MACHINE LEARNING PROJECT

TECHNICAL REPORT

tecnun
Universidad de Navarra

GROUP 4:
Daniel Aparicio
Pedro Arana
Juan Antonio Ruiz

INDEX

1.- BUSINESS PROBLEM DEFINITION

Augere is an innovative start-up in the agricultural sector that was born with the purpose of combating one of the most key challenges of humanity, food with limited resources.

Faced with a global society with a continuously growing population, shrinking land available for cultivation and the need to make the most yield from limited agricultural resources, Augere arises. For example, before 2050, a 60% increase in wheat production will be needed to meet the demand of the growing population.

The increase in demand for food has to be accompanied by an improvement in the production level of farmers. This is where Augere comes in, optimising crop yields and advising on the most suitable crops based on a series of meteorological parameters.

In the graphic below, there is a chronology of the cereal demand between the end of last century, the actual and the projection in the future.

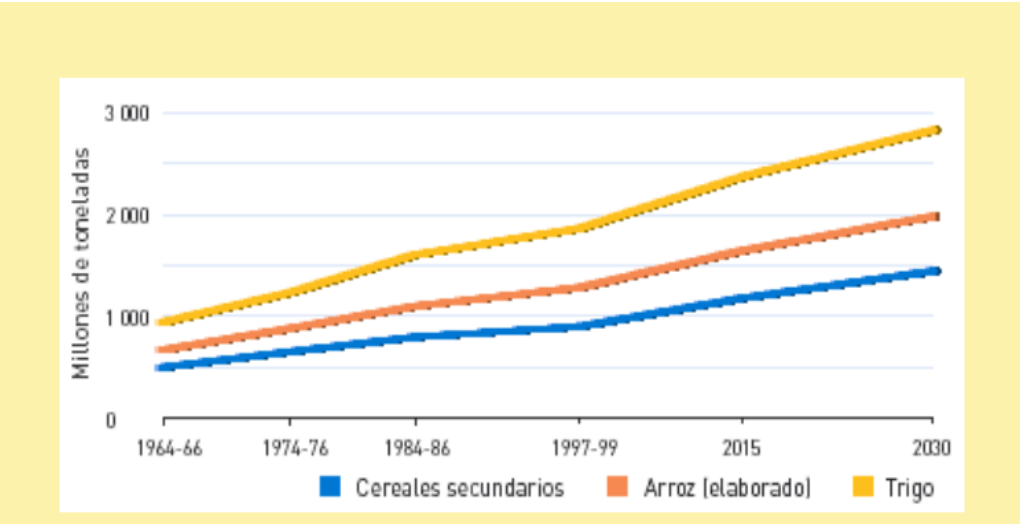

Fig 1. Expected Mundial Demand of cereals from 1965 to 2030.

In order to meet these demand requirements, the crop yield must be improved, as the next graphic reflects. This improvement is mainly related to the limited resources exploitation, and the climatological variables are one of the key responsibilities of these numbers. That is where the organisation focuses all its efforts in solving the problem.

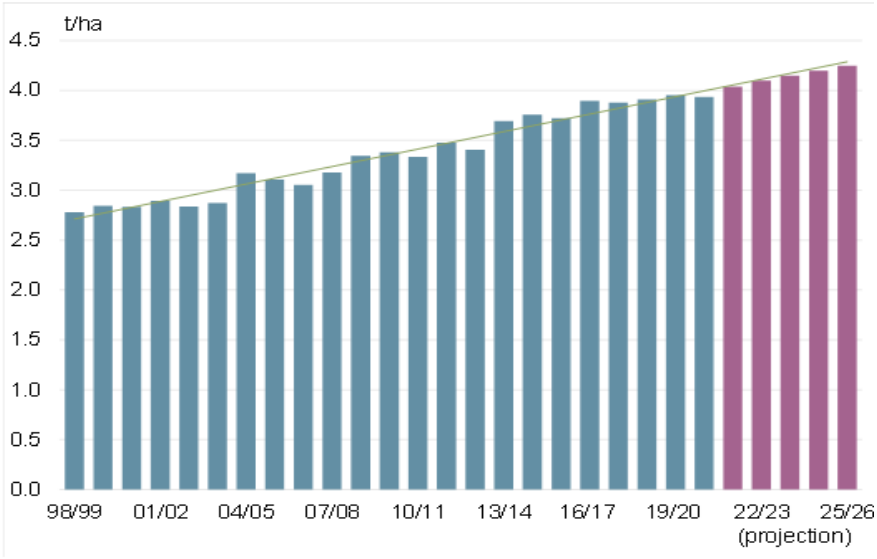

Fig 2. Total grains world average yields (t/ha).

In addition to this information, there is another graph which analises the level of wheat harvested and the potential that different countries around the world could produce, and only Germany has an optimum level, instead of the rest that have ample room for improvement.

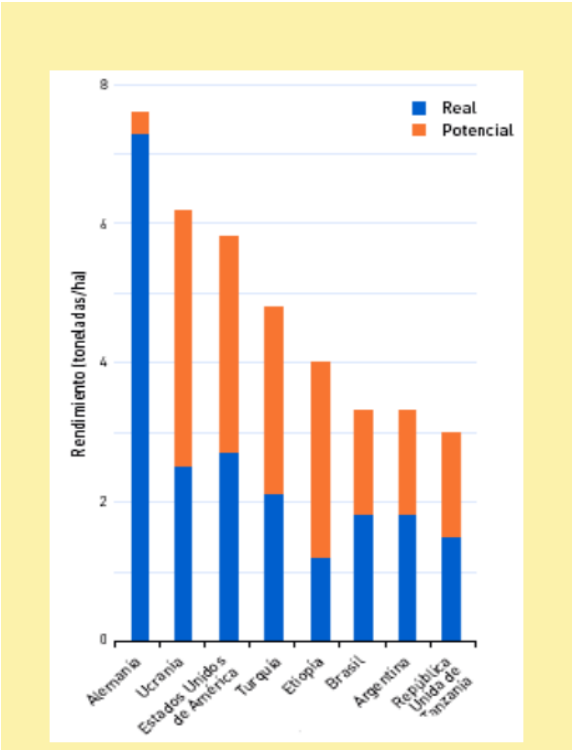

Fig 3. Real yield obtains vs Potential yield in all of the countries.

All those research reports reinforce the purpose of this organization, convincing evidence of their contribution to society in achieving better results for food producers.

1.1.- SPAIN

The starting point for developing the idea of Augere is the Spanish territory, divided in regions, each one with different weather conditions in terms of temperature and rainfall. Also, the idea is to focus on four of the main crops that can be cultivated in the majority of the regions: wheat, corn, potatoes and chickpeas, all of them taking part in a normal consumer's food diet.

Taking data into account, in 2020, cereal (wheat, corn) were the products most planted (6 million hectares) in Spain. In the world, both cereals are also the most cultivated. Corn is the most produced, reaching 1100 tonnes harvested in the 19/20 season.

In the next diagrams there is a representation per region of the production of the crops in which the company is focused on with the application of its algorithm.

- **<u>Wheat production per CCAA in 2019</u>:**

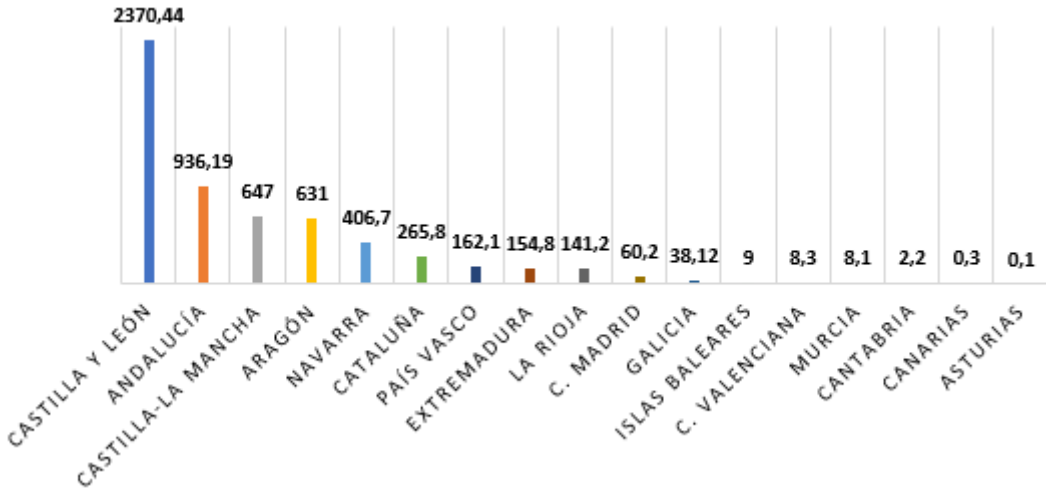

Fig 4. Volume of wheat production in thousands of tons.

It can be seen that Castilla y León is the autonomous community with the highest wheat production by far.

- **Corn produced per CCAA in 2019:**

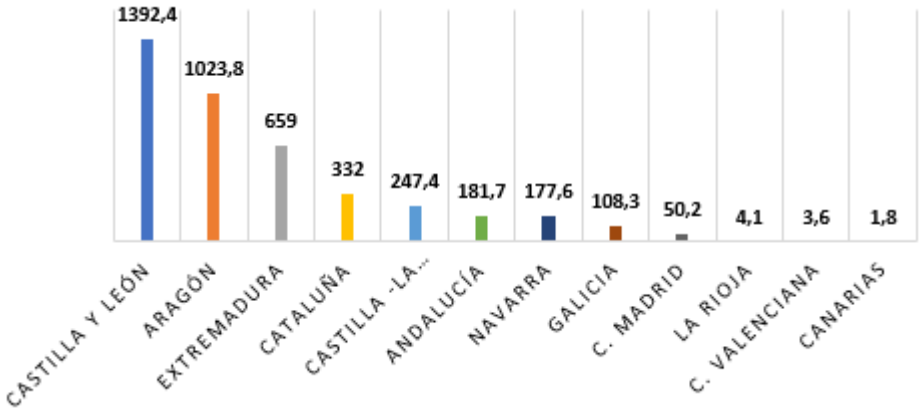

Fig 5. Volume of corn in 2019 production in thousands of tons.

Looking at the figure above, corn production reached 4182.4 thousand tonnes, representing 21% of cereal production and 3.5% of the agricultural total in Spain as a whole.

- **Potato production per CCAA in 2019:**

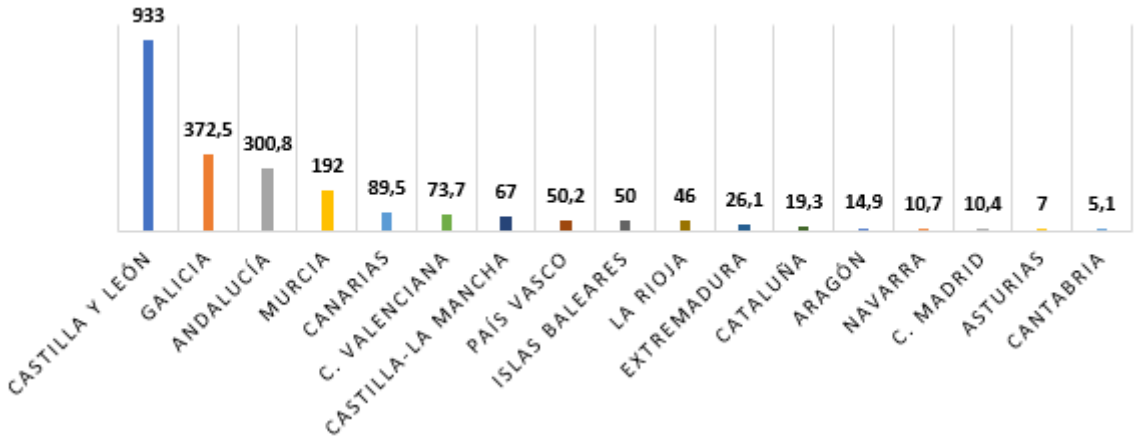

Fig 6. Volume of potato in 2019 production in thousands of tons.

It is true that Castilla y León production predominates over the rest, but there is a basic production in most of the regions, which is a positive thing, so, like other crops, Augere can reach many destinations.

- **Chickpea produced per CCAA in 2019:**

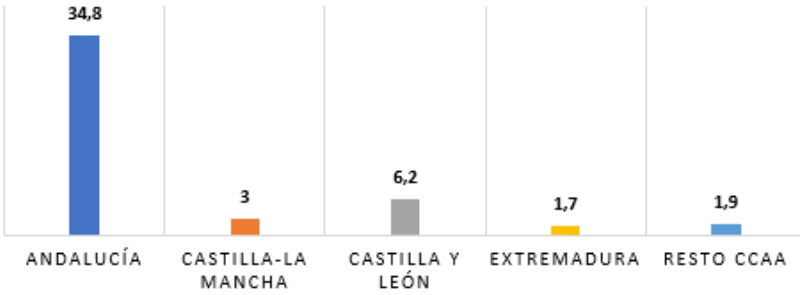

Fig 7. Volume of chickpea in 2019 production in thousands of tons.

Andalusian production is very defined, but there is a tendency in other communities to increase the seeding of this kind of grain legume, which reached in total more than 47 thousands of tonnes produced in the whole country.

Analyzing these statistics and the needs of the society, Augere can help with the business problem contributing with its idea based on Machine Learning technology in order to improve the food production around farmers.

2.- DESCRIPTION OF THE DATASET USED

With the aim to learn more about the case in question, the Machine Learning model was done end-to-end, that is, the dataset used was elaborated by hand by the team members, searching through numerous pages, from different government ministries (agriculture, industry), state meteorological agency (Aemet), national statistical institute (INE) and independent webpages, extracting the interesting data and doing some calculus, like sum and the mean, with the objective to have all the information in an unique dataset.

The elaboration of this dataset was a bit hard and difficult, sometimes to find the data, so the team had to send messages and telephone calls to get more detailed information, but with no success.

In general, the whole process of extracting the dataset was a bit time-consuming, but as a general reflection from the three group members, it was worthwhile, so the learning was more successful and the data was owned by the team, not copied. As a result, the database is not incomplete and does not have missing values.

The different features contained in the database are described below.

2.1.- FEATURES

These are the features contained in the database:

- **Comunidad_Autonoma:** Spain's 8 autonomous communities (Galicia and Pais Vasco from the north area, Cataluña and Comunidad Valenciana from the east area, Madrid and Castilla y Leon from the central area and Extremadura and Andalucia from the south area). These autonomous communities have been selected in order to cover as much of the national territory as possible and to get a general idea of the whole country.

- **Año:** years from 2012 to 2019.

- **Precipitacion_Mes:** there are 12 features of this type (each one referring to the mean precipitation registered in l/m2 in the entire region each month of the year).

- **Precipitacion_Anual:** sum of precipitation registered in l/m2 in the region throughout all the year.

- **Rendimiento_Trigo:** wheat production in kg per hectare of land in each of the regions per year.

- **Rendimiento_Patata:** potato production in kg per hectare of land in each of the regions per year.

- **Rendimiento_Maiz:** corn production in kg per hectare of land in each of the regions per year.
- **Rendimiento_Garbanzo:** chickpea production in kg per hectare of land in each of the regions per year.
- **TempMax:** maximum temperature recorded in the region throughout the year
- **TempMin:** minimum temperature recorded in the region throughout the year
- **MediaTemperatura:** mean of the observed temperature in each region per year.
- **DiasTempMenor0:** number of days with a temperature lower than 0 degrees in each region per year.

We have created a database for 17 autonomous communities in Spain, but we will work with the 8 previously mentioned to facilitate calculations and data visualisation.

2.2.- EXPLORATORY ANALYSIS

The exploratory analysis of the dataset contains the following items:

2.2.1.- Initial Database Analysis

It is interesting to note that all the variables in the database are numerical, except for the Autonomous Community, which is categorical. Also, the total number of elements in the dataset is 2852.

As mentioned above, the data in the database has been entered by hand by the members of the group by analysing different sources of information, so none of the columns contain any missing values.

What is more, precipitation is an environmental factor whose value can range from 0 to an unpredictable maximum due to punctual meteorological phenomenon such as storms, floods, etc. On the other hand, as the information in the database was entered by hand by the members of the group, we tried to make sure that both the values related to rainfall and crop yields were coherent. That said, we did not consider the existence of outliers.

2.2.2.- Visualization of Data

Let's explore the data by visualizing the distribution of values in some columns of the dataset.

To facilitate a series of operations, a data frame will be created for each autonomous community. The following graph shows the annual rainfall by autonomous community for the years 2012 to 2019.

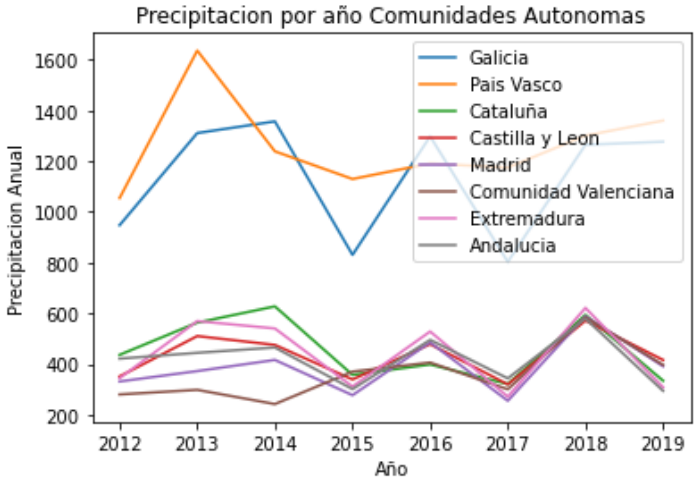

Fig 8. Annual precipitation from 2012 to 2019 in every region of Spain.

The following conclusions can be drawn from these graphs:

- On one hand, it can be seen that Galicia and Pais Vasco are the rainiest regions. Based on this fact, it can be seen that the northern regions have the highest precipitation levels. On the other hand, Comunidad Valenciana is the least rainy region.
- Secondly, the graphs show that the years 2013 and 2018 were the rainiest ones.

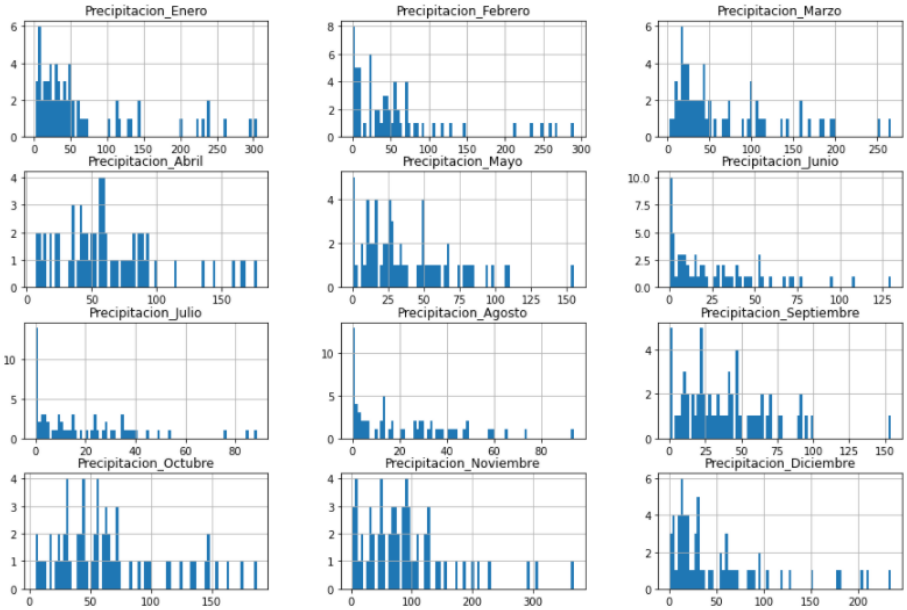

Fig 9. Rainfall distribution in every month for all regions.

As it can be seen in the graph above (Fig.9), January, February, March and November are the wettest months, while July and August are the driest.

Once the analysis of rainfall has been carried out, the next step is to proceed to study the yield of the different crops by region. First, the wheat yield will be studied.

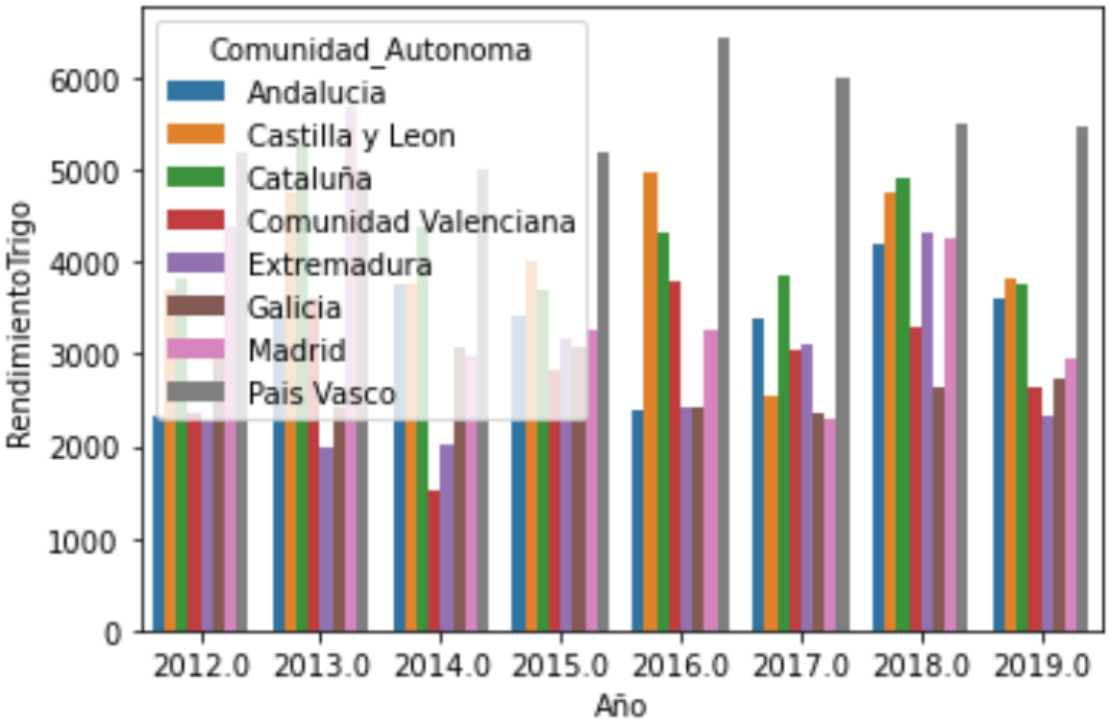

Fig 10. Wheat yield analysis for each region.

The above graph does not show this information very well, so a graph by autonomous community will be made to be able to compare the values in a better way.

Wheat yield analysis:

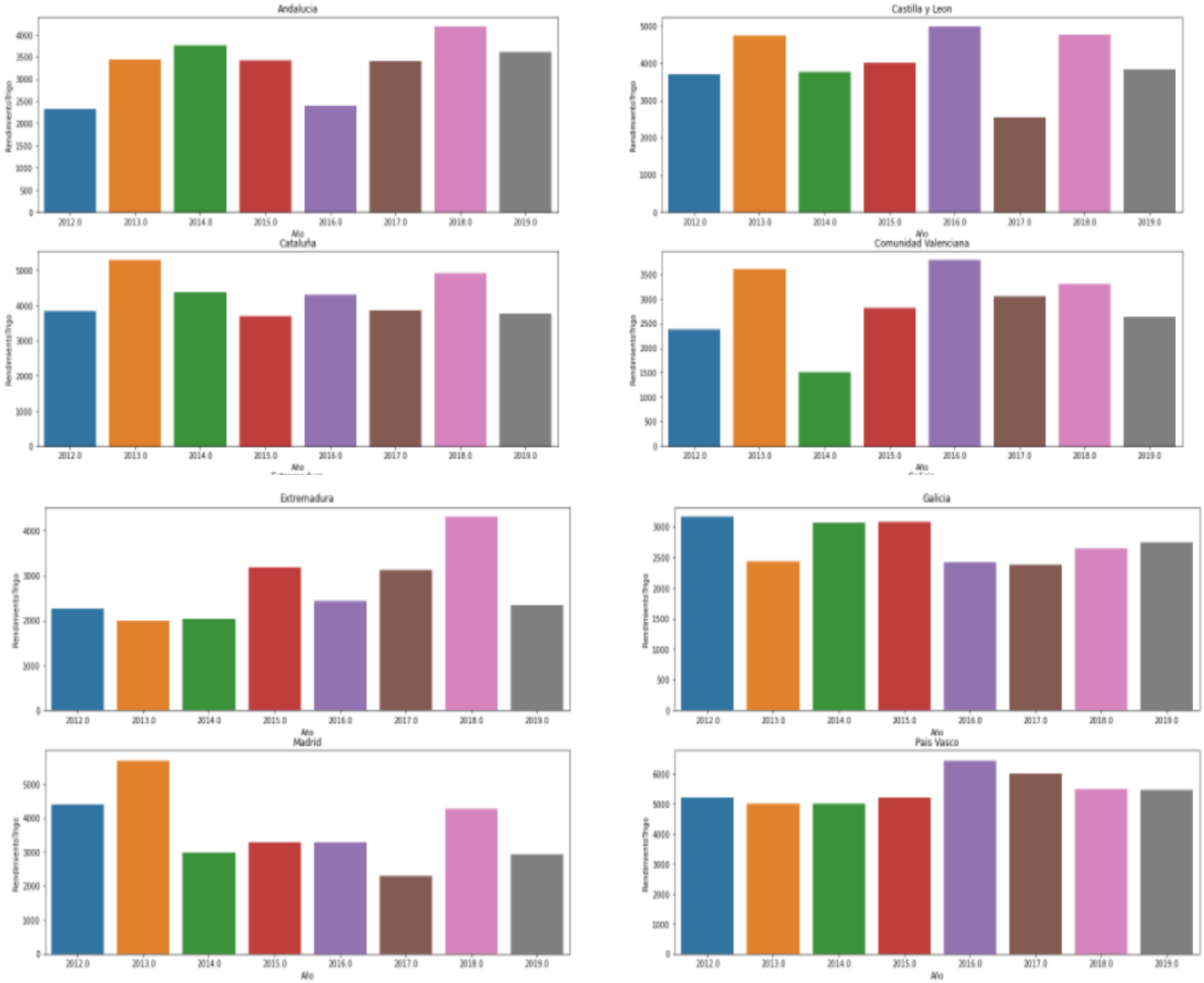

Fig 11. Yield of wheat per year in every region of Spain.

The following conclusions are drawn:

- Pais Vasco and Cataluña are the two regions with the highest wheat yields. Galicia is on the opposite side.
- Extremadura is a unique case. It can be seen that almost every year it has a low wheat yield, but in 2018 it doubles these values, reaching 4000kg/ha.

To verify these facts, the average per region and the following graph will be calculated.

```
Comunidad_Autonoma
Andalucia                3317.375
Castilla y Leon          4047.125
Cataluña                 4251.875
Comunidad Valenciana     2885.250
Extremadura              2710.750
Galicia                  2742.500
Madrid                   3639.125
Pais Vasco               5477.500
Name: RendimientoTrigo, dtype: float64
```

Fig 12. Average per region.

Also, to demonstrate the previous conclusions, the boxplot of each region with the wheat yield is performed:

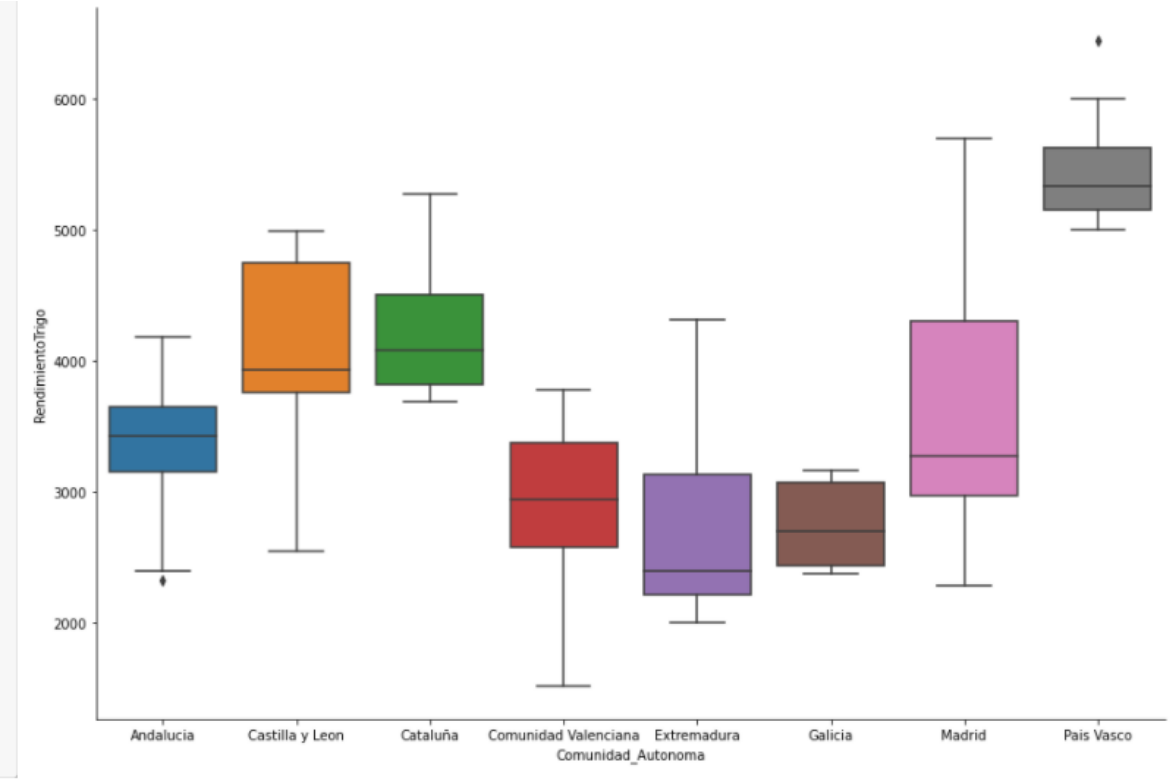

Fig 13. Yield distribution of wheat in every region of Spain.

Looking at this graph the previous conclusions are demonstrated.

Next, the case of the potato will be studied.

Potato yield analysis:

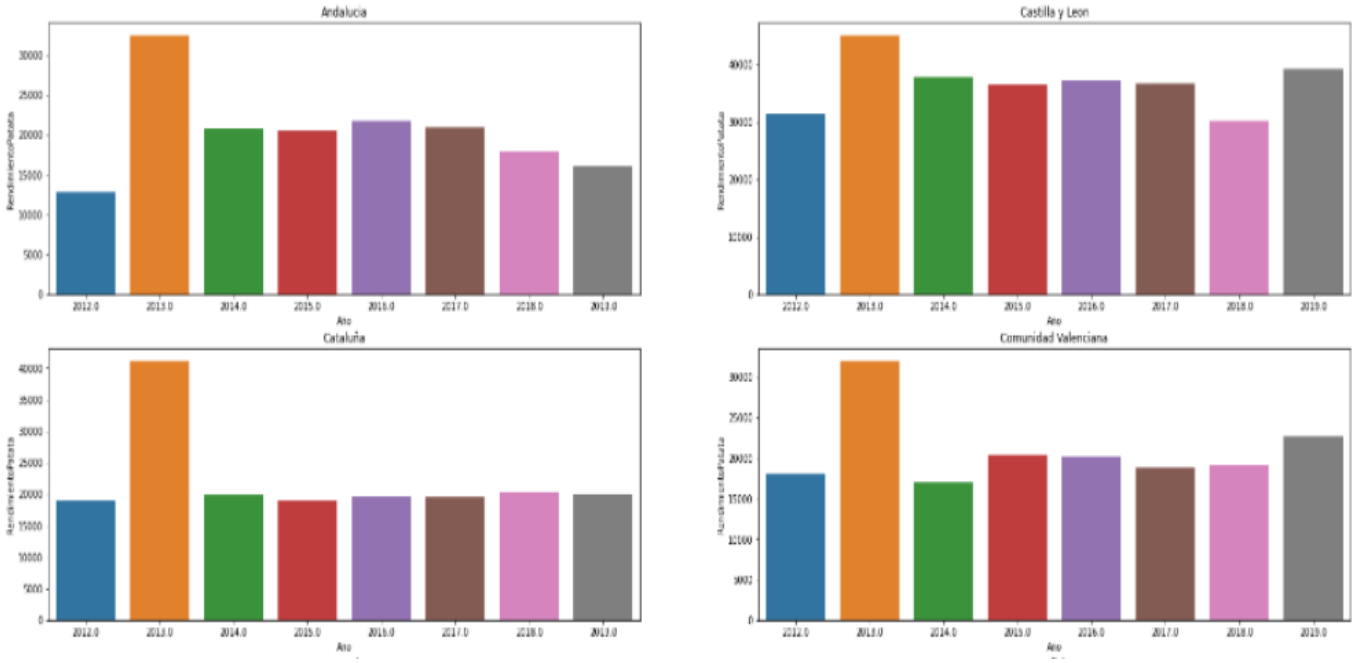

Fig 14. Yield of potato per year in every region of Spain.

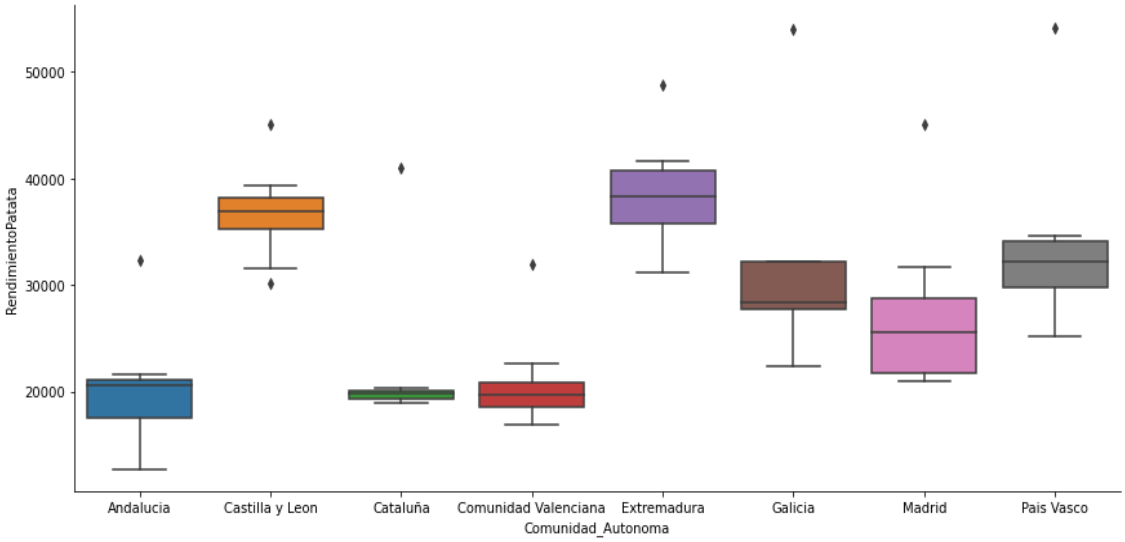

Fig 15. Yield distribution of potato in every region of Spain.

In contrast with the previous crop, potato yield per region is more similar. However, it seems that Extremadura is in the first place.

Doing the next operation, that affirmation is correct. It can be seen that Extremadura has the highest average with an approximate value of 39000kg/ha per year and that the difference between the rest of the averages is not so big.

```
Comunidad_Autonoma
Andalucia                20380.125
Castilla y Leon          36786.250
Cataluña                 22309.250
Comunidad Valenciana     21000.500
Extremadura              38577.875
Galicia                  31588.125
Madrid                   27515.250
Pais Vasco               33908.875
Name: RendimientoPatata, dtype: float64
```

Fig 16. Average per region.

Corn yield analysis:

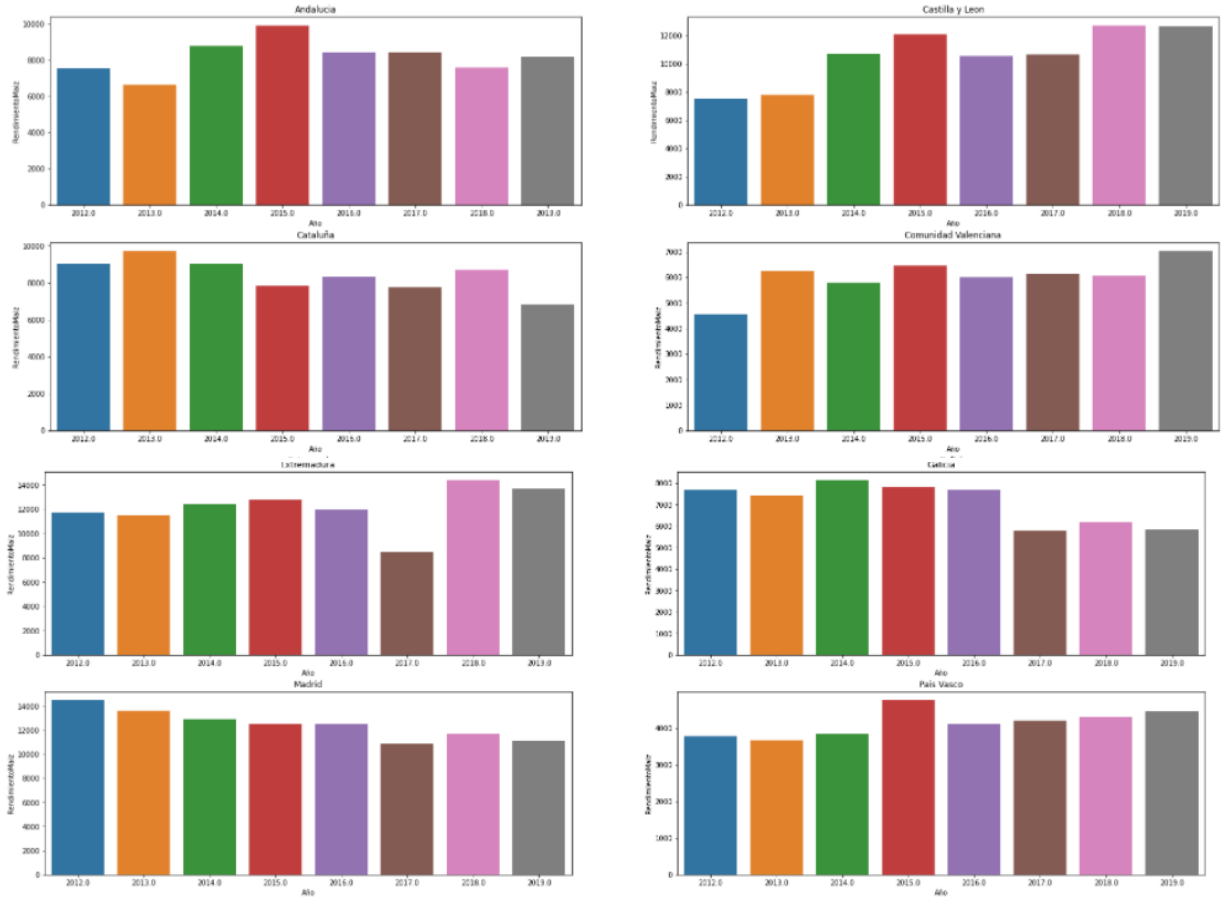

Fig 17. Yield of corn per year in every region of Spain.

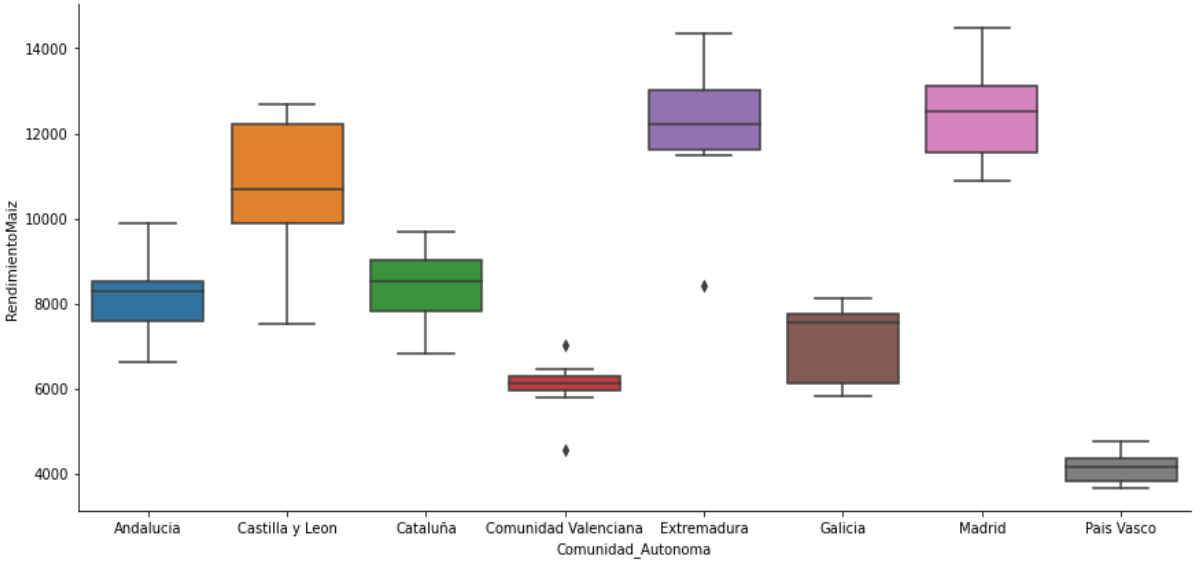

Fig 18. Yield distribution of corn in every region of Spain.

In terms of corn cultivation, Madrid and Extremadura are the two regions that make the best use of their resources for this type of crop.

```
Comunidad_Autonoma
Andalucia                8180.000
Castilla y Leon         10578.125
Cataluña                 8395.000
Comunidad Valenciana     6035.750
Extremadura             12101.000
Galicia                  7078.000
Madrid                  12471.375
Pais Vasco               4141.000
Name: RendimientoMaiz, dtype: float64
```

Fig 19. Average per region.

From these values the following interesting conclusion can be drawn: it seems that the climate in the northern regions of the country is not favourable for corn, as these are the places with the lowest corn yields, whereas the centre regions of the country are the best ones. However, this will be discussed later when comparing crop yields with rainfall levels.

Chickpea yield analysis:

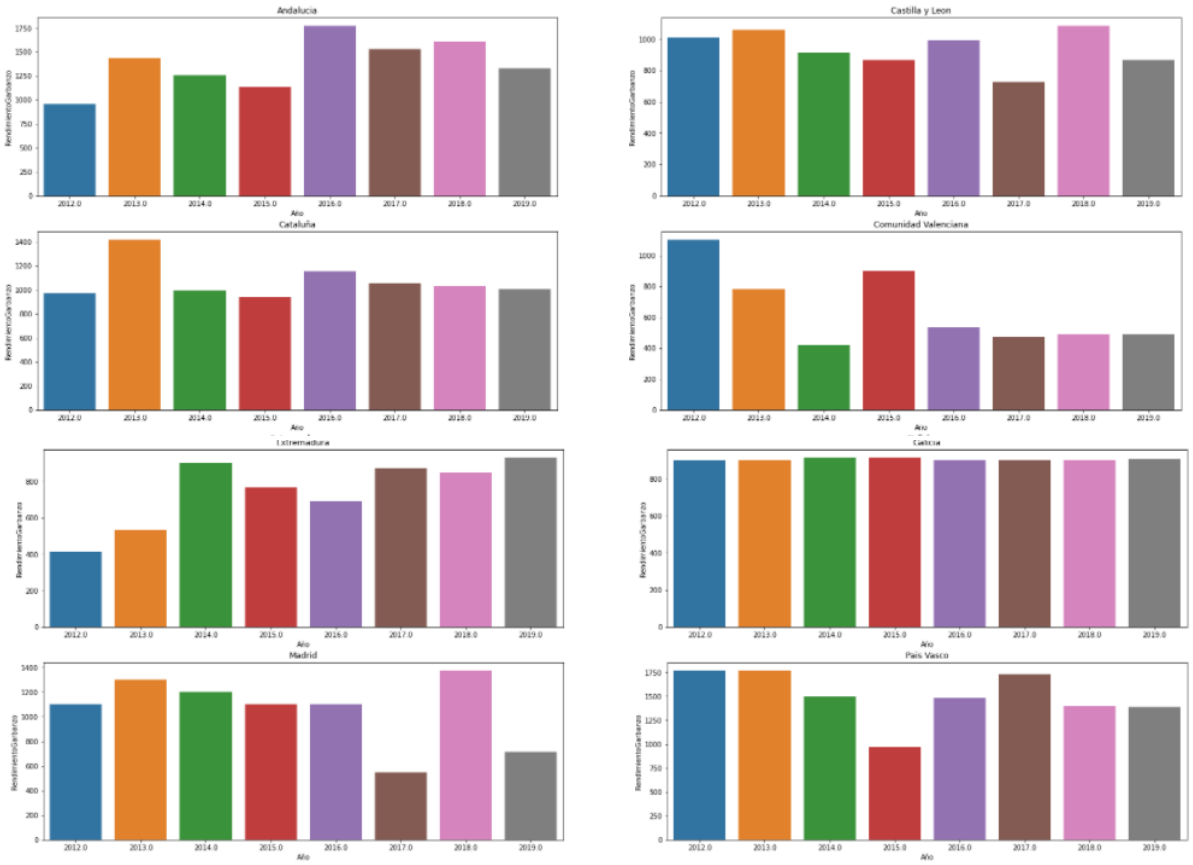

Fig 20. Yield of chickpeas per year in every region of Spain.

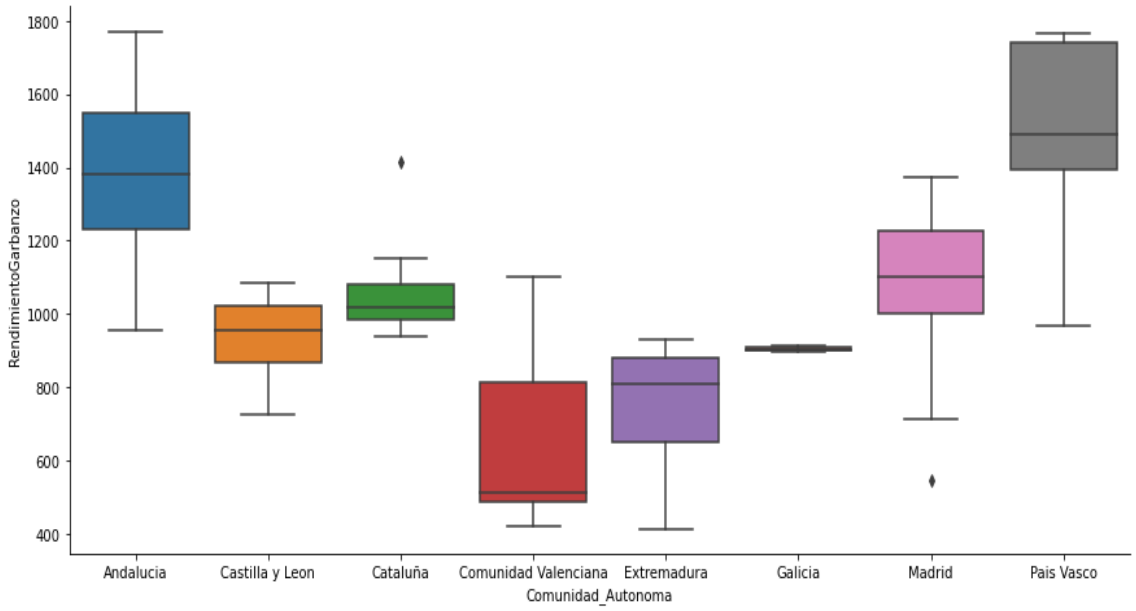

Fig 21. Yield distribution of chickpea in every region of Spain.

Andalucia and Pais Vasco are the two regions with the highest chickpea yields. In terms of Andalucia, as the production graph shows in the description of the business problem, the region with more tonnes harvested, has one of the best yields. So, this is a reinforcement that the weather in that place is favourable for this type of crop.

```
Comunidad_Autonoma
Andalucia                1378.375
Castilla y Leon           940.000
Cataluña                 1070.000
Comunidad Valenciana      649.750
Extremadura               745.125
Galicia                   903.500
Madrid                   1054.375
Pais Vasco               1498.625
Name: RendimientoGarbanzo, dtype: float64
```

Fig 22. Average per region.

Among the four crops studied, it is clearly the least cultivated in the country, although each year is being cultivated more.

2.2.3.- Correlation between features

The following types of correlation are going to be calculated:

- Types of climates in Spain.
- Correlation between the level of precipitation and different crops.

The following graph shows the types of climate in the country.

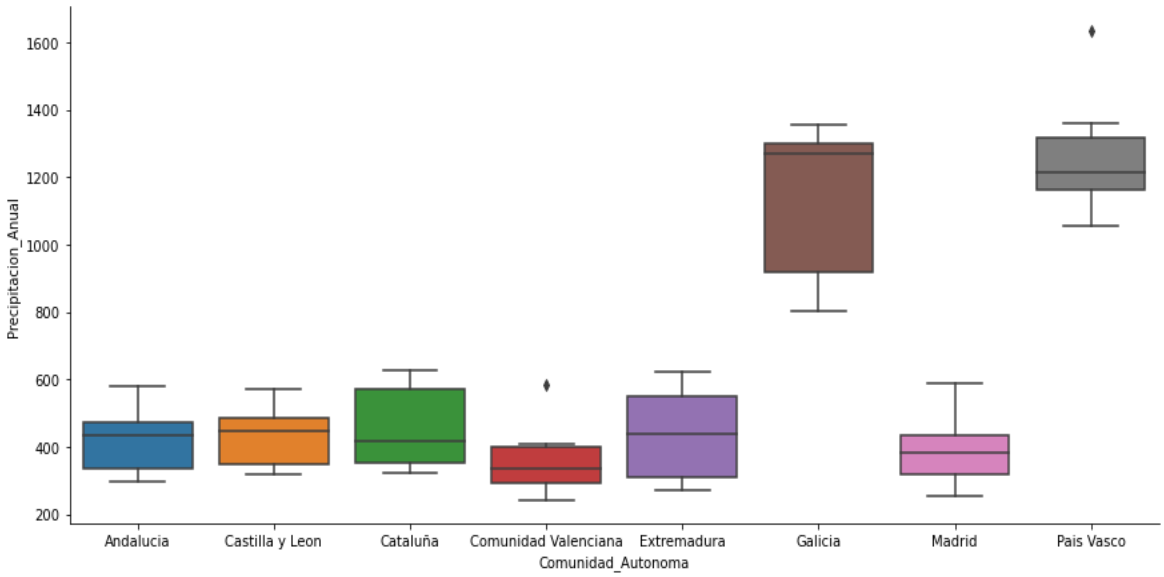

Fig 23. Distribution of annual precipitation in every region of Spain.

The conclusion is that Galicia and Pais Vasco have a rainier climate than the rest of the autonomous communities. For this reason, Spain's climate is divided into these two groups: the north regions are clearly correlated because of its rain levels.

Correlation between the level of precipitation and different crops:

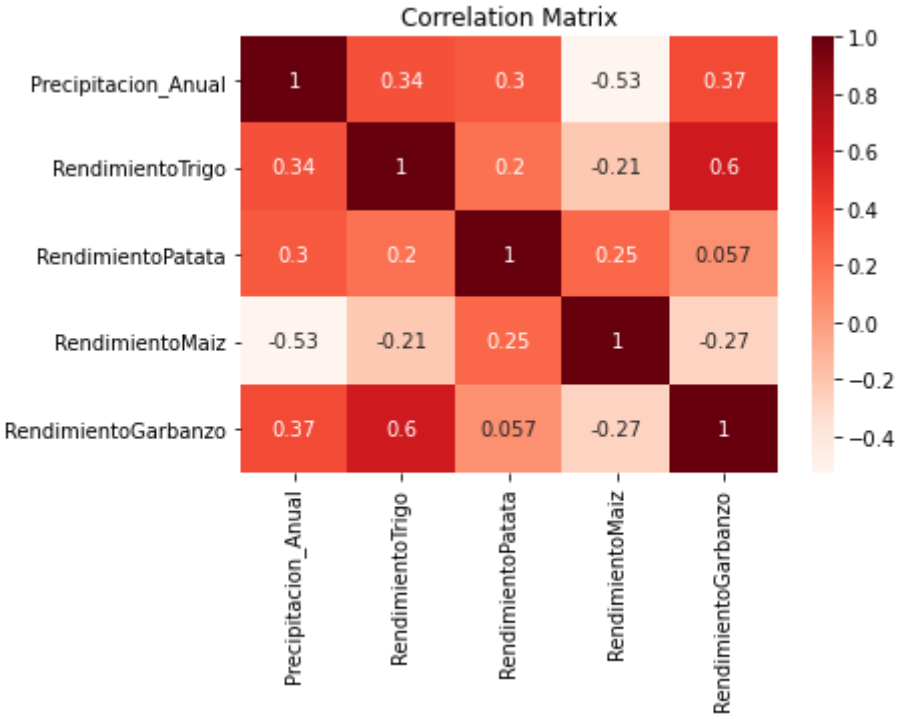

Fig 24. Correlation Matrix between of all the yields.

The following conclusions can be drawn from the upper matrix:

- Although the correlation between annual rainfall and wheat, potato and chickpea yields is small, it is positive. On the other hand, the correlation between annual rainfall and corn yield is negative.
- The correlation between wheat and chickpea yields is the highest.

In order to observe these facts, the following graphs contain the correlation between the annual precipitation and each of the crops:

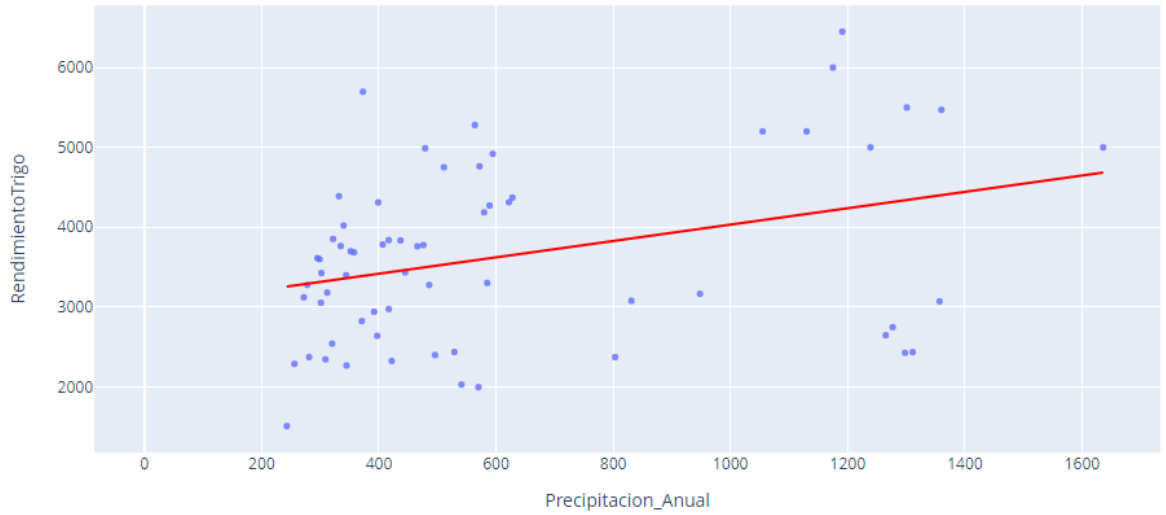

Fig 25. Annual precipitation vs Wheat Yields.

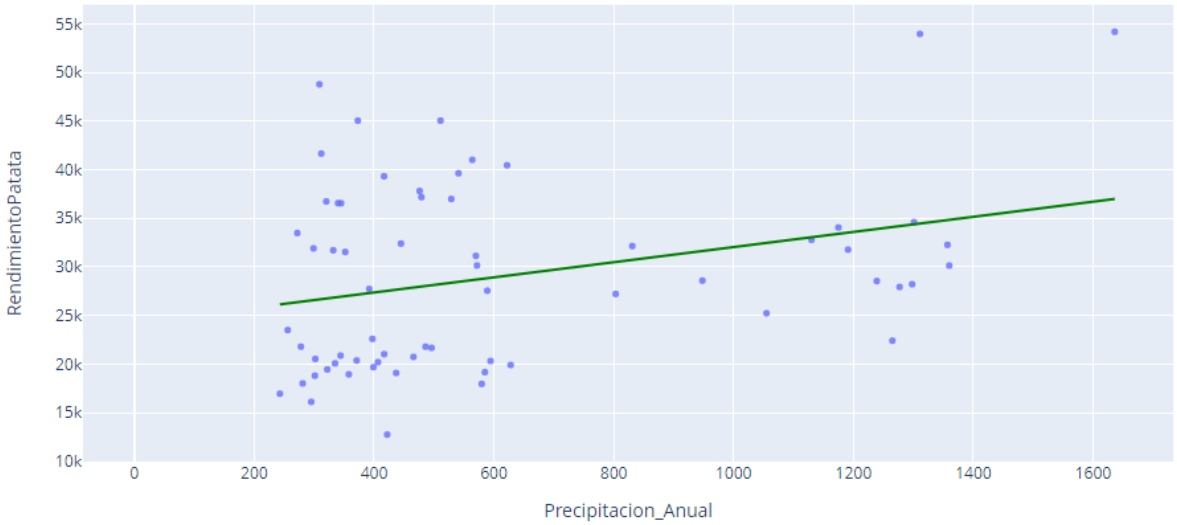

Fig 26. Annual precipitation vs Potato Yields.

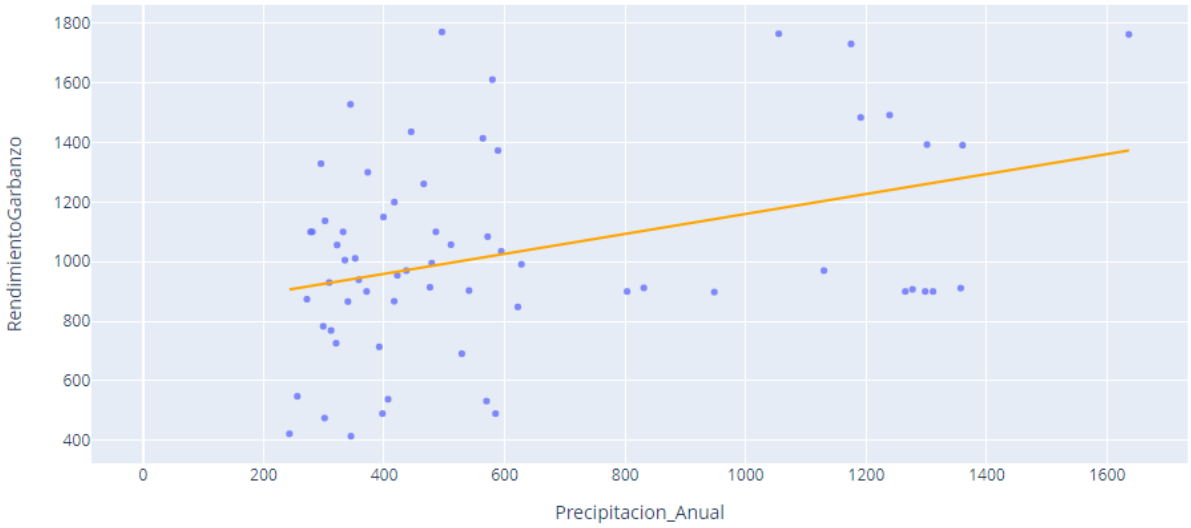

Fig 27. Annual precipitation vs Chickpea Yields.

The correlations in the crops above show a positive line.

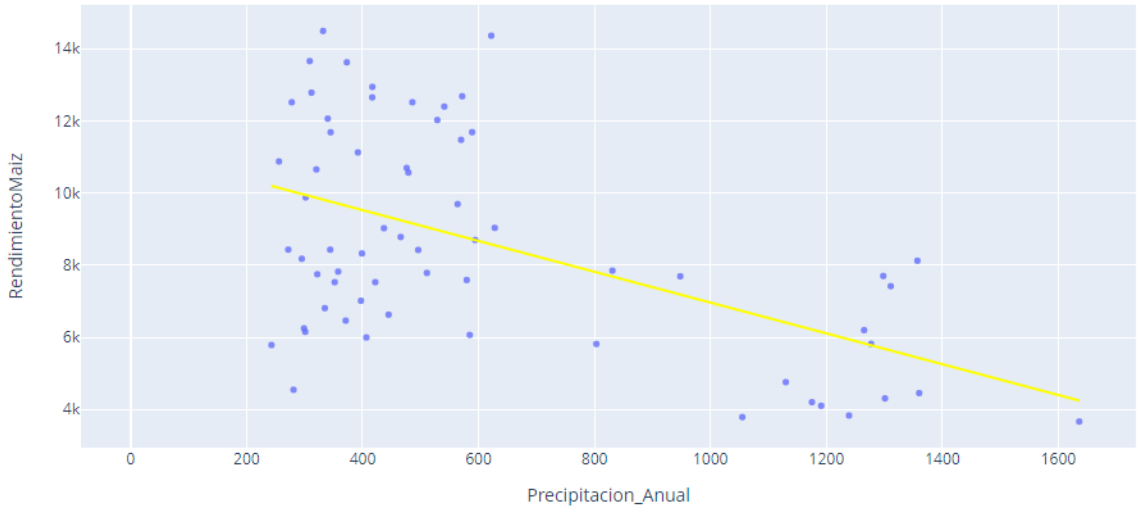

Fig 28. Annual precipitation vs Corn Yields.

In contrast to the other cases, as shown in the correlation matrix, the correlation is negative: when rainfall increases, corn yield decreases.

In the following, two correlation matrices will be created to make the values more visible in the matrix and observe the influence of monthly rainfall on the different crops.

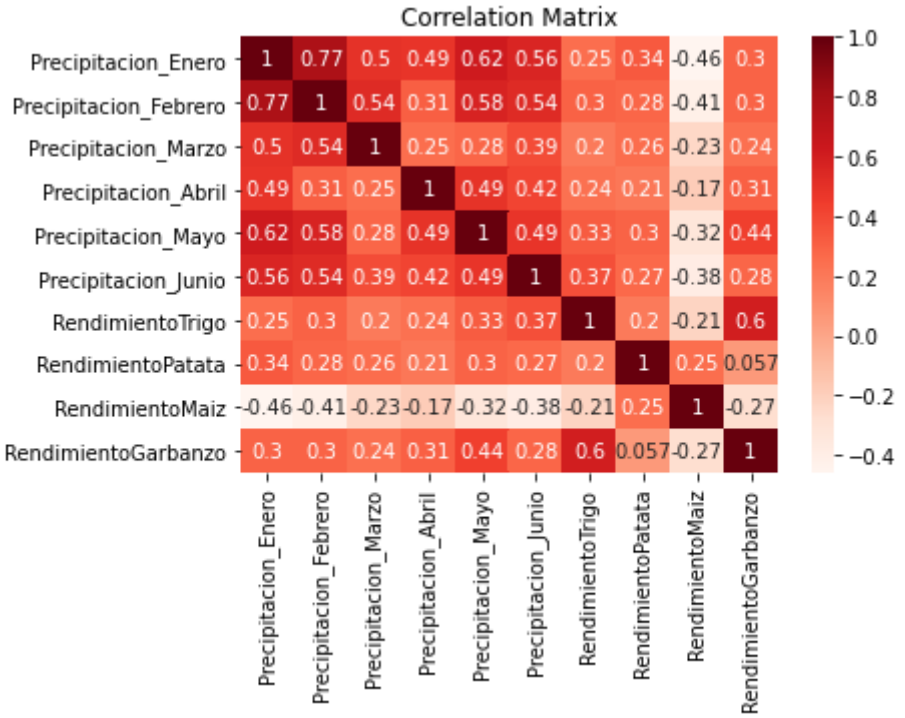

Fig 29. Matrix of correlation between of all the yields and precipitation levels from January to June.

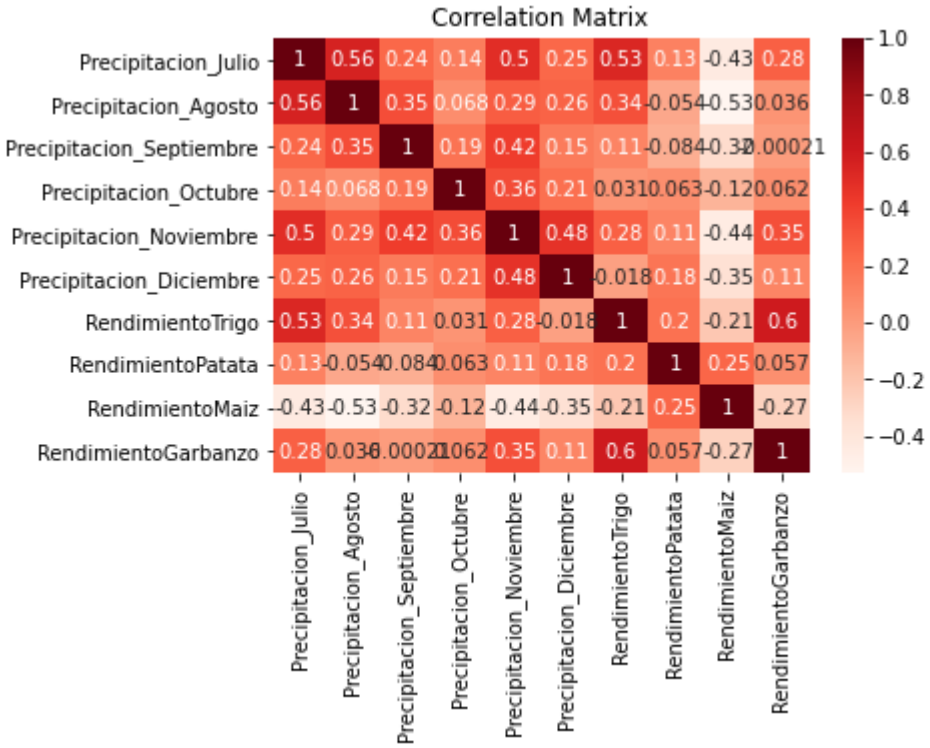

Fig 30. Matrix of correlation between of all the yields and precipitation levels from June to December.

The following results are obtained from the two correlation matrices above:

- July's rainfall level has the greatest influence on wheat productivity.
- For potatoes, even if there are no predominant months, January's rainfall level has the biggest impact.
- It can be seen that in the case of corn yields, a higher level of precipitation in the month of August negatively affects this feature.
- May's rainfall level has the biggest influence on chickpea productivity.

2.2.4.- Conclusions of the exploratory analysis

- This has been the exploratory analysis of the database.
- As a way of making the database more complete and more real, there are going to be added four more variables related to temperature, as defined in the features.

2.3.- EVALUATION STRATEGY

The evaluation strategy is to outline the yield prediction with 4 different models, as there are 4 crops. Maybe it would be necessary to perform more than one predictor, in order to get the best result possible.

2.4 DATA-DRIVEN HYPOTHESIS

The preliminary hypotheses considered in tackling the problem are detailed below:

- Galicia and País Vasco are the autonomous communities with the highest levels of precipitation.
- Wheat yields are less affected by rainfall than other crops.
- Cataluña and Andalucia have very little correlation with respect to the level of precipitation. However, Madrid and Extremadura have a high correlation.

3.- DESCRIPTION OF THE STEPS USED OF THE ML ANALYSIS

3.1- TRAIN AND TEST DIVISION

First of all, the **train and test division** has been done. It is interesting to note that 80% of the data will be used in the training set and the remaining 20% in the test set.

We have created the following variables:

- X_train
- X_test
- y_train_trigo
- y_train_patata
- y_train_maiz
- y_train_garbanzo
- y_test_trigo
- y_test_patata
- y_test_maiz
- y_test_garbanzo

There are four variables related to y_train and other four related to y_test because our model has to be able to predict the yields of the four crops.

3.2- PREPROCESSED

Once we have created those variables, the second step is to **preprocess** the data.

It has to be mentioned that Comunidad_Autonoma is the only qualitative variable. The rest of the variables are numerical ones, so these variables have been divided into two groups.

Once the variables have been divided into two groups, the following operations have been executed:

- Binarization of qualitative variable (Comunidad_Autonoma).
- Standardization and scaling of numerical variables.

With the standardization and scaling of numerical variables, the predictors are equal in some way, so there will be no predominant feature or most influential feature for the response variable.

It can be seen also that the qualitative variable has been divided into 8 columns, thanks to the binarization (range 0-1).

3.3- LINEAR REGRESSION

Then, we have started with **linear regression.** The steps that have been carried out are the following ones:

- Define the model.
- Do the fit.
- Make the prediction.
- Estimate the error.

With this result obtained from the linear regression of wheat, it is feasible to perform the calculations with another type of model, as the error is 43.4%. It has been decided to study the error with the Random Forest to see if this value improves.

3.4- RANDOM FOREST

With the **Random Forest**, the steps are the same but the error obtained improves significantly (0,2). However, the yield prediction error of the other three crops have been calculated with these two methods, to see if this decrease in error was a one-off event or if the Random Forest is more efficient.

Having said that, the next figure shows the error of the predictions with the two methods previously mentioned.

ERROR	TRIGO	PATATA	MAÍZ	GARBANZO
LINEAR REGRESSION	0,43	0,53	0,27	0,37
RANDOM FOREST	0,19	0,267	0,14	0,34

The model that best fits the yield data of the 4 crops is Random Forest, with a lower percentage error than Linear Regression.

3.5- CREATION AND VALIDATION OF THE MODEL

The next step consists on creating the model. Two different techniques (Ridge and Random Forest Regressor) have been considered for this action. It has to be mentioned that this action has been done for all the four crops.

3.5.1-RIDGE

Firstly, the wheat yield is explained. The next graph shows that the train error is always smaller than the validation error. This means that there is an overfitting problem.

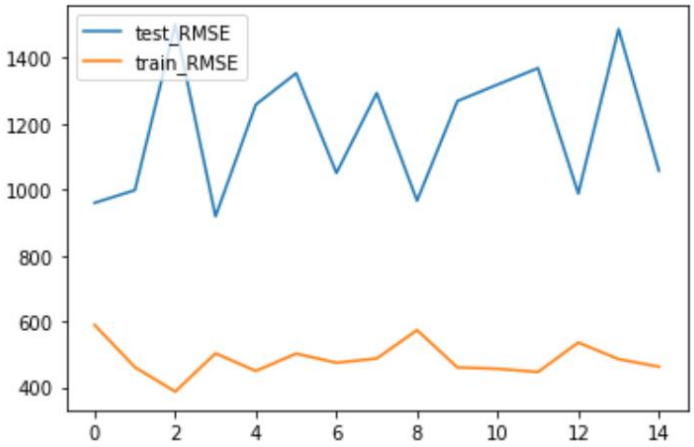

Fig 31. Train error vs validation error (wheat-Ridge).

The average root_mean_squared_error estimated by cross-validation for the ridge model is 929. This value will be checked later when the error of the model is calculated with the test set.

The same process is then carried out for the other three crops.

In the case of potato yield, we can observe the following figure:

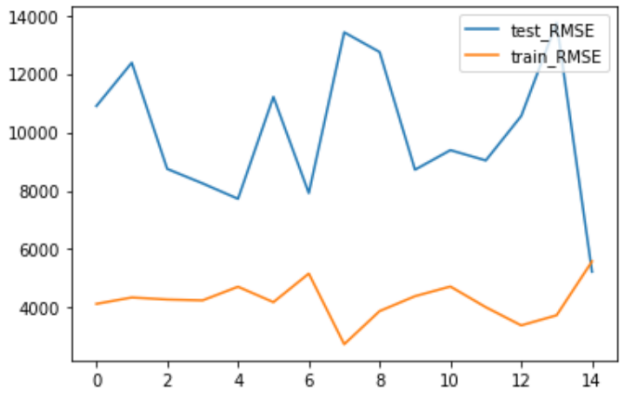

Fig 32. Train error vs validation error (potato-Ridge).

Unlike in the case of wheat, it can be observed that there is considerably less difference between the training error and the validation error, even crossing at some points. This means that the overfitting is less pronounced. On the other hand, the average root_mean_squared_error is 8811,62.

For the third type of crop (corn):

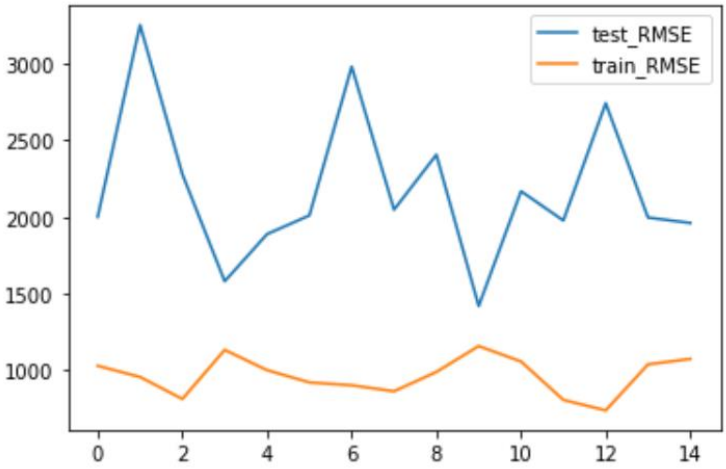

Fig 33. Train error vs validation error (corn-Ridge).

This graph has a similar behaviour as in the case of wheat, i.e. a clear problem of overfitting is observed, and the average root_mean_squared_error obtained is 1779,68.

Finally, for the case of chickpea:

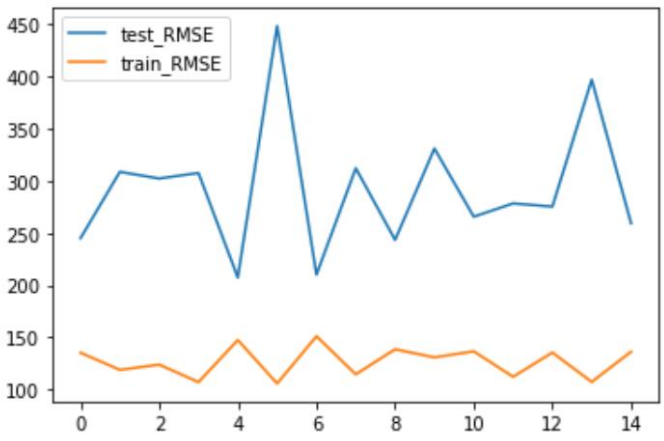

Fig 34. Train error vs validation error (chickpea-Ridge).

The same conclusions can be drawn as for wheat and corn. What is more, the average root_mean_squared_error in this case has a value of 242,32.

3.5.2- RANDOM FOREST REGRESSOR

Next, to compare the results, the same process will be carried out but with the Random Forest Regressor, as it has been mentioned before. To avoid putting up the graphs (in the Notebook they can be seen), it is said that the problem of overfitting is still present.

In the case of wheat yield, the average root_mean_squared_error estimated for the Random Forest Regressor is lower than the value of the Ridge regularization technique (872<929).

As well as in the case of wheat, the average root_mean_squared_error estimated is lower with the Random Forest Regressor (8137,55<8811,62) in the case of potato yield.

The behaviour in the case of corn yield is a different because the value is a little bit bigger in the case of Random Forest Regressor (1834,85>1779,68).

Finally, taking into account the chickpea yield, the value decreases a little bit (235<242,32).

In conclusion, it can be said that the average root_mean_squared_error estimated is lower with the Random Forest Regressor than with the Ridge. However, an overfitting problem is still observed.

3.6- PREDICTION AND TEST ERROR

Once the process of training has been completed, the **prediction** of the different crops has been carried out and the **test error** has been calculated.

It is interesting to note that for each crop, on the one hand, a table with different predictions will be shown, and, on the other hand, the test error will be visualized.

- **Wheat**

	RendimientoTrigo	prediccion
22	4919.0	4214.199110
32	2267.0	2833.785253
33	1996.0	3692.686463
8	3698.0	4367.904161
6	4187.0	3960.352210

Fig 35. Wheat prediction vs real value.

In the validation section (Random Forest Regressor), it was estimated, by repeated cross-validation, that the rmse of the model was 872. However, in the test the value decreases to 677,02.

So, this is a good notice. Let's see if it happens the same with the other crops.

- **Potato**

	RendimientoPatata	prediccion
22	20316.0	29406.27
32	36539.0	31913.90
33	31130.0	36869.87
8	31532.0	38015.57
6	17962.0	29684.69

Fig 36. Potato prediction vs real value.

In the validation section (Random Forest Regressor), it was estimated, by repeated cross-validation, that the rmse of the model was 8137. The test error decreases to 7862,65.

- **Corn**

	RendimientoMaiz	prediccion
22	8699.0	7770.49
32	11685.0	11405.52
33	11475.0	10462.10
8	7530.0	11440.79
6	7589.0	8740.05

Fig 37. Corn prediction vs real value.

In the validation section (Random Forest Regressor), it was estimated, by repeated cross-validation, that the rmse of the model was 1834,85. The test error decreases to 1287,99.

- **Chickpea**

	RendimientoGarbanzo	prediccion
22	1035.0	1010.86
32	414.0	910.30
33	532.0	1033.78
8	1011.0	959.76
6	1611.0	1159.19

Fig 38. Chickpea prediction vs real value.

In the validation section (Random Forest Regressor), it was estimated, by repeated cross-validation, that the rmse of the model was 235. The test error increases to 347,83.

3.7- CONCLUSION

As a final conclusion, it can be said that the model is quite effective in predicting new variables.

Also, summarising the point three, it is interesting to note that for the data available, the Random Forest is a better model than the Linear Regression and that the Random Forest Regressor is more efficient than the Ridge.

4.- CONCLUSIONS

This project has been a very enriching experience as it has been possible to execute a Machine Learning model from a database that has been created by the three members of the group, and all the fundamental aspects to be included have been analysed: description of the dataset, exploratory analysis, visualisation of the features, correlation between crops, training and testing analysis for the ML model and error testing.

Also, it has to be mentioned that it has been a challenging project because of the lack of data in some regions and mainly, because we have spent a lot of time finding and filling our dataset.

On the other hand, the biggest limitation of this model is that weather variables are very unpredictable.

Based on the notions learned, the group has understood that this idea has a lot of projection for possible future steps:

- **Improvement of the predictive model:** including more types of crops (sunflower, canola, sugar beet, olive, grape, etc.), and more environmental variables (humidity rates, soil composition, etc.). In this way, it is possible to cover a larger number of service users and to have more accuracy in the model.

- As a result of the previous point, the idea is to **expand the target market**: going international and not just focusing on the domestic market.

- In order to improve the algorithm model, the **introduction of physical tools** is an excellent complement: sensors, drones and AGVs for data collection, as a way of covering the concept of the future “Precision Agriculture”.

- **Sell the results obtained from applying the machine learning model, on which crops are the most suitable to seed and fertiliser companies**. In this way, they can better know which is the most suitable raw material for each region based on a prediction of soil composition and weather variables. The consequences are positive for all parties: reduction of production costs for companies and the farmers can buy the product as a more exact amount, reducing investment.

- **Development of a mobile application**: creation of a user interface for the clients, easy to use and able to allow farmer-specific planning and monitoring.