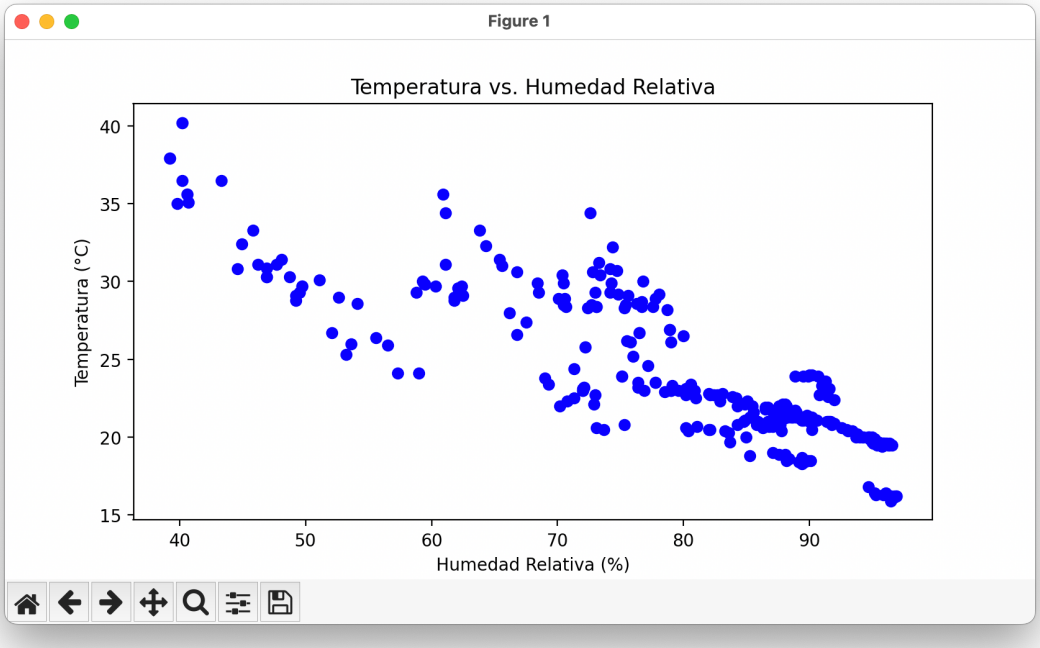
Analysis

1. Greenhouse

Measurements of temperature T (units in °C) and relative humidity R (units in %) in a real greenhouse of a rose crop are given in the greenhouse.txt file. The first column is temperature and the second is humidity.

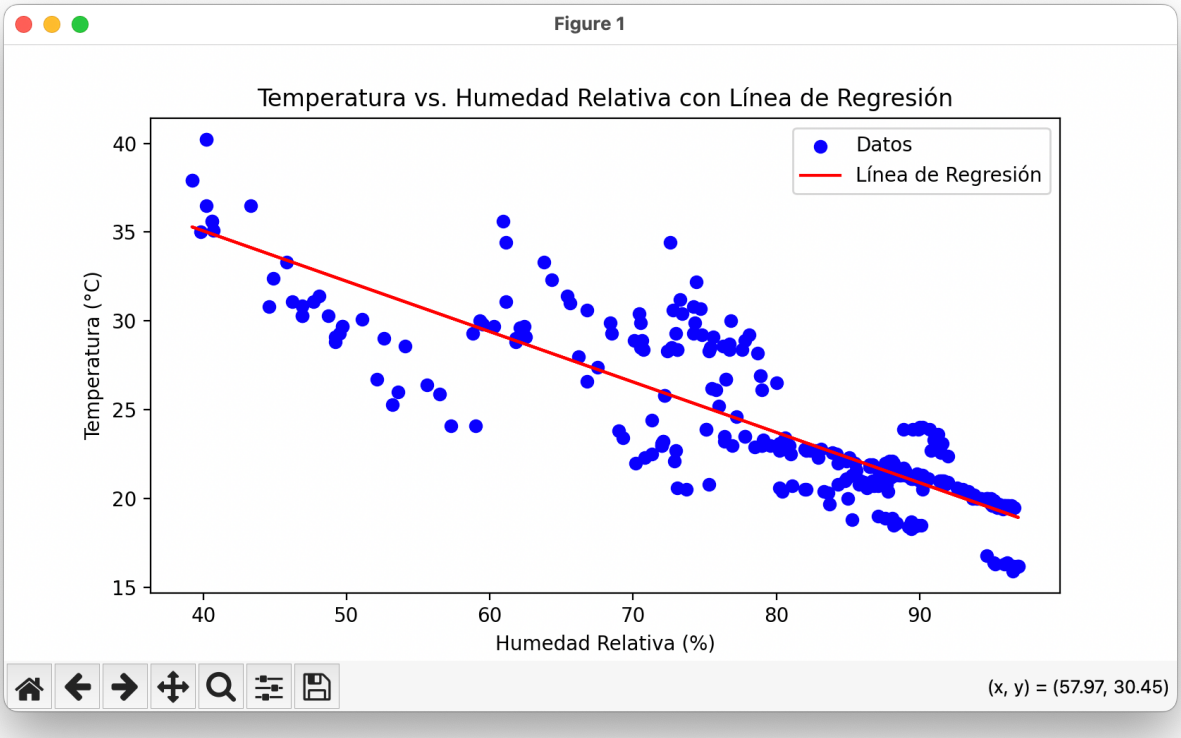
a) Plot the temperature vs. humidity on a scatter plot. Analyze this graph.



From the graph you can see an inverse relationship between both variables, as the relative humidity increases, the temperature decreases. This type of relationship is common in controlled environments such as greenhouses where humidity and temperature are regulated to optimize growing conditions. The dots do not follow an exact linear pattern, but show a general downward trend, which suggests that there is a negative correlation, this inverse relationship may be due to humidification

systems that, by increasing humidity, also help to lower the temperature, or to external environmental conditions that simultaneously affect both variables.

b) Assume that the relationship between humidity and temperature is given by the equation $T = \beta 0 + \beta 1R$. Using the data in greenhouse.txt, estimate $\beta 0$ and $\beta 1$ using least squares. Graph the line found with the overlapping temperature vs. humidity data.



The slope of the regression line is negative ($\beta<0$), confirming the inverse correlation between temperature and relative humidity. This suggests that, in this controlled environment, an increase in humidity is associated with a decrease in temperature.

The value of $\beta 0$ represents the expected temperature when humidity is zero (theoretically), while $\beta 1$ indicates the rate of change of temperature with respect to humidity. Although in this practical context, $\beta 0$ may not have physical significance if the near-zero moisture values are not realistic in the greenhouse, which is likely.

The approximation seems quite accurate, but there are some dispersions around the line, which is normal due to environmental factors or other variables not considered in the model.

c) Using the model obtained in statement b), define a rule to determine whether a new measurement of the pair (T, R) is abnormal or not. That is, if we have a new measurement of temperature and humidity with the greenhouse sensors, we can use the model $T = \beta 0 + \beta 1 R$ to know how much the new measurements follow that model. If they deviate too much, we could consider those measurements as anomalous.

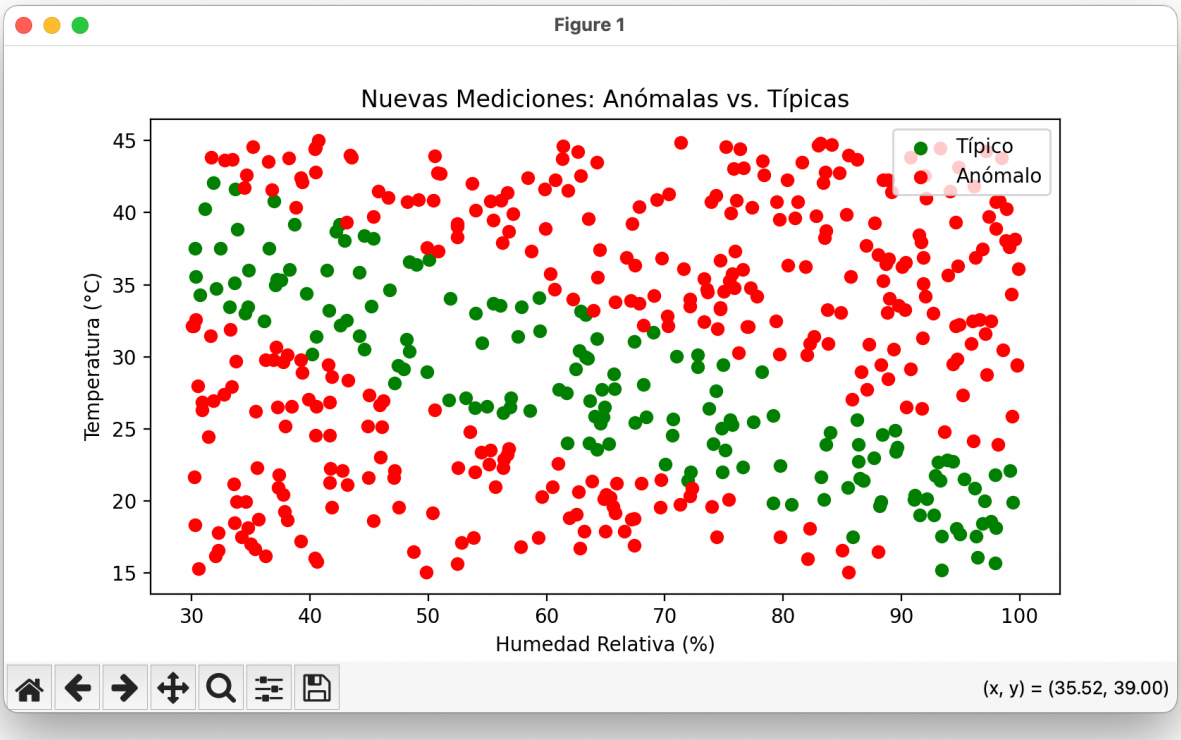
To determine if a new measurement (T, R) is anomalous, we will calculate the residual (difference between the observed temperature and the temperature predicted by our model):

$$\text{Residual} = T_{observado} - (\beta 0 + \beta 1 \times R_{observado})$$

We will consider a measurement to be anomalous if the absolute value of its residual is greater than twice the standard deviation of the residuals in the training data. The result of executing it was:

```
Estimated parameters: β₀ = 46.4062, β₁ = -0.2835
Waste standard deviation: 2.4582
Threshold for anomalies: ±4.9164
```

d) There are new measurements of the temperature and humidity pairs in the datosNuevos.txt file. Using the rule obtained in statement c) determine which pairs of measurements are anomalous and which are considered typical (normal). Plot these points in a scatter plot of temperature vs. humidity, and paint the anomalous ones with one color and those considered typical with another color. Identify the pattern in the detection that can be identified in the graph.



The scatter plot demonstrates that most "typical" conditions (green dots) occur within a specific band of moderate humidity (about 40-80%) and temperature (around 20-40 °C). This shows a comfort zone for the greenhouse where the temperature-humidity ratio is stable and within the expected range of the model.

Most of the anomalies (red dots) are concentrated at the extremes of the humidity scale (very high and very low humidity levels), suggesting that greenhouse temperature regulation might not align well with humidity changes in these ranges. This indicates a problem maintaining constant temperatures when humidity is at atypical levels. Another group of anomalies appears when the temperature is too high or too low for a given range of humidity, this points to possible external influences (such as sudden changes in ventilation, heating, or cooling) or sensor errors during periods when environmental conditions fluctuate sharply.

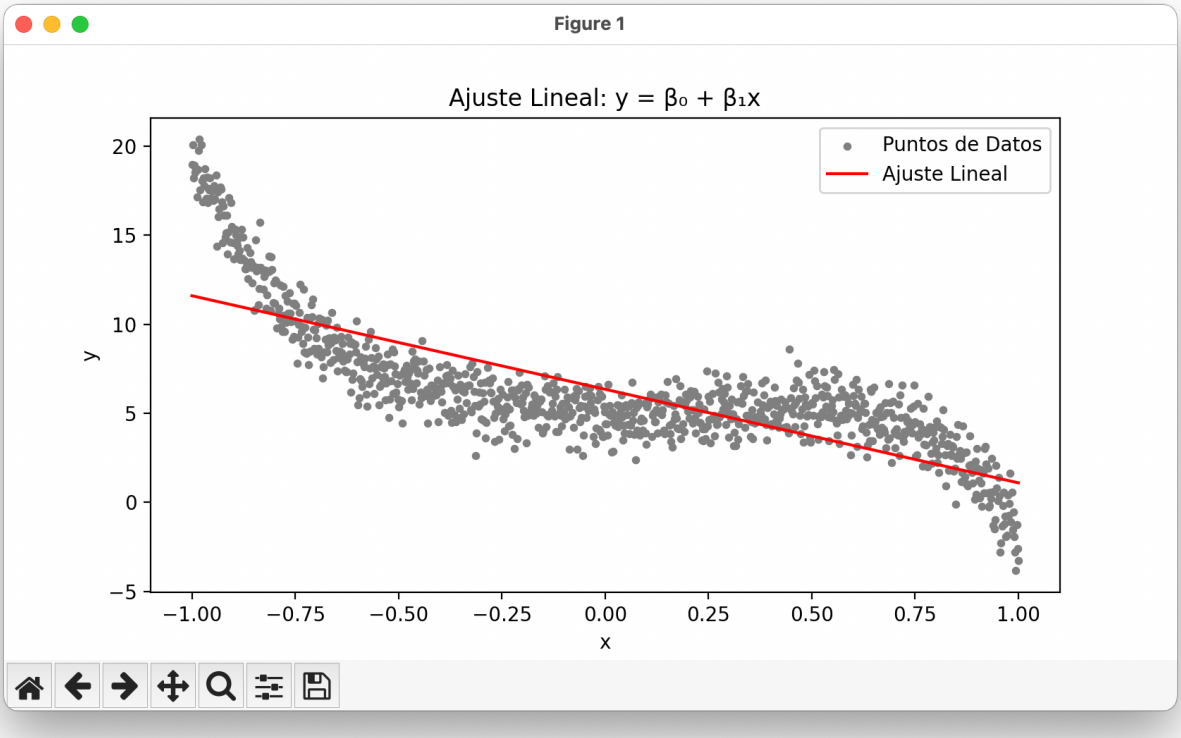
What we would recommend to the Greenhouse is that it adjust the heating, cooling and ventilation based on the anomalies observed at extreme humidity levels to maintain constant temperatures. It would be good if they set up an alert system that signals when humidity or temperature readings are approaching areas of known anomalies (identified in the scatter plot) to allow for real-time interventions.

2. Toy Data

Consider the data with the pairs (x,y) in the datosToy.txt file.

a) Assume that these data are related through the relationship $y = \beta_0 + \beta_1 x$.
Using least squares, estimate the parameters β_0 and β_1.
Plot the data on a scatter plot (y vs x) and plot the estimated line.
Analyze the results.

Estimated parameters for the linear model:
$b_0 = 6.3466$
$b_1 = -5.2480$



The red line represents linear fit, but it is noticeable that this model does not follow the curvature of the data. This shows that the data do not fit well into a simple linear relationship and that a more complex model will be able to offer a more accurate approximation. By itself, the way the data points are scattered around the adjustment line suggests a nonlinear relationship, because it starts going down, then goes up a bit, and then goes back down.

b) Assume that these data are related through the relationship $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

Using least squares, estimate the parameters β_0, β_1, and β_2.

Plot the data on a scatter plot (y vs x) and graph the estimated parabola.

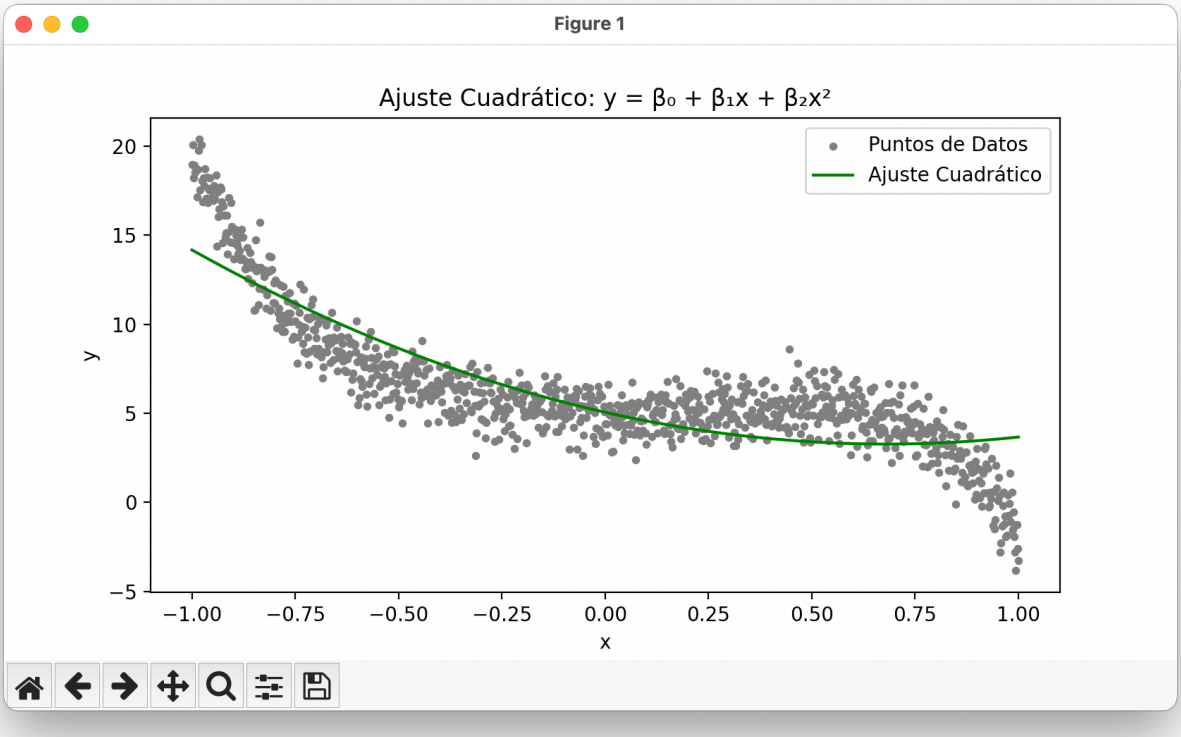
Analyze the results.

Estimated parameters for the quadratic model:

$b_0 = 5.0580$

$b_1 = -5.2480$

$b_2 = 3.8581$



The quadratic fit better follows the overall trend of the data compared to the linear model. In the range of x between about −1 and 0.75, the quadratic model seems to fit reasonably well to the decreasing trend of the data.

However, at extreme values of x, especially near x=1, there is still some deviation, showing that even the quadratic model is not completely adequate in these areas. This clearly shows the need for higher-order terms (such as a polynomial fit) to fully capture the behavior of data at the endpoints.

c) Assume that these data are related through the relationship $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

Using least squares, estimate the parameters β_0, β_1, β_2, and β_3.

Plot the data on a scatter plot (y vs x) and plot the estimated polynomial.

Analyze the estimated results. Analyze the results.

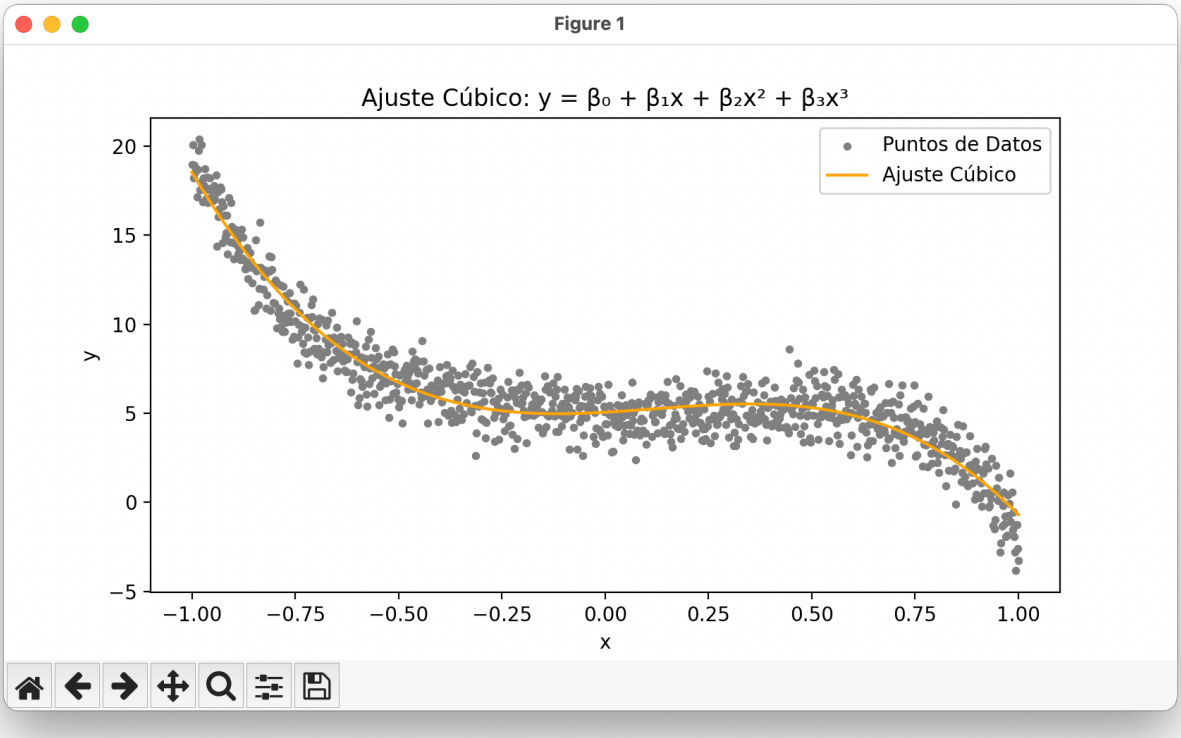
Estimated parameters for the cubic model:

$b_0 = 5.0580$

$b_1 = 1.3177$

$b_2 = 3.8581$

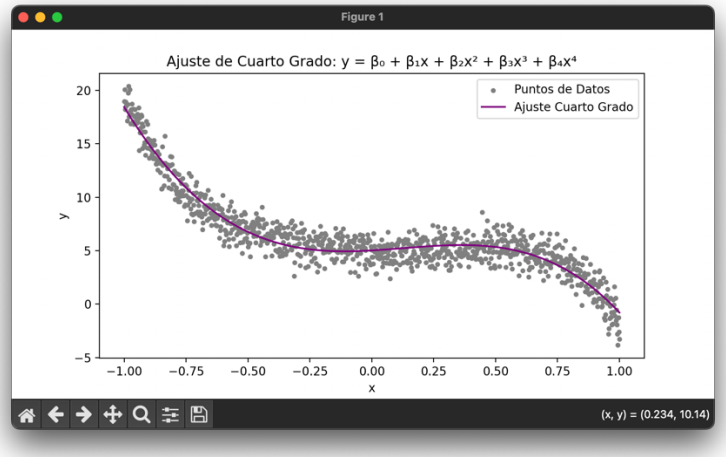
$\beta_3 = -10.9210$



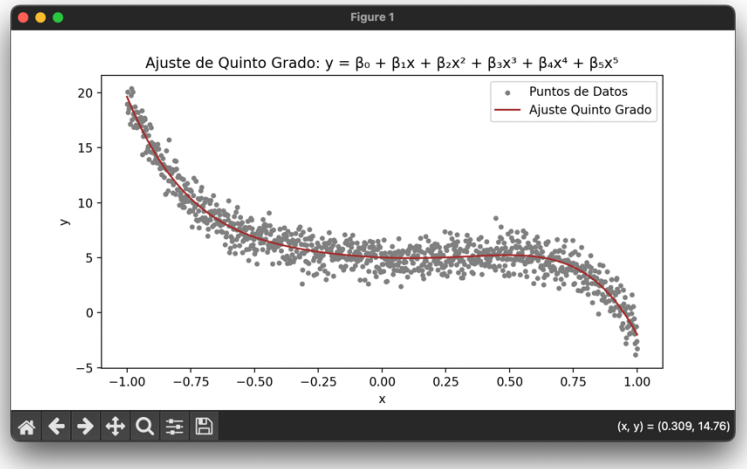
The flexibility of the term allows the model to adapt to the curved shape and change direction at these ends, achieving a more precise fit. The coefficients B2 and B3 are responsible for capturing the additional curvatures in the data, while B0 and B1 provide the linear basis of the model. The negative value of B3 is crucial, as it allows the curve to decrease towards the extremes, adapting to the decrease in y when x moves away from the center. The cubic model provides the best fit yet, capturing not only the overall trend but also the subtlest variations in the data along the range of $\boldsymbol{x^3}$x. This indicates that the relationship between x

and y is highly nonlinear and fits best with a polynomial of order three, which offers the flexibility needed to model the complex changes observed in the data.

Additionally, to corroborate that this is the best approximation, we tested with 2 more cases, an approximation of degree 4 polynomial:



And a grade 5 approximation.

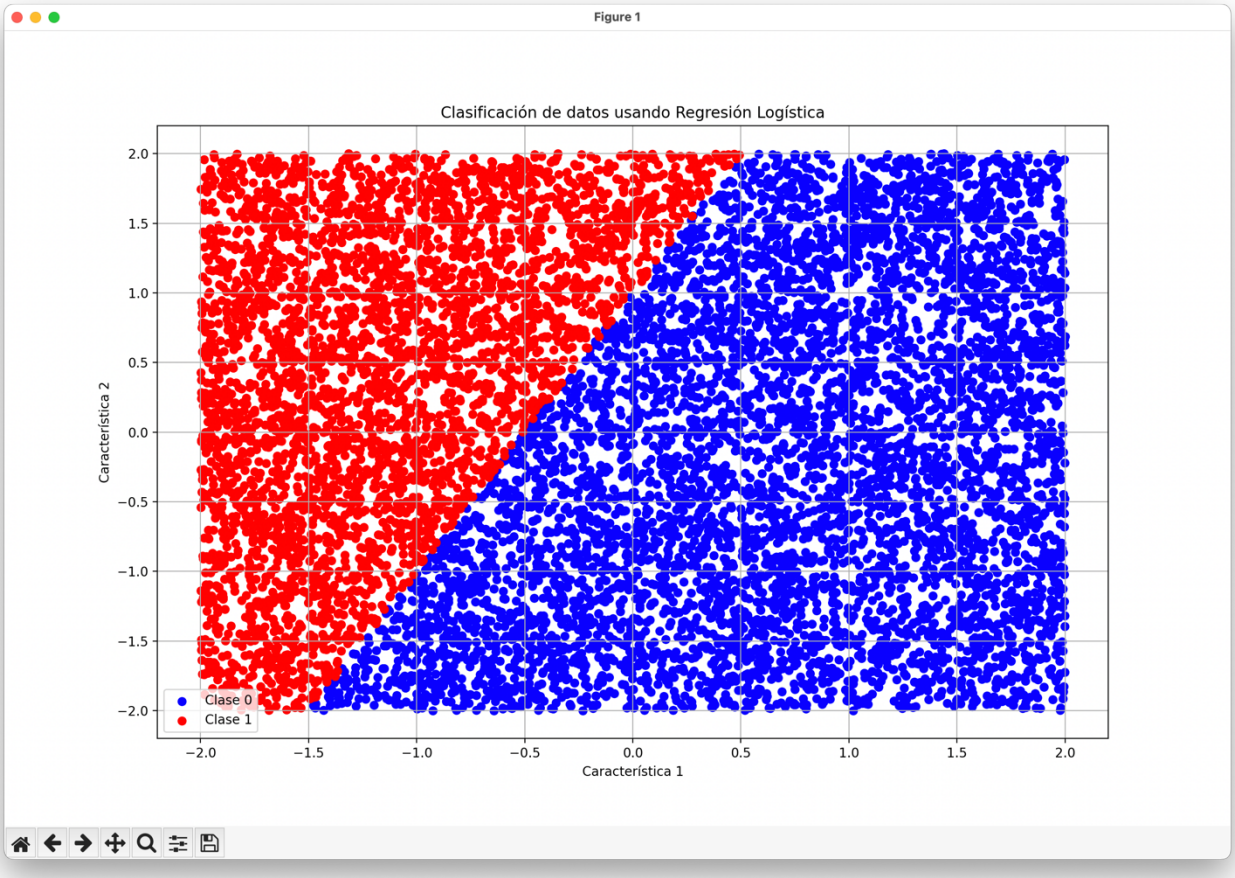


What was found was that the approach does not change much, so there is no point in continuing to move up a grade, since the 3 approximates it quite well.

3. Logical Regression

After applying logistic regression on a set of labeled data defined in a space of 2 characteristics, the following weights were obtained: w=[-2,1]^T and b=-1. The datosNN.txt file contains 1000 2-dimensional vectors.

For each of these 2D points, apply the decision rule (each of these vectors would be x), and classify them according to the values obtained from y. Generate a scatterplot of the data on a 2D plane where the color coding depends on the respective labels (i.e. only two colors).



In the scatter plot we clearly see a line separating the two classes, which is the "decision frontier" created by the logistic regression model. This boundary is the place where z = 0 (i.e., when). The sigmoid function assigns a probability value close to 0.5 to the points on this boundary. Since the weights are w = [-2, 1] and the bias is b = -1, this means that the decision boundary leans more towards the x-axis, reflecting how these weights influence the classification and how the negative weight for the first characteristic causes the impact of this characteristic to be inverse on probability, This segments the data into two regions, each associated with a different class, based on the characteristics and weights obtained, achieving a perfect separation.$w^t x + b = 0$

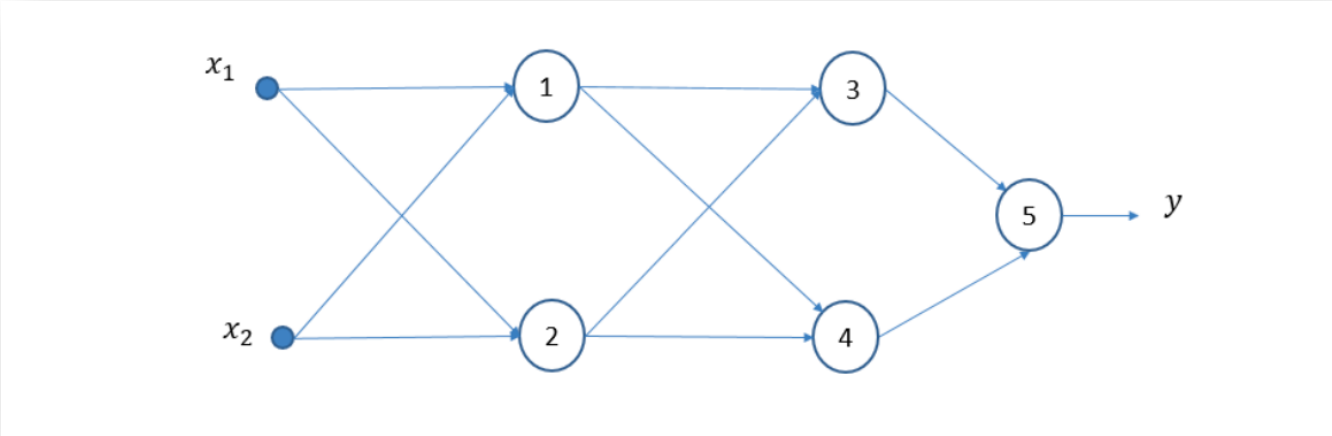
4. Neural Network

Consider the neural network in Figure 1, with input $x = [x1, x2]^T$, and output y. This neural network is designed for binary classification of the input x: if the output y is >= 0.5, it assigns x a category 1, while if it is <0.5 it assigns a category 0. The weights and firing function of each of these neurons are available in Table 1.

a) Implement a routine that takes a vector $x = [x1, x2]^T$ and propagates it through the neural network to obtain y. You can't use predefined neural network packages or functions to do this routine. Use the standard vector multiplication and function definition operations in Python to do this.

b) The archivodatosNN.txt contains 1000 2-dimensional vectors. Pass all of these through the neural network (each of these vectors would be x), and classify them according to the obtained values of y. Generate a scatterplot of the data on a 2D plane where the color coding depends on the respective labels (i.e. only two colors).

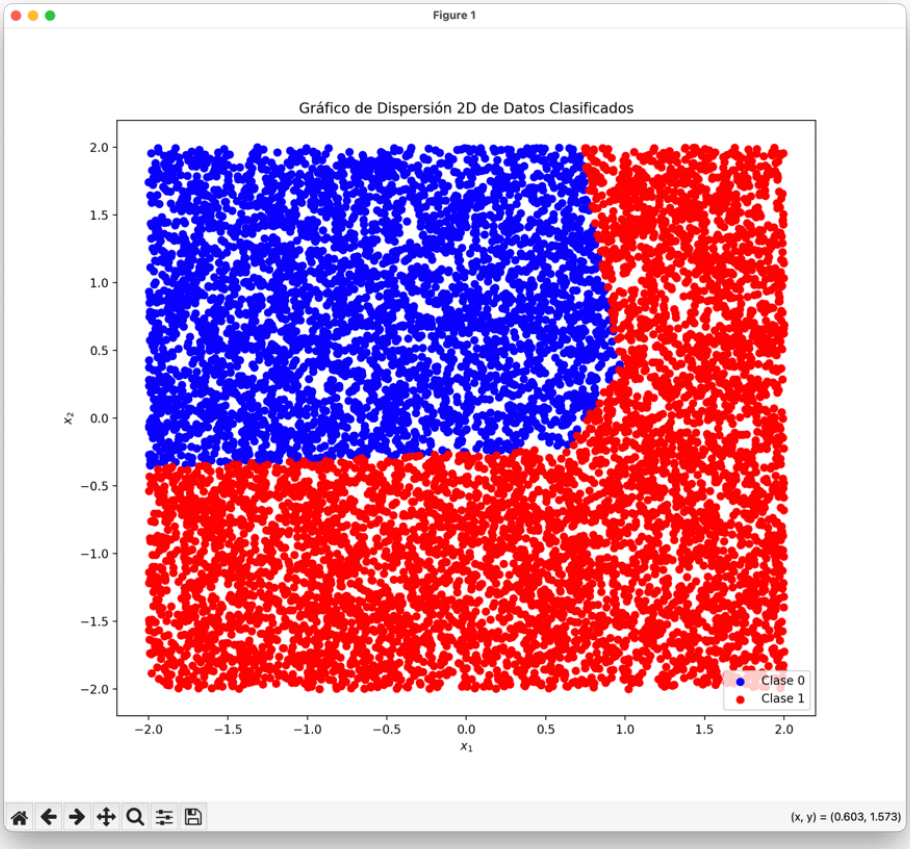
c) Analyze the results in b). In particular, discuss the decision boundary for classification.



Weights and activation functions of each neuron in the neural network in Figure 1.

Neuron 1	$w = [0.1188, -2.1436]^T$, $b = 0.7065$ ReLu Activation Function
Neuron 2	$w = [1.6878, 0.2615]^T$, $b = -1.0121$ ReLu Activation Function
Neuron 3	$w = [-0.4082, -0.8420]^T$, $b = 0$ ReLu activation function
Neuron 4	$w = [1.1433, 1.9843]^T$, $b = -0.1435$ ReLu Activation Function
Neuron 5	$w = [-1.3052, 1.6268]^T$, $b = -2.1021$ Sigmoid activation function

The decision boundary is clearly visible on the graph as a horizontal separation between the Class 0 and Class 1 points. Most of the dots in the upper half of the plane are classified in Class 0 (blue), while those in the lower half are in Class 1 (red). This suggests that the neural network has learned a roughly linear decision boundary that separates points on the vertical axis based on their x2 values.



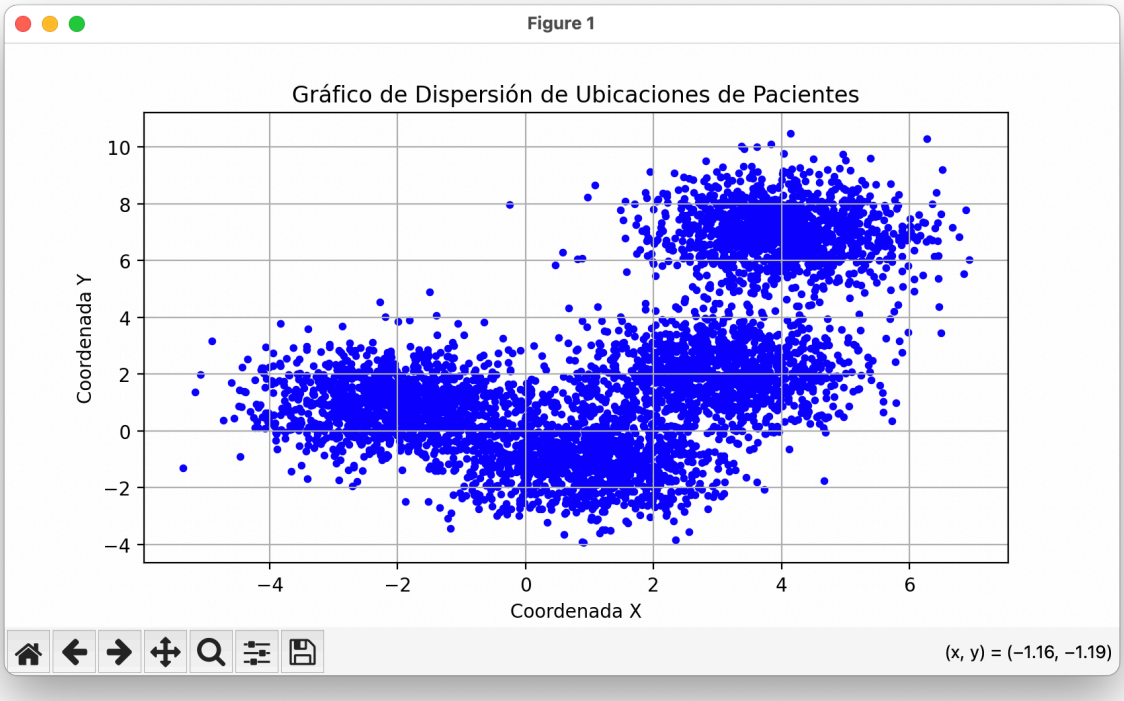
The weights of the neurons in the first layer have a significant impact on how the network processes the x1 and x2 features, the neurons in the first layer (which use ReLU) pass activated information in an intermediate layer structure, while the output neuron (which uses sigmoid) generates a probability that is compared to the threshold of 0.5, The decision boundary results from the combination of these activations and reflects a stronger dependence on the X2 characteristic due to the relative weight of the connections. ReLU facilitates the creation of nonlinear and complex decision boundaries, better suited to high-dimensional data and nonlinear patterns, on the other hand, the sigmoid function tends to create smoother decision boundaries and may struggle with complex patterns, as its output flattens out when the input values are very positive or very negative. In our network, the sigmoid function in the output creates a binary classification that, as we see in the scatter plot, produces a more linear or simpler decision boundary. This fits well with the current data, but if the data had a more complex classification structure, the network might not capture these patterns accurately, there would be a need to increase the complexity of the model (more layers and neurons) or adjust the hyperparameters.

5. K-Means

In the midst of a pandemic, the geographical location of infected people is known, and it is desired to locate several care centers from where rapid and timely care can be provided. For this, the centers should be located where the greatest sources of infection are. In other words, the centers must be located and the infected must be assigned to the care centers in such a way that the distance between the centers and the position of the infected assigned to the respective centers is as minimal as possible. Suppose that you have access to a total of 4 care centers, and that a total of 5000 infected people were detected. So, this problem could be formulated mathematically as follows:

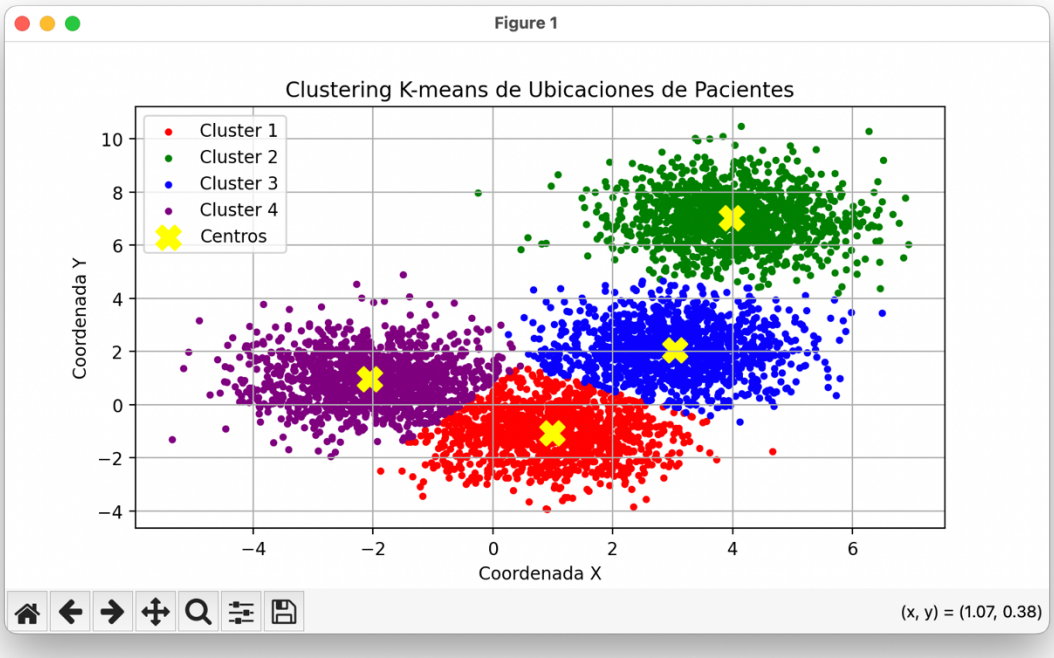
Let $x_i \in R2$ be the geographical position of the ith infected, and $\mu_j \in R2$ be the position of the j-th center of attention. The degree of "belonging" of the infected i-th to the j-th center is quantified by the variable $r_{ij} \in \{0.1\}$, where $r_{ij}=0$ means that the infected i is not assigned to the j center, and $r_{ij}=1$ means that the infected i is assigned to the j center. The variables of the problem are μ_j and r_{ij}. In this problem, the aim is to locate and assign the μ_j centres in such a way that the sum of the distances between the infected to each assigned centre is as minimal as possible.

a) Consider the coordenadasPacientes.txt file. This file contains the coordinates of the 5000 infected patients. Plot the geographical locations of the patients on a scatter plot. Analyze the distribution of locations.



The graph reveals at least 4 well-defined point clusters, with high concentrations around certain coordinates. The main clusters appear to be aligned horizontally along the Y-axis, with relatively clear spacing between each cluster and one scattered higher up than the rest. This shows the different geographical areas where infections are most frequent, which would be ideal for the placement of care centers if they were distributed properly.

b) Implement the K-means algorithm to solve this problem. You can use your own implementation or some Python package to solve this (for example, the sklearn library has an implementation of this algorithm). Plot a scatter plot with the location of the patients and color each dot with a color code that depends on the dot's belonging to a care facility (i.e., they must paint four different colors). In the same plot, graph the location of the centers. Analyze the results obtained.

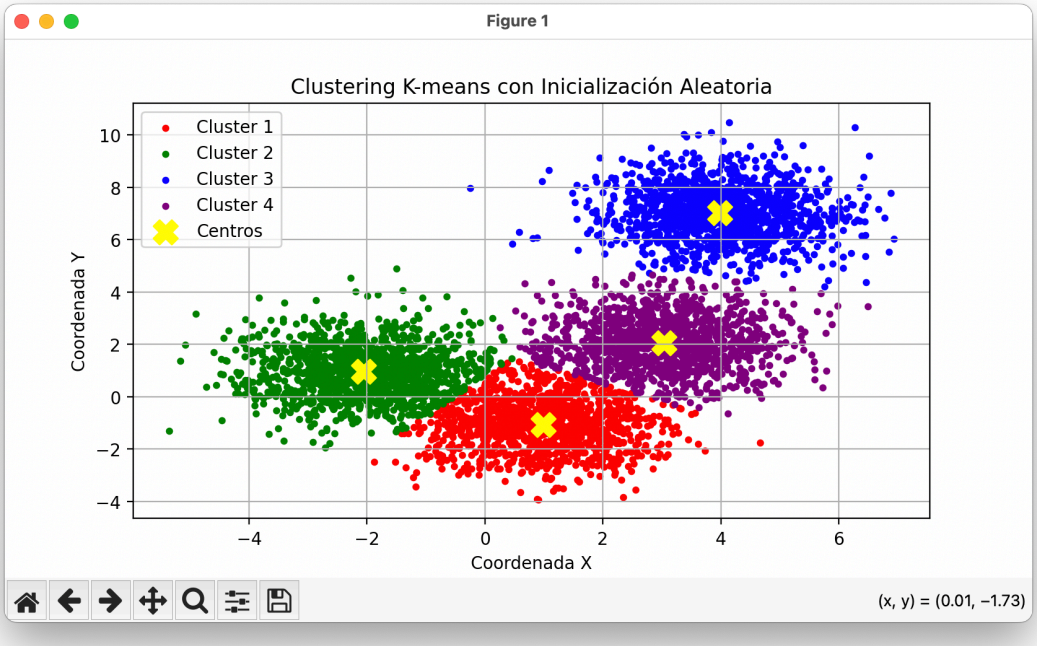


The patients were grouped into four well-defined clusters (red, green, blue and purple), the K-means algorithm managed to effectively segment the infected ones based on their proximity to the locations of the centers, we see this because the location of each center in the middle of the clusters shows that each group has a more or less uniform shape and density around its centroid, which is optimal to minimize the response time of each center to the assigned patients.

This type of segmentation facilitates the allocation of resources, since each center can focus on serving a specific set of patients within its area of influence. In addition, it minimizes the distance that patients or healthcare personnel must travel, thus improving the efficiency of response time, with a distribution of centers in this way,

resources can be organized and mobilized more efficiently to serve the patients of each cluster.

c) Run the algorithm with a different initial condition, and compare the results obtained in b) (in the implementation of the sklearn library you can use the initialization method as 'random').



The distribution of the clusters in this run is similar to the previous one in terms of general location, with four well-defined groups. However, due to random initialization, cluster hubs have slight differences in their positions compared to previous execution. In this run, the care centers (marked with a yellow "X") have been adjusted somewhat differently. For example, the center in the red cluster appears to be slightly offset compared to the previous run. This is due to the stochastic nature (which operates using probabilistic methods to troubleshoot) of the algorithm when using random initialization.

Normal k-means initialization tends to be more efficient at convergence, reaching a near-optimal solution in fewer iterations. With random initialization, the algorithm may need more iterations to converge and is more susceptible to getting stuck in local lows. Robustness: In general, the clustering maintains a similar structure, which indicates that the K-means algorithm can adapt and find valid partitions even with different initializations, although k-means without random initialization provides greater consistency.

References

- Ideal greenhouse temperature and humidity. (2022, julio 28). Atlas Scientific. https://atlas-scientific.com/blog/ideal-greenhouse-temperature-and-humidity/
- ReLU vs Sigmoid Function in Deep Neural Networks. Recuperado el 5 de noviembre de 2024, de W&B; Weights & Biases, Inc. https://wandb.ai/ayush-thakur/dl-question-bank/reports/ReLU-vs-Sigmoid-Function-in-Deep-Neural-Networks--VmlldzoyMDk0MzI

- kmeans. (n.d.). Unioviedo.es. Retrieved November 3, 2024, from https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html
- Wikipedia contributors. (s/f). Estocástico. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Estoc%C3%A1stico&oldid=161969788