.2.1. Correlated variables

First, an initial correlation analysis was made, considering the commodities prices, for selecting key features that were not related one with each other. It is observed in Figure 3 that the features

'extra_quality_price' (EQP) and 'prime_quality_price' (PQP) are highly correlated (0.82 > 0.75). These two variables are correlated due to the way that the prices are assigned to the products. If one regular price is assigned, it is normal that the prime and extra quality prices are assigned with a percentage increase over the regular price. Therefore these features were removed from the final dataset.

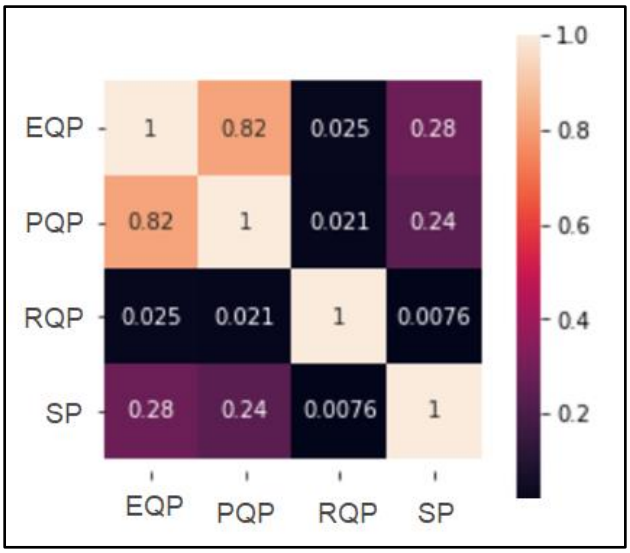


Figure 3. Correlation heatmap for commodities prices dataset.

●.2.2. Statistically significant variables

A ttest of the three prices of the defined products is carried out in order to verify the distribution of the standardized prices of the same, it is identified that

p-value in T-test between MAIZ and CEBOLLA: 0.9999999999999994

p-value in T-test between MAIZ and PAPA: 0.9999999999999959

p-value in T-test between CEBOLLA and PAPA: 0.9999999999999954

due p - value (0.99) is higher than α (0.05), hence, we have to retain the null hypothesis that the three groups have the same average...This helps us define the model to use for the three products, it can be the same

●.2.3. Plotting

In the Figures 4 to 6, some behaviours of the analyzed variables are plotted. This plotting procedure was done for merely data analysis and to give us an idea of the variables found in the datasets.

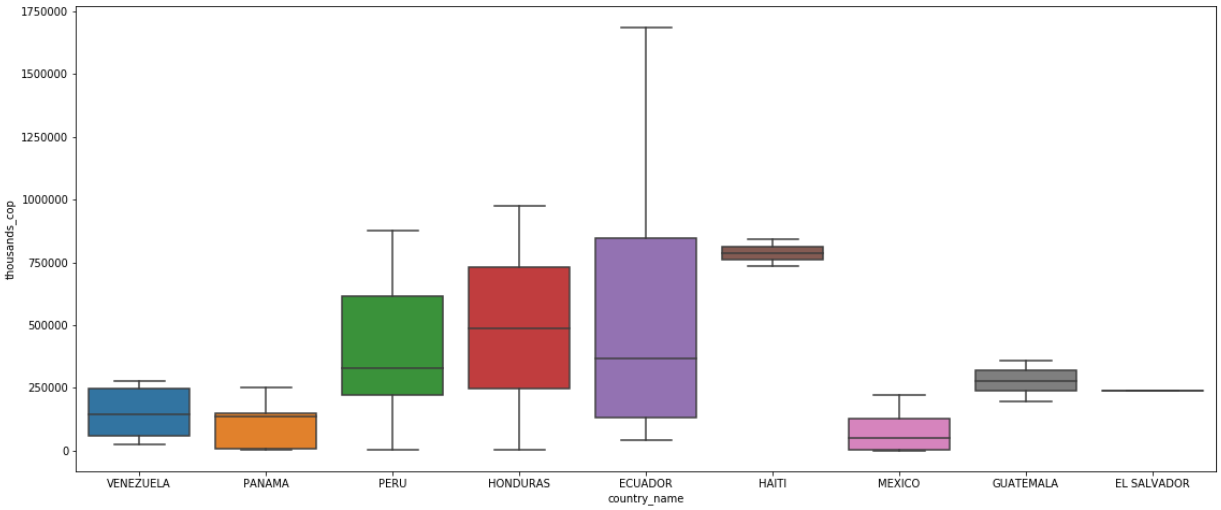


Figure 4. Top corn export countries.

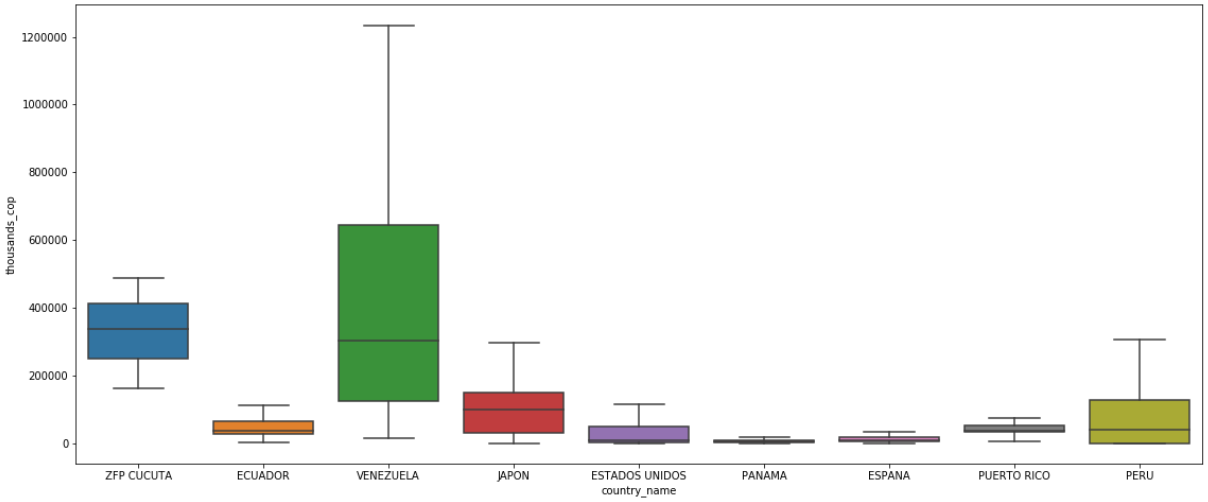


Figure 5. Top potato export countries.

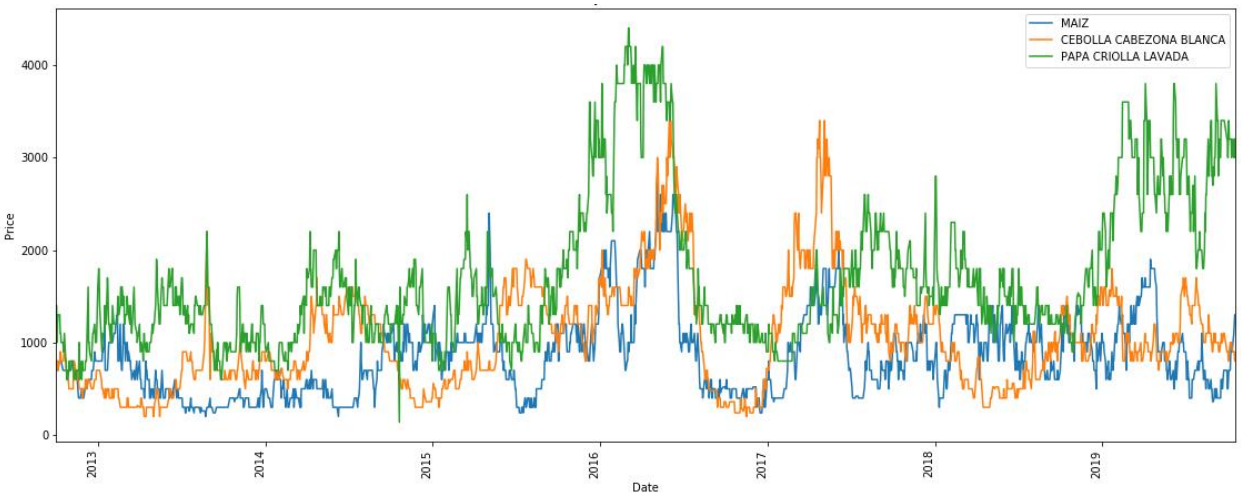


Figure 6. Commodity price distribution.

This last figure is particularly interesting, since it is the one that describes the behavior of the regular price of the three products chosen, in the department of Cundinamarca. This price turns out to be the variable of interest for the subsequent modeling study.

●.2.4. Model Selection

In the process of predicting future values of time series, it is not desirable to use a binary classification model, since the output values are not discrete. In this way, models such as logistic regression and Random Forest were initially discarded. A linear regression model was not taken into account for price predictions because the behavior of the variables considered is nonlinear, which makes association to a linear behavior difficult. A conventional neural network model can be adjusted to meet the prediction needs raised in this project. However, feature extraction is a complex process in time series and in most cases not all confounders that explain the behavior displayed in the series are considered. In addition, time series are difficult to forecast. Consequently, we chose a model adjusted to the processing of time series that allows to learn about the intrinsic patterns of the series. This model is that of LSTM (Long Short Term Memory) algorithm.

LSTM networks are particularly efficient with time series because they have feedback connections, which allows predictions to be executed based on stored historical data. Conversely, their architecture can learn to recognize and robustly generate precisely timed events separated by significant time lags. In this way, LSTM networks are well-suited for classifying, processing and predicting time series, and therefore, it was the model selected for the prediction of the prices of the proposed agricultural products.

●.2.5. Modeling

Three learning Sequential LSTM models were implemented, one for each product chosen. The models have the following characteristics: (a) 4 hidden nodes, (b) 1 prediction time step and (c) a lookback window of 120 samples. The models were evaluated in the training process based on the mean squared error metric, taking into account an Adam optimizer for the learning rate. The following hyperparameters were defined: 1000 training epochs and a batch size of 200 observations.

The Mean Absolute Percentage Error (MAPE) was used as a metric for the evaluation of the models in the testing set. This metric was used because it allows a clearer view on the performance of the models and a way to compare them. The MAPE prediction error of the potato model for the prediction of a sample is 2.99%, the onion model had an error of 3.56% and the corn model an error of 3.17%.

Figures 7, 8 and 9 show the behavior of the price of the daily product (in blue), the behavior of the daily prediction of the training set of the model (in orange) and the behavior of the daily prediction of the test set of the model (in green).

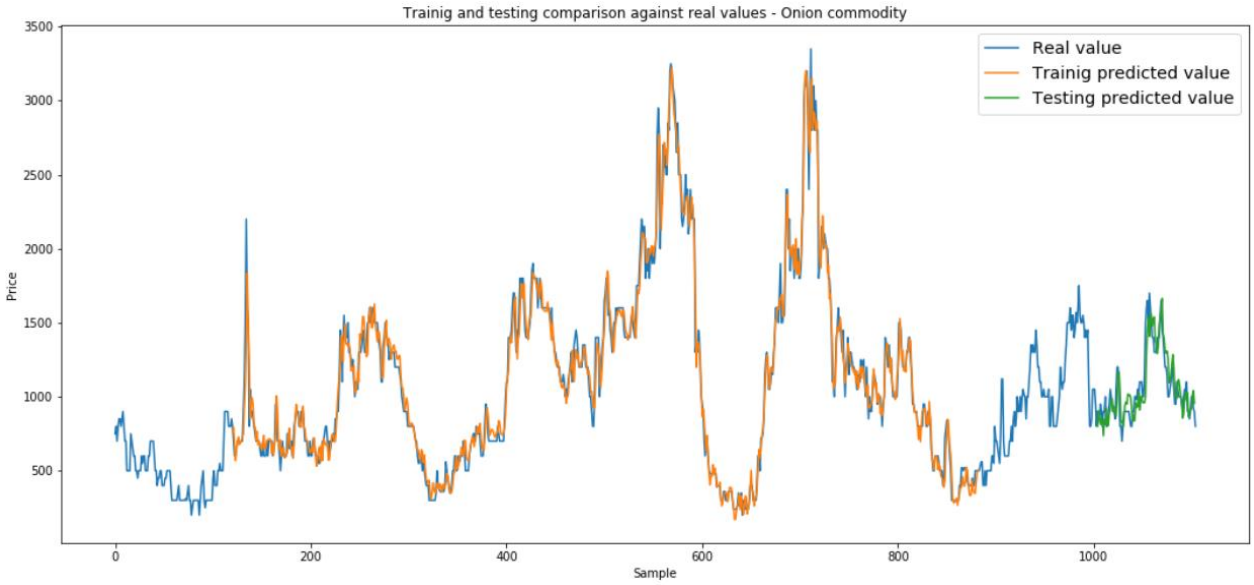


Figure 7. Trainig and testing behaviour comparison against real values on the onion product.

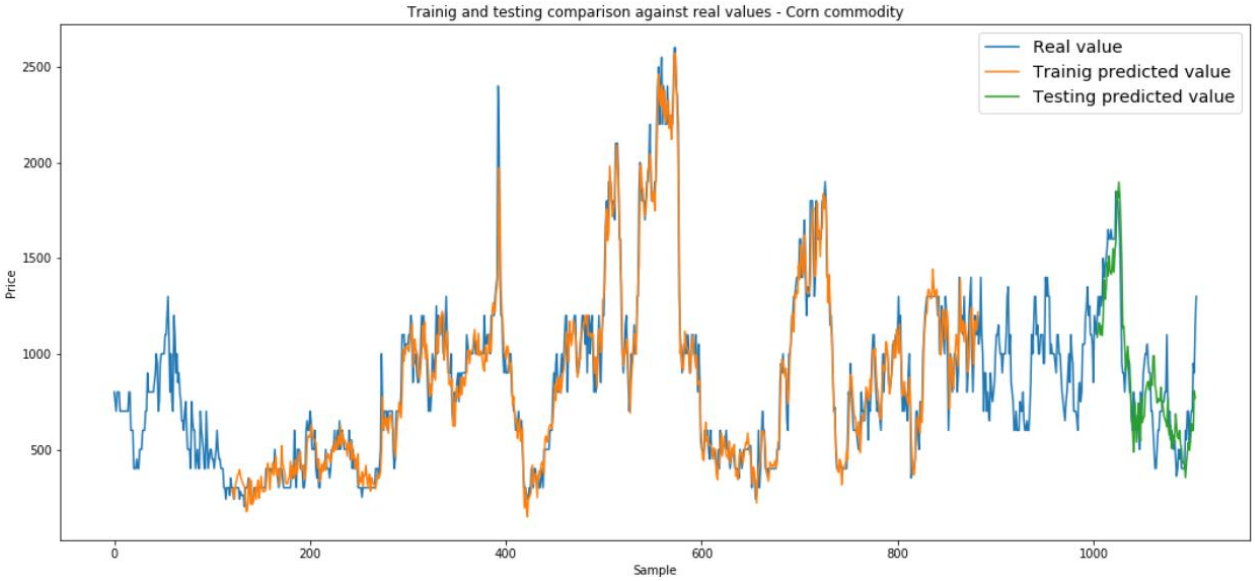


Figure 8. Trainig and testing behaviour comparison against real values on the corn product.

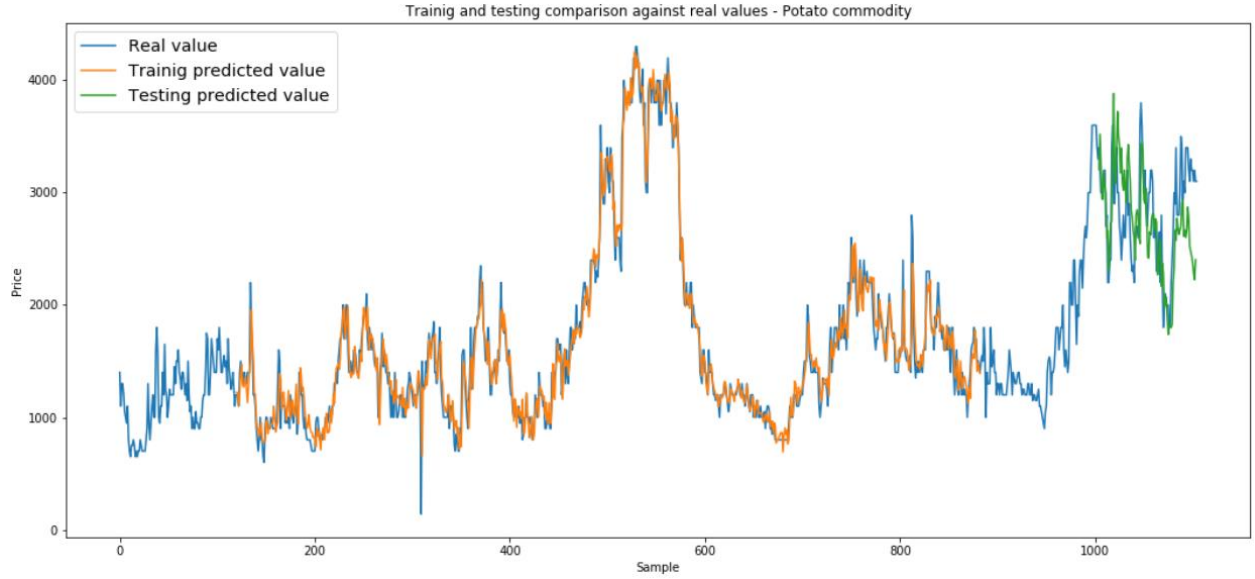


Figure 9. Trainig and testing behaviour comparison against real values on the potato product.

Figures 10, 11 and 12 show the behavior of the price of the daily product (in blue) in a sample of 100 days, the behavior of the daily prediction of the training set of the model (in orange) and the behavior of the prediction daily test set of the model (in green). It can be noted that the prediction has an error of less than 7% in all predictions for the products considered.

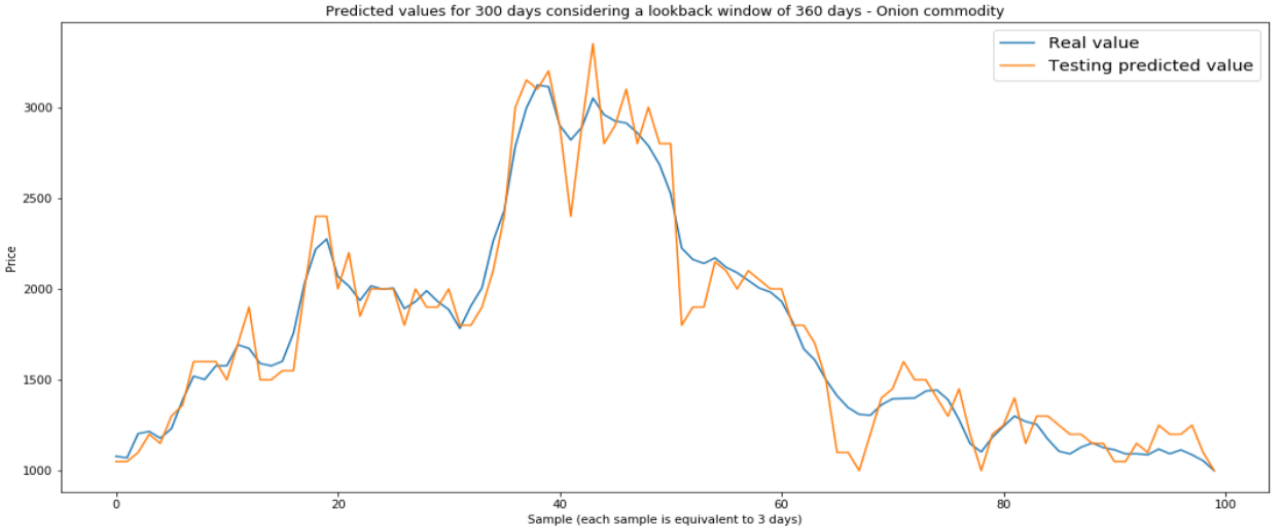


Figure 10. Predicted values for 300 days considering a lookback window of 360 days for onion product.

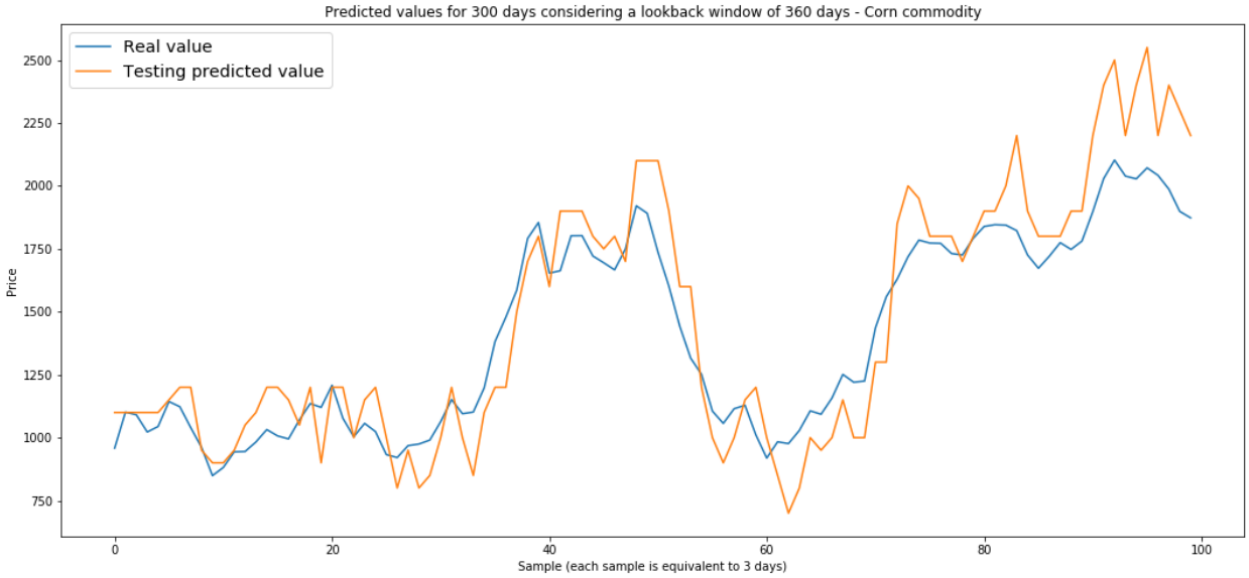


Figure 11. Predicted values for 300 days considering a lookback window of 360 days for corn product.

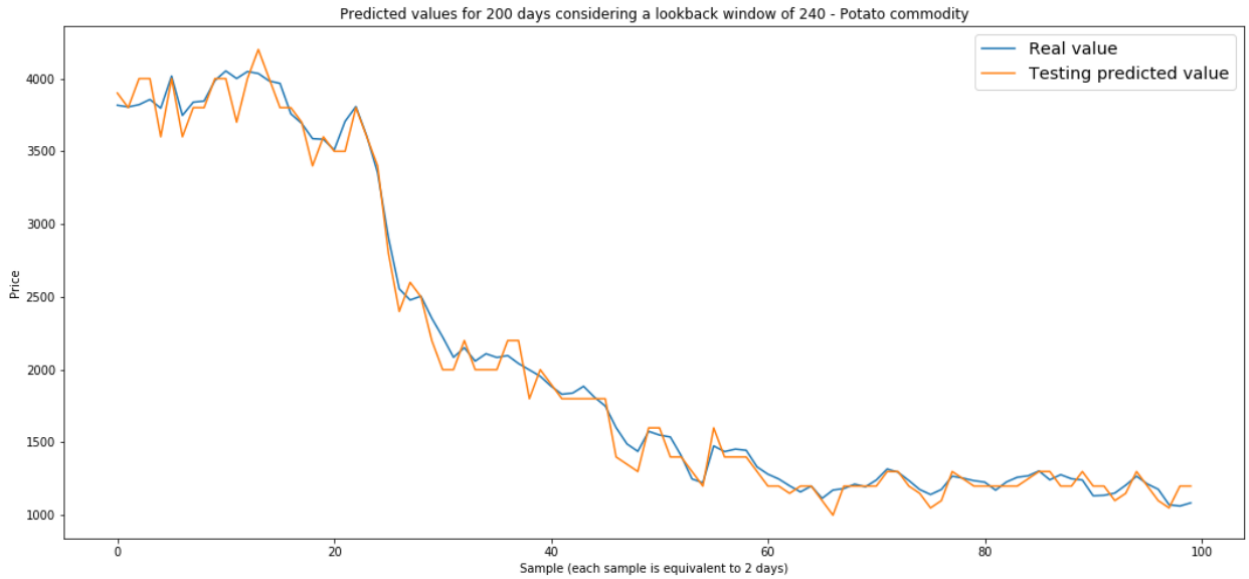


Figure 12. Predicted values for 300 days considering a lookback window of 360 days for potato product.

Figures 13, 14 and 15 show the behavior of the MAPE when the number of days of prediction is increasing. It can be seen that the more days are predicted, the more the prediction error increases. This is because the error spreads as a future prediction is running over a past prediction.

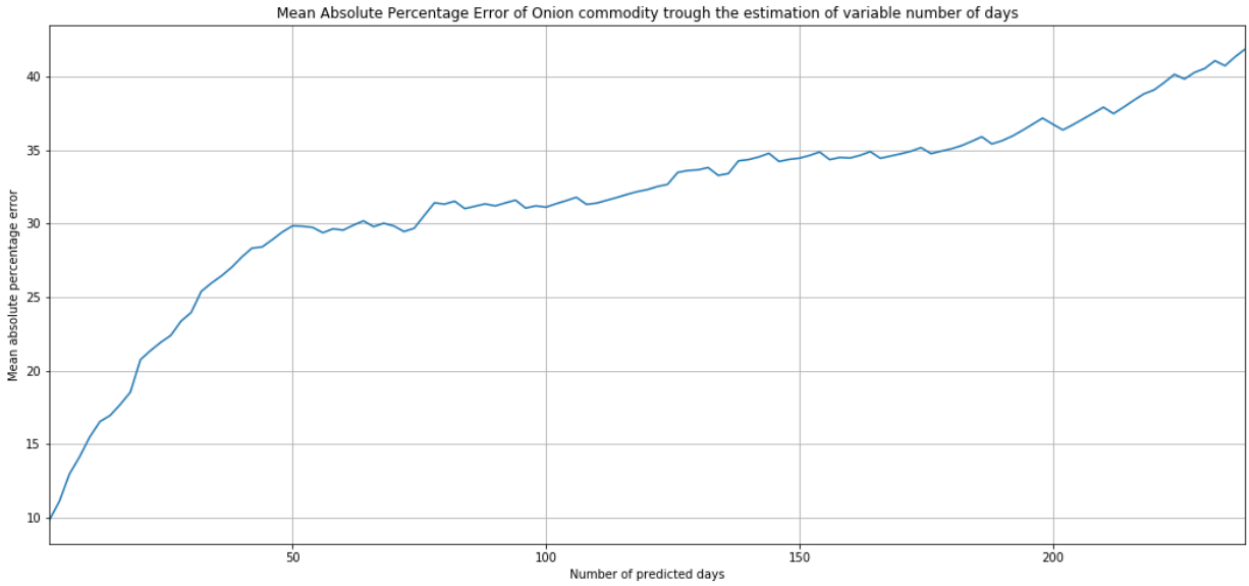


Figure 13. Mean Absolute Percentage Error of onion product through the estimation of variable number of days.

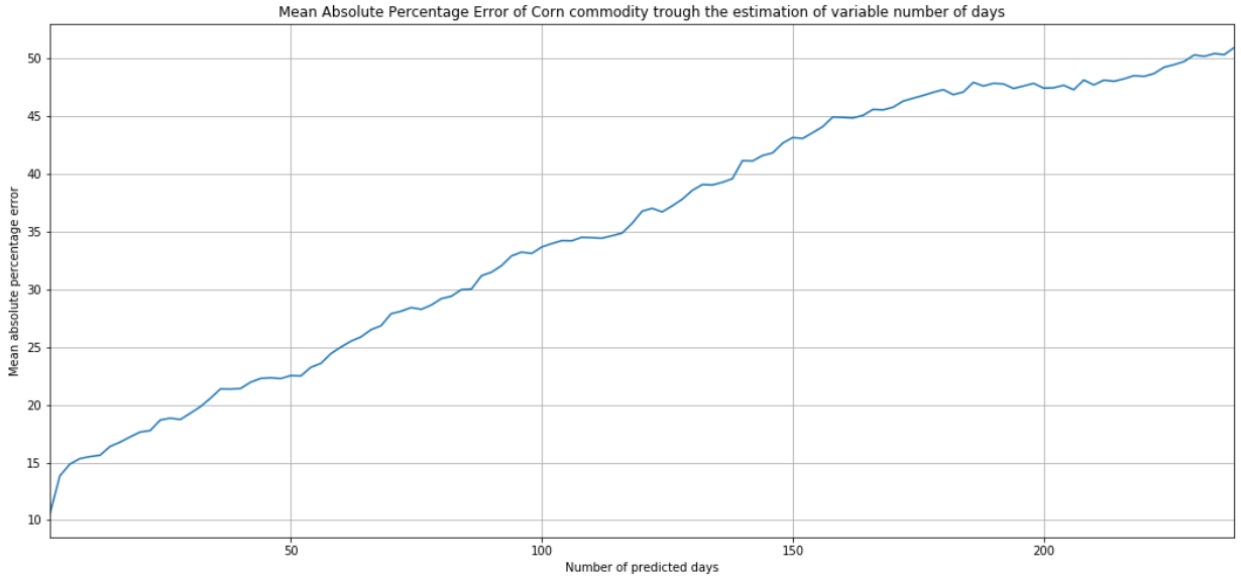


Figure 14. Mean Absolute Percentage Error of corn product through the estimation of variable number of days.

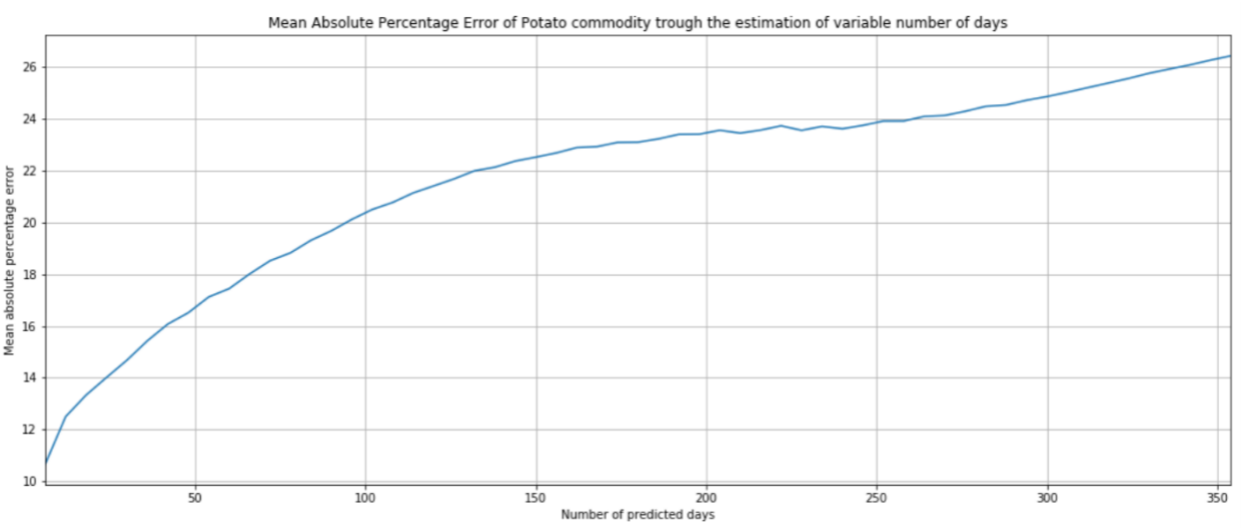


Figure 15. Mean Absolute Percentage Error of potato product through the estimation of variable number of days.

Figure 16 shows the complete behavior of a one-year prediction of the price of corn. The harvest period for a 30-day planting period is demarcated with two yellow lines. The green point corresponds to the peak of the price within the harvest window and as a green vertical line is the time of planting to obtain the maximum price found.

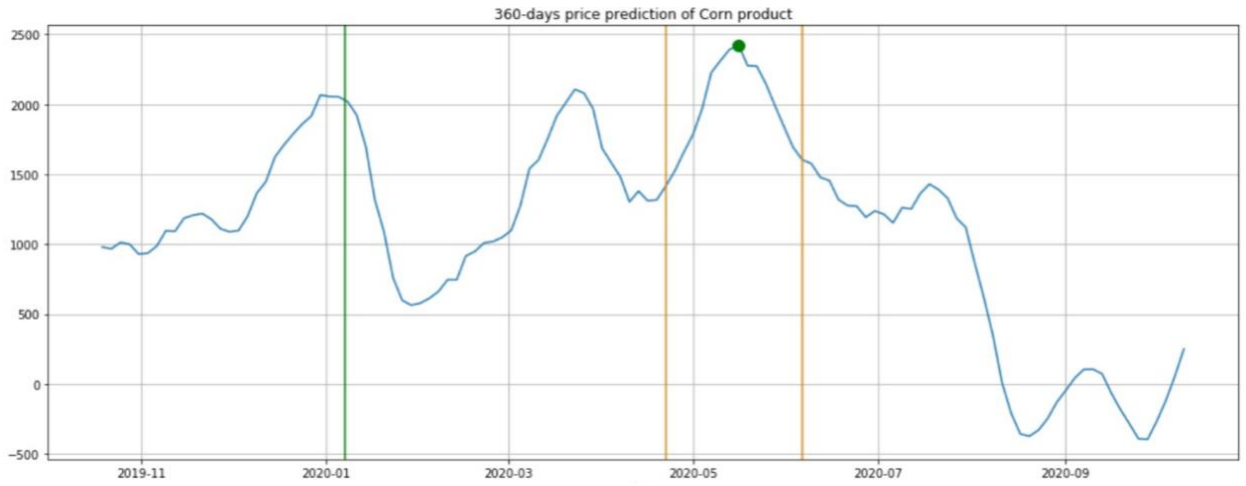


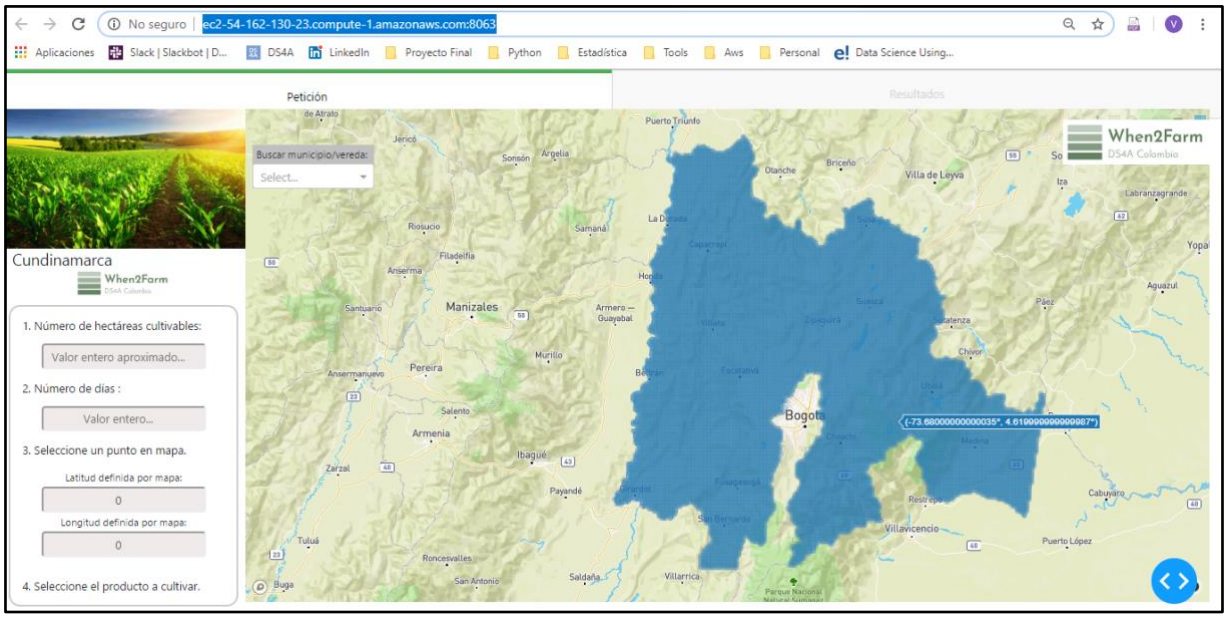
Figure 16. Complete prediction of corn price and window framing..

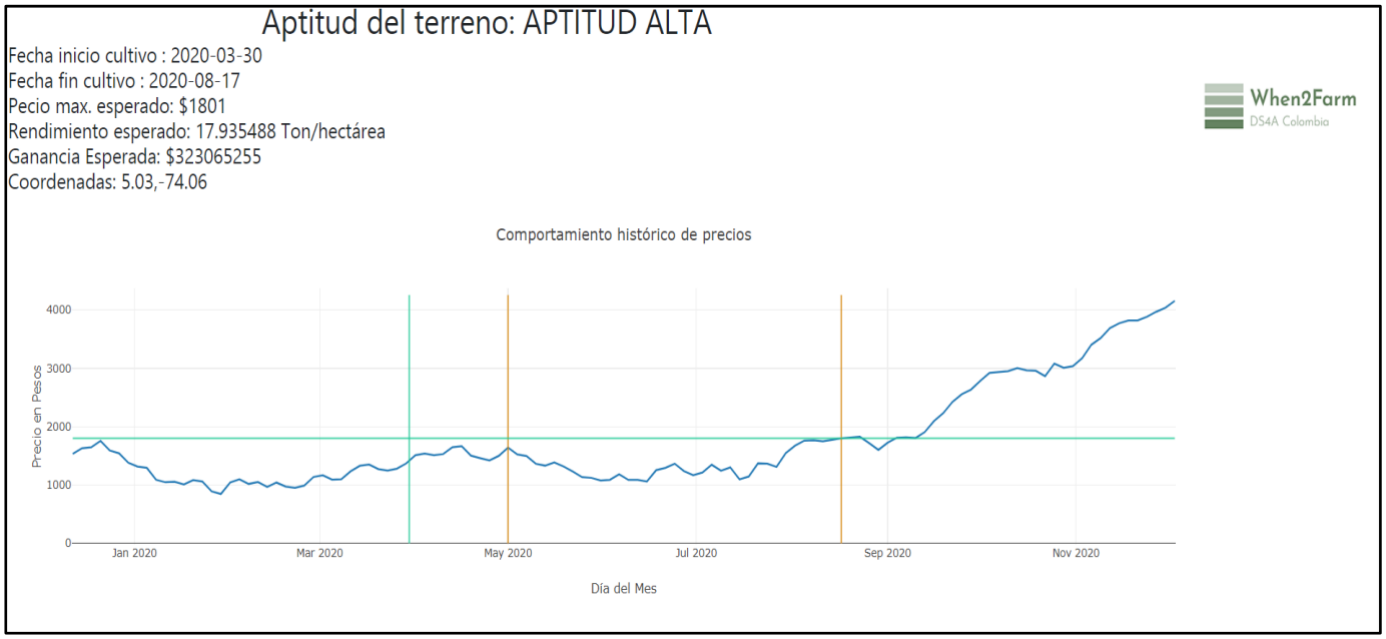
● Functional Dashboard

Our dashboard is published on the following link; some data is already connected to it, using the following connection URL:

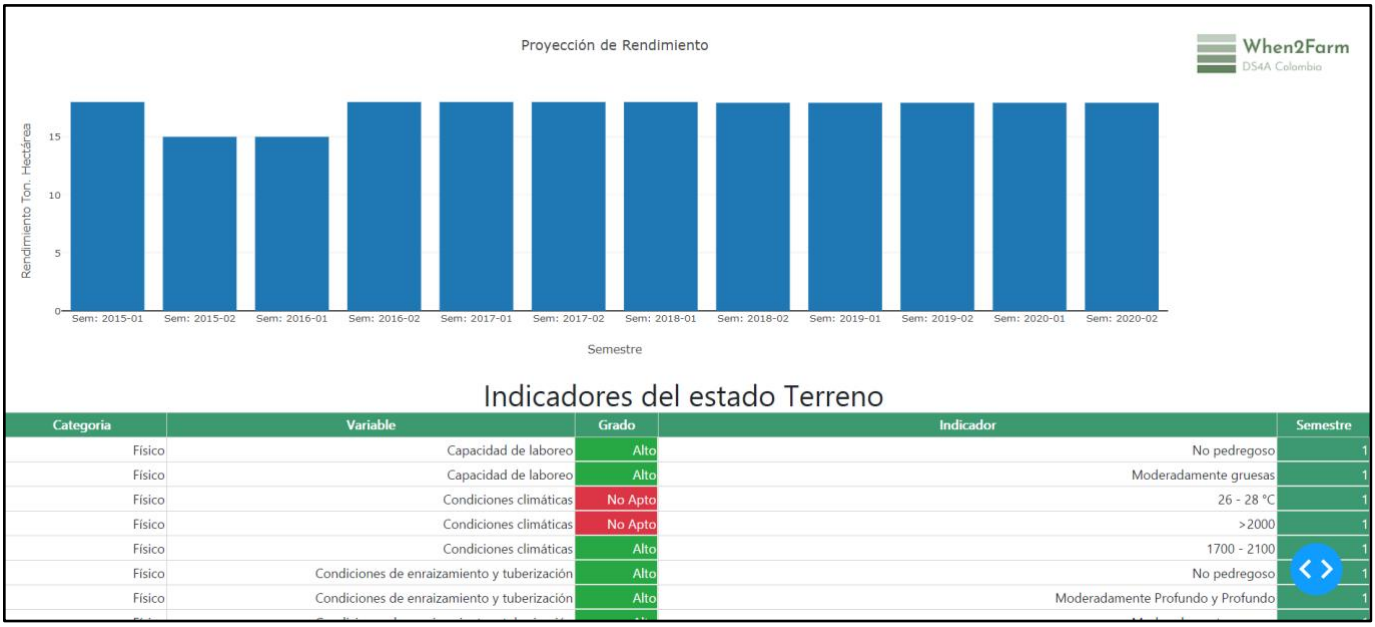
postgresql://<user>:<password>@nps-demo-instance.clhkro5grjw5.us-east-2.rds.amazonaws.com/ds4a_project_bd

●.1. Request interface:





●.2. Response interfaces:



Indicadores del estado Terreno

Categoria	Variable	Grado	Indicador	Semestre
Físico	Capacidad de laboreo	Alto	No pedregoso	1
Físico	Capacidad de laboreo	Alto	Moderadamente gruesas	1
Físico	Condiciones climáticas	No Apto	26 - 28 °C	1
Físico	Condiciones climáticas	No Apto	>2000	1
Físico	Condiciones climáticas	Alto	1700 - 2100	1
Físico	Condiciones de enraizamiento y tuberización	Alto	No pedregoso	1
Físico	Condiciones de enraizamiento y tuberización	Alto	Moderadamente Profundo y Profundo	1

● Conclusions

Agriculture in Colombia still uses very ancient methods, which have been used generation after generation. These methods are still valid and it is the purpose of this work to make a contribution

to improve decision making in this process. The foregoing because we believe that using new technologies such as data science can help to solve some of the problems that impact Colombian farmers on a daily basis.

The conclusions that are drawn from the scope of this work are:

Data science can help improve decision making and understand the complexities of the agricultural business context, by presenting time slots where farmers can define which products can represent a greater benefit

Advanced data science techniques such as neural networks can be used successfully to predict variables such as price and quantities for different products

Based on the Exploratory data analysis EDA, the price variation depends on diverse variables like weather (rain, temperature, humidity) of the location, imports, exports, local production of commodities.

The public data used in this exercise allowed us to make a model that allows us to predict the price of agricultural products quite accurately.

● Future work

In the development of this project, several additional use cases for the agricultural sector could be identified. Machine learning is everywhere throughout the whole growing and harvesting cycle. It begins with a seed being planted in the soil, soil preparation, seed water feed measurement and the harvest determining the ripeness with the help of computer vision.

We believe that additional uses cases that may extend the use of this work are related to crop management and livestock production are explained below:

Crop management

- Yield Prediction: Yield prediction is one of the most important and popular topics in precision agriculture as it defines yield mapping and estimation, matching of crop supply with demand, and crop management. it may also incorporate other tools and techniques from machine learning as a computer vision technologies to provide data on the go, comprehensive multidimensional analysis of crops, weather, etc.
- Crop Quality: The accurate detection and classification of crop quality characteristics can increase product price and reduce waste, improving the research and human labor.

Livestock management

- Livestock Production: Machine learning provides accurate prediction and estimation of farming parameters to optimize the economic efficiency of livestock production systems.

For example, weight predicting systems can estimate the future weights 150 days or animal species providing farmers information to modify diets and weights of animals.

References

[1] Witold Kula, An Economic Theory of the Feudal System: Towards a Model of the Polish Economy, 239 pp, ISBN: 978-0-86-09185-16, 1974.

[2] Maurice Dobb, Transition from feudalism to capitalism, ISBN: 978-8-47-42301-78, 278 pp. 1977.

[3] Perry Anderson, Lineages of the absolutist state, ISBN: 978-84-323-0362-3. 569 pp. 1974.

[4] FAO Regional Office For Latin America And The Caribbean. Factors affecting agricultural production, Annex 11. Available at: http://www.fao.org/3/i0515e/i0515e18.pdf.

[5] José Carlos Cardoso. Agrometeorología, la importancia de su desarrollo técnico y los sistemas de información y cooperación internacional. Available on: http://parlatino.org/pdf/comisiones/agricultura/exposicion/xv-agrometeorologia-pma-24-mar-2011.pdf

[6] FAO Regional Office For Latin America And The Caribbean. Family farming and inclusive food systems for sustainable rural development. Eradicate hunger and poverty, strengthening rural sector. Available on: http://www.fao.org/americas/prioridades/agricultura-familiar/en/. 2019.

[7] Unidad de Planificación Rural Agropecuaria (UPRA). Sistema Para la Planificación Rural Agropecuaria.Available on: https://sipra.upra.gov.co/. 2018.