Model-agnostic: Partial Dependency Plot (PDP)

Sebastián Gómez, Martín Juanes, Jose Valero

Attention

This practice has been created using git as a version control system. To see the commands used, see appendix (4)

1. Introduction

In some areas (those with a higher risk), it is crucial not only to obtain models with high accuracy, but also to understand how the independent variables influence the dependent variable. One of the most useful tools for this purpose is the Partial Dependence Plot (PDP). This report explores the application of one- and two-dimensional PDPs in two different scenarios: bicycle rental forecasting and housing price forecasting.

2. Discussion

2.1. Bicycle rental forecasting

In the first scenario, the problem of predicting the number of bicycles rented from meteorological variables (temperature, humidity...) and temporal variables (season, days since 2011...) is posed.

Before building the model, we performed the same preprocessing as in previous reports (denormalization of variables, creation of one-hot encoding variables with the stations...). Once this was done, we proceeded with the creation of the model, a ***Random Forest***, in which the independent variable is *cnt* (number of bicycles rented) and the explanatory variables are:

- *workingday*. If day is neither weekend nor holiday is 1, otherwise is 0.
- *holiday*. Weather day is holiday or not.
- One-hot encoding variables for stations: *season_spring*, *season_summer*, *season_fall*.
- *MISTY* that is 1 when weathersit is 2. In other cases it will be 0.
- *RAIN* that will be 1 when weathersit is 3 or 4. It will be 0 in other case.
- *temp*. Temperature in Celsius.
- *hum*. Humidity.
- *windspeed*. Wind speed.
- *days_since_2011*. Number of days from 1-1-2011.

Once the model is created, using the PDPs, we can analyze the influence of days elapsed since 2011, temperature, humidity and wind speed on the predicted bicycle counts. In addition, we can also analyze the joint influence of two variables (e.g. humidity and temperature) on the number of bikes rented.

2.1.1. One-dimensional PDPs

First, it is interesting to analyze whether the number of rented bicycles has been increasing since 2011. For this we can use the PDP in figure 1. In this we can see how in the first 100 days, there is an increase in the number of bikes rented. Between approximately 100 and 300 days, the number of rented bicycles remains relatively stable with a slight decrease at the end. Around day 400, there is a very noticeable increase in the number of bikes rented, after which there is a period of slight growth between days 400 and 650, followed finally by a decrease in the number of bikes rented after day 650.

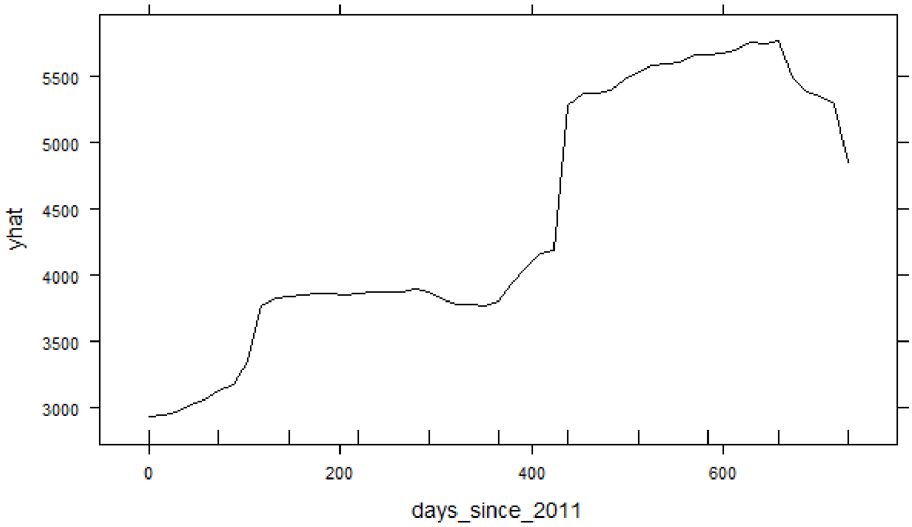

Figure 1: PDP: *cnt* vs *days_since_2011*

Given that the values of *days_since_2011* are uniformly distributed, the above PDP is quite reliable and representative. There are no sections of the graph where predictions are influenced by sparse data, which ensures that the variations observed in the graph reflect actual trends.

On the other hand, it is interesting to know the influence of weather variables on the number of bicycles rented (since, for example, different actions can be applied depending on the weather forecast). In this report we will study the influence of temperature, humidity and wind speed on this variable:

- **Temperature**. In figure 2, the PDP relating the variable *temp* to *cnt* can be observed.

  The following conclusions can be drawn from this PDP:

  - At very low temperatures (below 0 degrees to about 5 degrees), the number of rented bicycles is relatively low and constant. However, since there are few observations in this range, the model prediction at these temperatures is less reliable.
  - As the temperature rises from 5 to 20 degrees, there is a noticeable increase in the number of bicycles rented. This indicates that warmer weather conditions are more favorable for bicycle use. This range is well represented in the data, so predictions in this range are reliable.
  - The number of rented bicycles peaks when the temperature is between 20 and 25 degrees. This temperature range appears to be the most comfortable and attractive for bicycle users. The density of data in this range is also high, which reinforces the confidence in this prediction.
  - When the temperature exceeds 25 degrees, the number of rented bicycles starts to decrease. Although there are fewer observations in the very high temperature range

(over 30 degrees), it is still possible to observe a downward trend. However, prediction in this range should be taken with some caution due to the smaller amount of data.

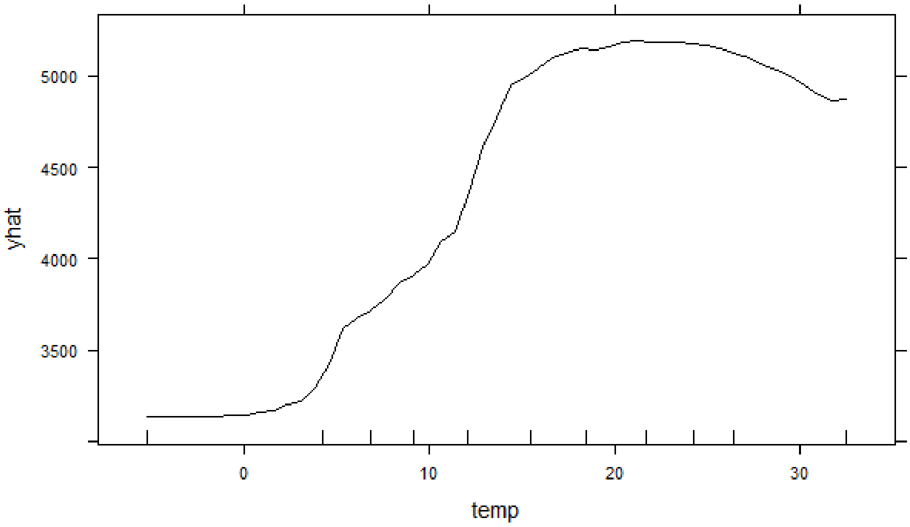

Figure 2: PDP: *cnt* vs *temp*

- **Humidity**. In figure 3, the PDP relating the variable *temp* to *cnt* can be observed.

  The following conclusions can be drawn from this PDP:

  - When humidity is between 0 % and 50 % the number of bikes rented is high and remains constant over this range. These conclusions are unreliable for the range 0 % - 40 % (since there are few observations in this range), but are reliable between 40 % and 50 %.
  - From 50 % humidity, the number of bikes rented decreases progressively up to 100 % humidity. The confidence of these predictions is high when the humidity is between 50-80 % humidity because the data density in this range is high. However, from 80 % humidity and up to 100 % humidity, there are few observations, so the reliability is low.

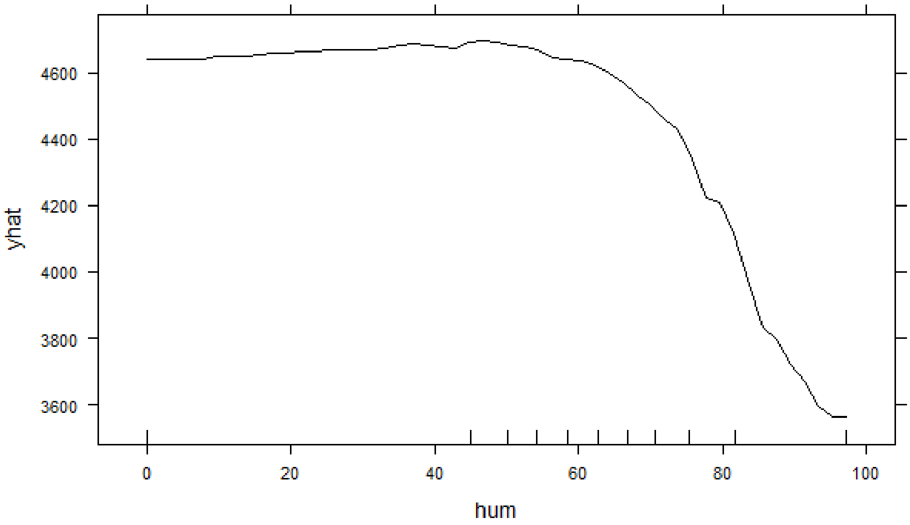

Figure 3: PDP: *cnt* vs *hum*

- **Wind speed**. In figure 4, the PDP relating the variable *temp* to *cnt* can be observed.

  The following conclusions can be drawn from this PDP:

  - In the 0 to 5km/h range, there is a high amount of rented bicycles. However, there is little data representation in this range, which makes the predictions less reliable.
  - Between just over 5 to just over 15km/h, a pronounced decrease in the number of bikes rented is observed. These conclusions are reliable as the data density in this range is high. From just over 15 to 25, the downward trend continues but the predictions are less reliable due to the lower amount of data.
  - Above 25km/h, the number of rented bicycles stabilizes, indicating few rented bicycles. However, there is little data in this range, making predictions unreliable.

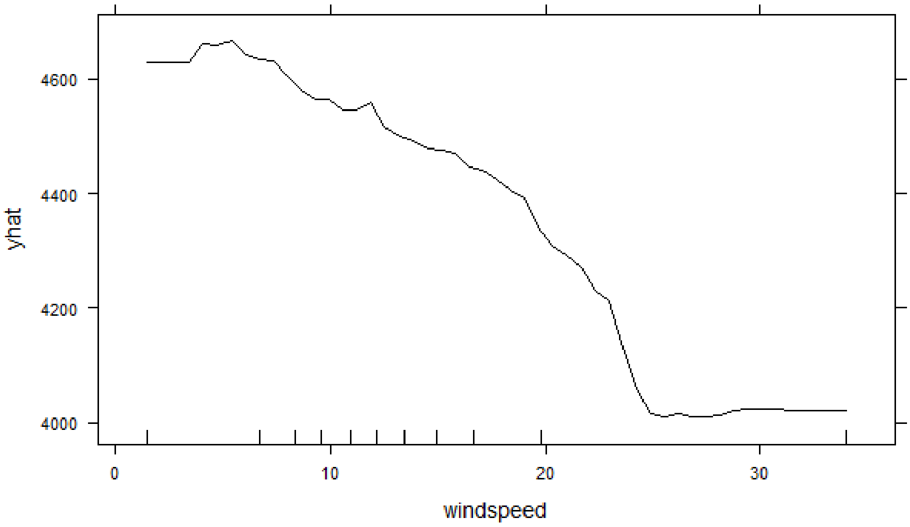

Figure 4: PDP: *cnt* vs *windspeed*

2.1.2. Two-dimensional PDP

In addition to these one-dimensional analyses (in the number of dependent variables), it is also possible to perform two-dimensional analyses, which allow us to know the influence of the combination of two variables on the independent variable.

Figure 5 shows a two-dimensional PDP, from which the influence of humidity and temperature on the prediction of the number of rented bicycles can be studied. In order to correctly interpret this graph it will be necessary to consult the figures 2 and 3, to know the distribution of the dependent variables.

Having made this note, some conclusions that can be drawn from Figure 5 are as follows:

- When the temperature is between a little more than 10 degrees and a little more than 30 degrees, the predicted number of bicycles is high, and the difference of this according to the humidity is not very high (between a humidity of 0 % to 80 %, the prediction is a little higher than from 80 % onwards). The predictions of some areas of these two ranges (humidity between 45 % and 80 %, and temperature between 10 and 25) are more reliable than others (humidity between 0 % and 40 %, and temperature from 25) due to the density of the data in each of these areas.

- On the other hand, when the temperature is low (between slightly less than 0 degrees and 10 degrees), the number of rented bikes predicted is lower than in the previous case. What does remain the same is that when the humidity is between 0 % and 80 % the prediction is higher than above 80 %.

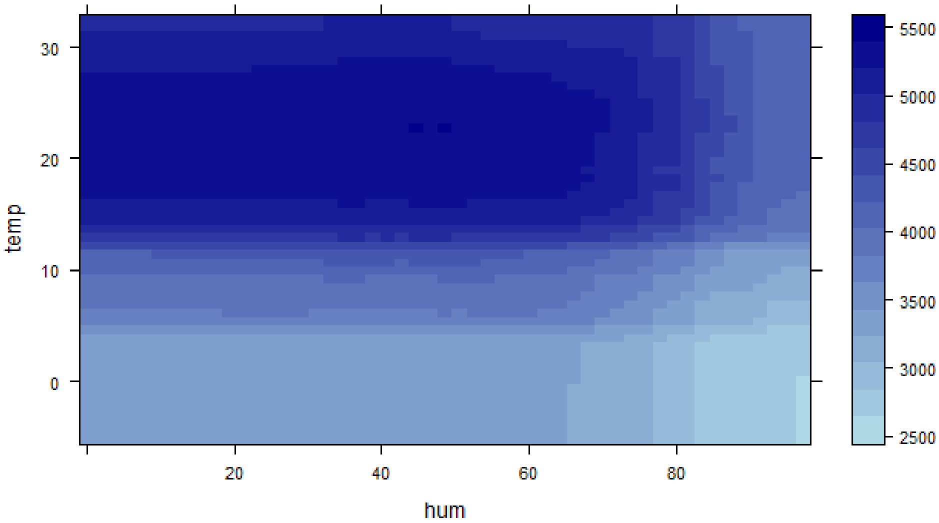

Figure 5: PDP: *cnt* vs *temp* and *hum*

2.2. Home price forecasting

In the second scenario, the problem of predicting the value of a house from the following variables is posed:

- *bedrooms.* Number of bedrooms.
- *bathrooms.* Number of bathrooms, where .5 accounts for a room with a toilet but no shower.
- *sqft_living.* Square footage of the apartment interior living space.
- *sqft_lot.* Square footage of the land space.
- *floors.* Number of floors.
- *yr_built.* The year the house was initially built.

In the same way as in the previous case, we are going to use a ***Random Forest*** as a predictive model. From this, we can generate (among others), the following PDPs, which, when interpreted, will provide us with information on how some of the previous variables affect the price of housing:

- **Number of bedrooms**

  First, we will analyze how the number of bedrooms affects the prediction of the house price. In figure 6, the PDP relating the variables *price* and *bedrooms* can be seen.

  In this graph, the really important part is the initial part. Since in this we can see how the predicted price of the house is lower when it has one bedroom and increases as the number of bedrooms increases up to 2-3 bedrooms. From here, up to 5-6 bedrooms, the predicted price of the house decreases considerably, something that would not be expected.

  From 5-6 bedrooms upwards, the price remains constant but this prediction is not at all reliable as there are practically no homes with that many rooms.

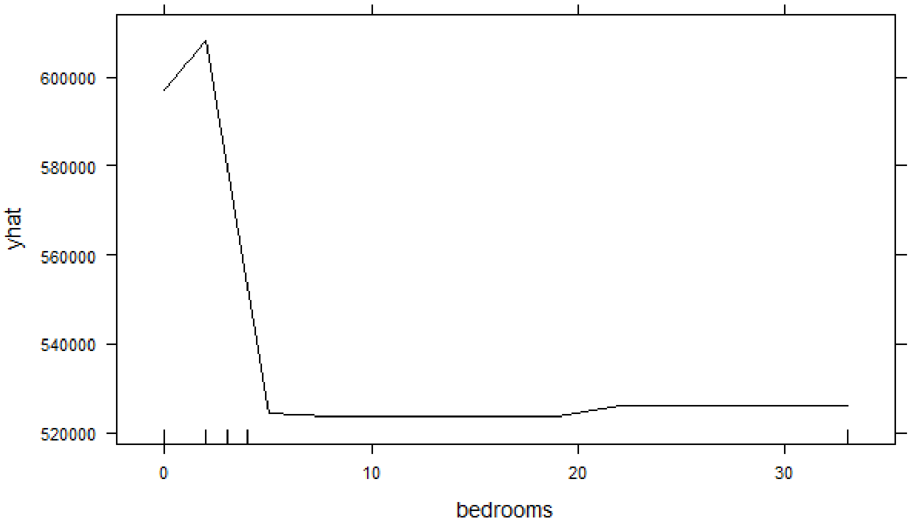

Figure 6: PDP: *price* vs *bedrooms*

- **Number of bathrooms**

  On the other hand, from figure 7, we can see how the number of bathrooms influences the price of the house. From this figure, we can conclude the following:

  - If the house has no bathroom or has a bathroom with toilet but no shower, the house price prediction is the lowest possible. However, given the low density, this conclusion is not very reliable.
  - If the house has between 1 bathroom and "3.5" bathrooms, the predicted price of the house is increased. This conclusion is reliable due to the large number of examples in this value range.
  - From "3.5" bathrooms onwards, the predicted house price increases, although, as in the first point, this conclusion is not at all reliable.

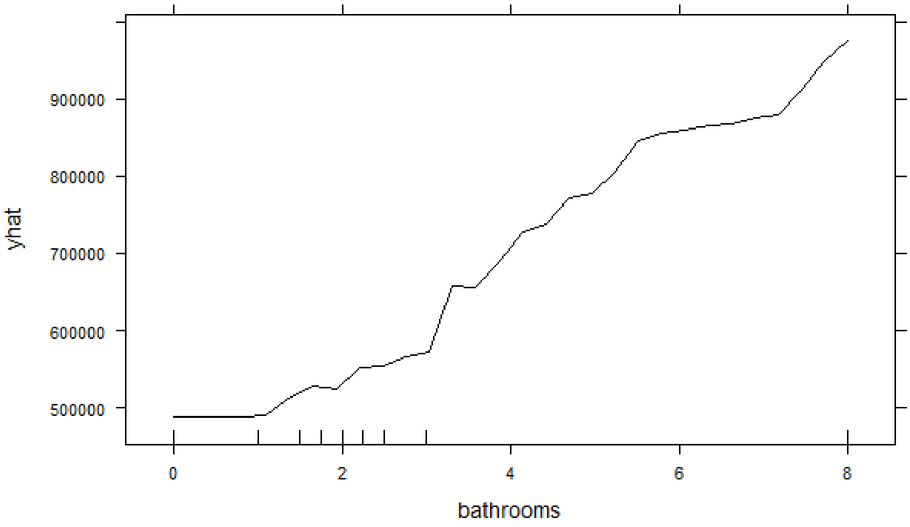

Figure 7: PDP: *price* vs *bathrooms*

- **Square footage of the apartment interior living space**

  As for the square footage of the house, from figure 8, it can be concluded that the predicted price of the house is increasing as the livable square footage does the same. When the square footage is between just over 0 and about $3000m^2$ the prediction is reliable, while from that price it is no longer reliable due to the low density of houses with so many square meters.

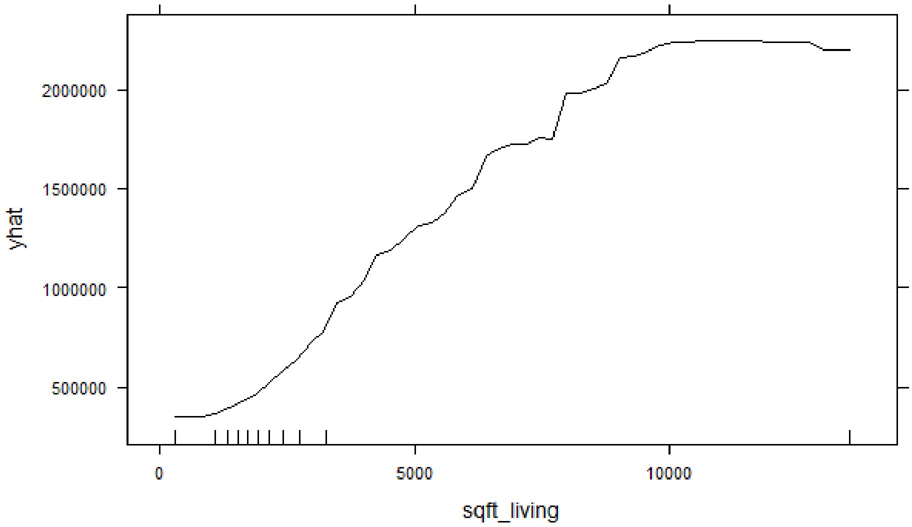

Figure 8: PDP: *price* vs *sqft_living*

- **Number of floors**

  Finally, as for the number of floors of the house, from figure 8, it can be concluded that the predicted price of houses with one floor is lower than that of houses with 2 floors. This conclusion is reliable due to the high density of examples with these numbers of houses.

  From 2 floors up to ¿3.5? floors, the predicted house price increases considerably, but the number of examples is not very high, so we must be cautious with this conclusion.

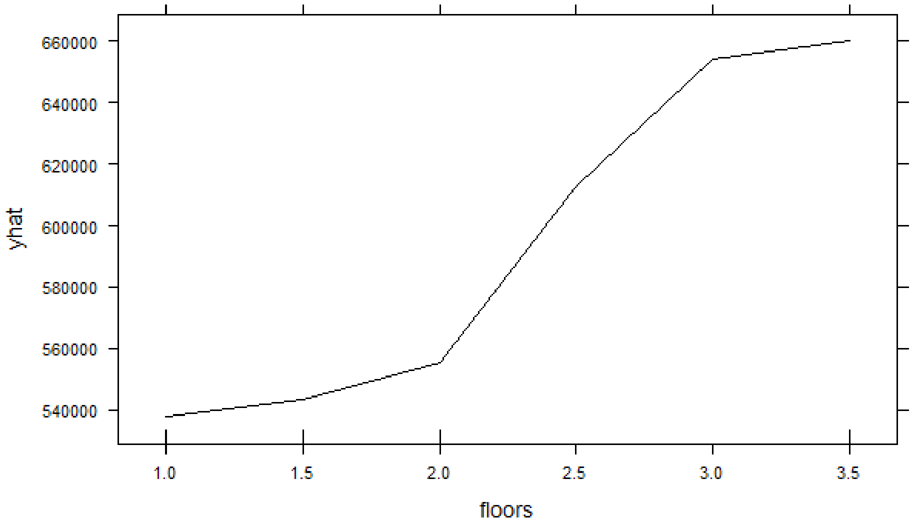

Figure 9: PDP: *price* vs *floors*

3. Conclusion

Thanks to the PDPs that have been presented throughout this report, we have been able to gain insight into how different variables influence the prediction of both the number of bicycles and housing prices. These visualization tools have allowed us not only to identify the relationships between the independent variables and the target variables, but also to understand the magnitude and direction of their impact.

Algunas conclusiones que pueden ser extraidas para cada caso son:

- **Bicycle rental forecasting**. It is concluded that the use of bicycles has, in general, been increasing since 2011. Moreover, users prefer to use this means of transport when weather conditions are moderate: when the temperature is not extreme (neither above nor below), when humidity is reasonable or when wind speed is moderate; the predicted number of rented bicycles is higher than in other cases.
- **Home price forecasting** It is concluded that the number of bedrooms, bathrooms, interior living space and the number of floors of a house have a significant impact on the price of the house. For example, price tends to increase with the number of bathrooms. Conversely, the price tends to decrease with the number of bedrooms (although with a reasonable number of bedrooms, 2 or 3, the opposite is true). In addition, price also increases with the increase in living space, and one-story homes tend to have a lower predicted price compared to those with two stories or more.

4. Appendix

As we have previously worked with *git*, it is assumed that we have git installed on our device and that it is linked to our GitHub account.

Having made this comment, the process of creation of the repository where the RMarkdown is stored is as follows:

1. The first thing to do is to create a repository on GitHub. Once this is done, we save the URL, which will be necessary to clone the repository on our device.

   This URL is: https://github.com/jose-valero-sanchis/edm_practica_5.git

2. To clone the repository locally, follow the steps below:

   a) Using the command `cd`, we locate ourselves in the directory where we want to have the clone of the repository.

   b) Using the following command:

   ```
   git clone https://github.com/jose-valero-sanchis/edm_practica_5.git
   ```

   we clone the GitHub repository locally. From now on, we will work in this directory and any modification made on it can also be updated in GitHub.

3. At this point, the repository is just a normal folder, so the first thing to do is to copy all the files needed to perform the practice (those available in the zip of *PoliformaT*). To make these changes show up on GitHub (so that the rest of the group members can access them), we use the following commands:

   a) `git cd edm_practica_5`, to move to the new repository.

   b) `git add .`, which adds the changes made in the repository to the staging area.

   c) `git commit -m "Añadiendo los datos"`, which create a new commit containing the current contents of the index and the given log message describing the changes.

   d) `git push`, which uploads local repository content to a remote repository.

   At this point, anyone who clones the repository as above and runs a `git pull`, will be able to see the updated resources.

4. Although it is usually a good practice to use different branches so that each component of the group can make the changes they consider without disturbing the rest, in this case, being a "project" of very small dimensions, we have all worked on `main`.

   Once this comment is made, we create an *Rmd* normally. We have not been able to do the practice in a single sitting, so every time a member of the group finished his part, he used the last 3 commands from the previous point so that the rest of the members (by doing a `git pull`) could work with the latest version.

5. Once the code is ready and uploaded to GitHub, we proceed to the creation of the report. As we have used *Overleaf* for its creation, we have only uploaded the latest version of *PDF*.