Data science

Data science is the use of statistics, through data analysis. It uses plenty of methods to extract useful information from the data and make decisions.

Goal definition.

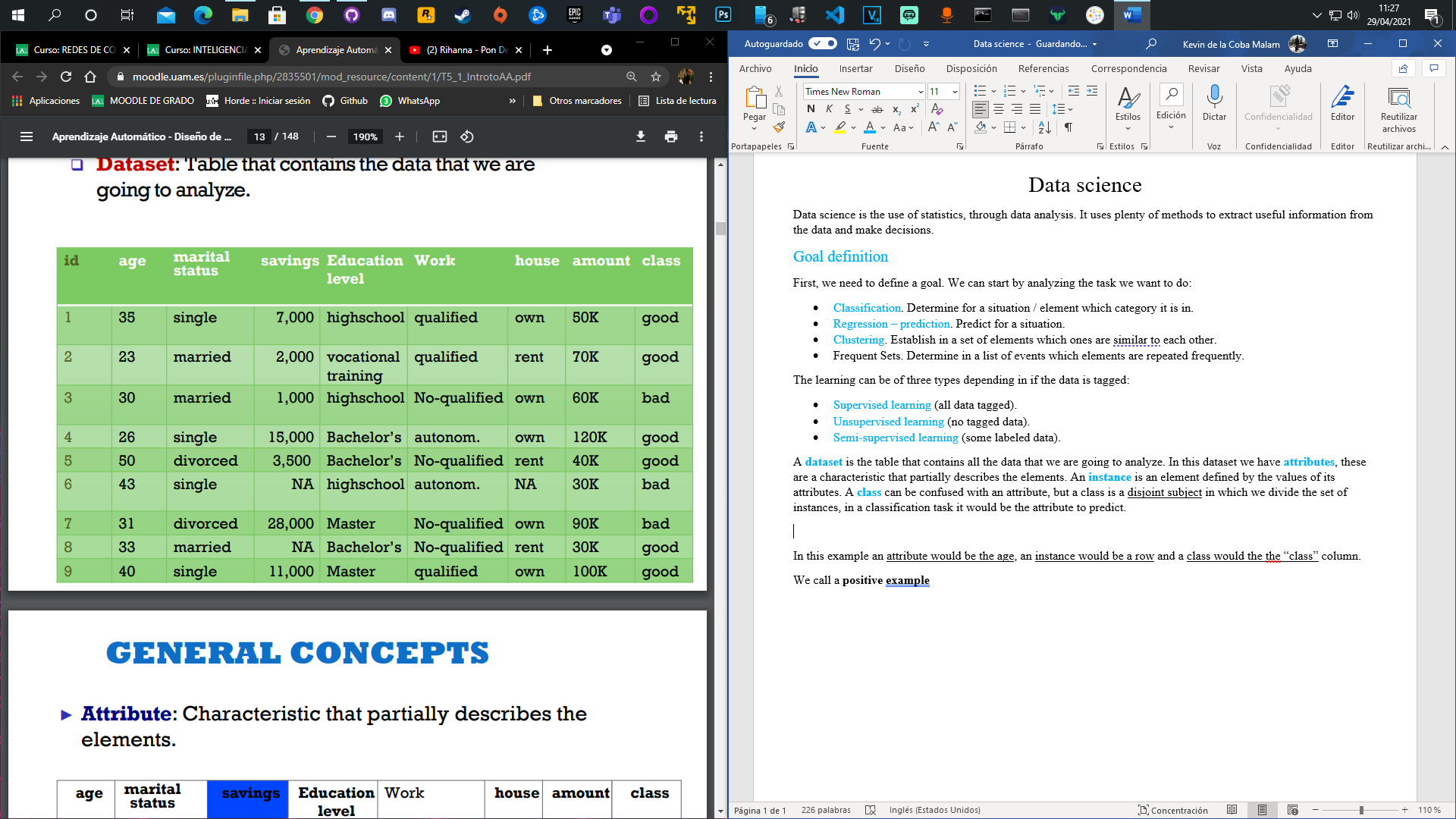
First, we need to define a goal. We can start by analyzing the task we want to do:

* Classification. Determine for a situation / element which category it is in.
* Regression – prediction. Predict for a situation.
* Clustering. Establish in a set of elements which ones are similar to each other.
* Frequent Sets. Determine in a list of events which elements are repeated frequently.

The learning can be of three types depending in if the data is tagged:

* Supervised learning (all data tagged).
* Unsupervised learning (no tagged data).
* Semi-supervised learning (some labeled data).

A **dataset** is the table that contains all the data that we are going to analyze. In this dataset we have **attributes**, these are a characteristic that partially describes the elements. An **instance** is an element defined by the values of its attributes. A **class** can be confused with an attribute, but a class is a disjoint subject in which we divide the set of instances, in a classification task it would be the attribute to predict.



In this example an attribute would be the age, an instance would be a row and a class would the the “class” column.

We call a **positive example** an instance that belongs to the subset defined by the class. A **negative example** of a class would be an instance that does not belong to the subset defined by the class.

Collection and understanding of data.

First, we must look at all the columns and read them. Then we can generate basic stats about the dataframe (each column). Stats such as the **mean**, **standard deviation**, **quartiles**.

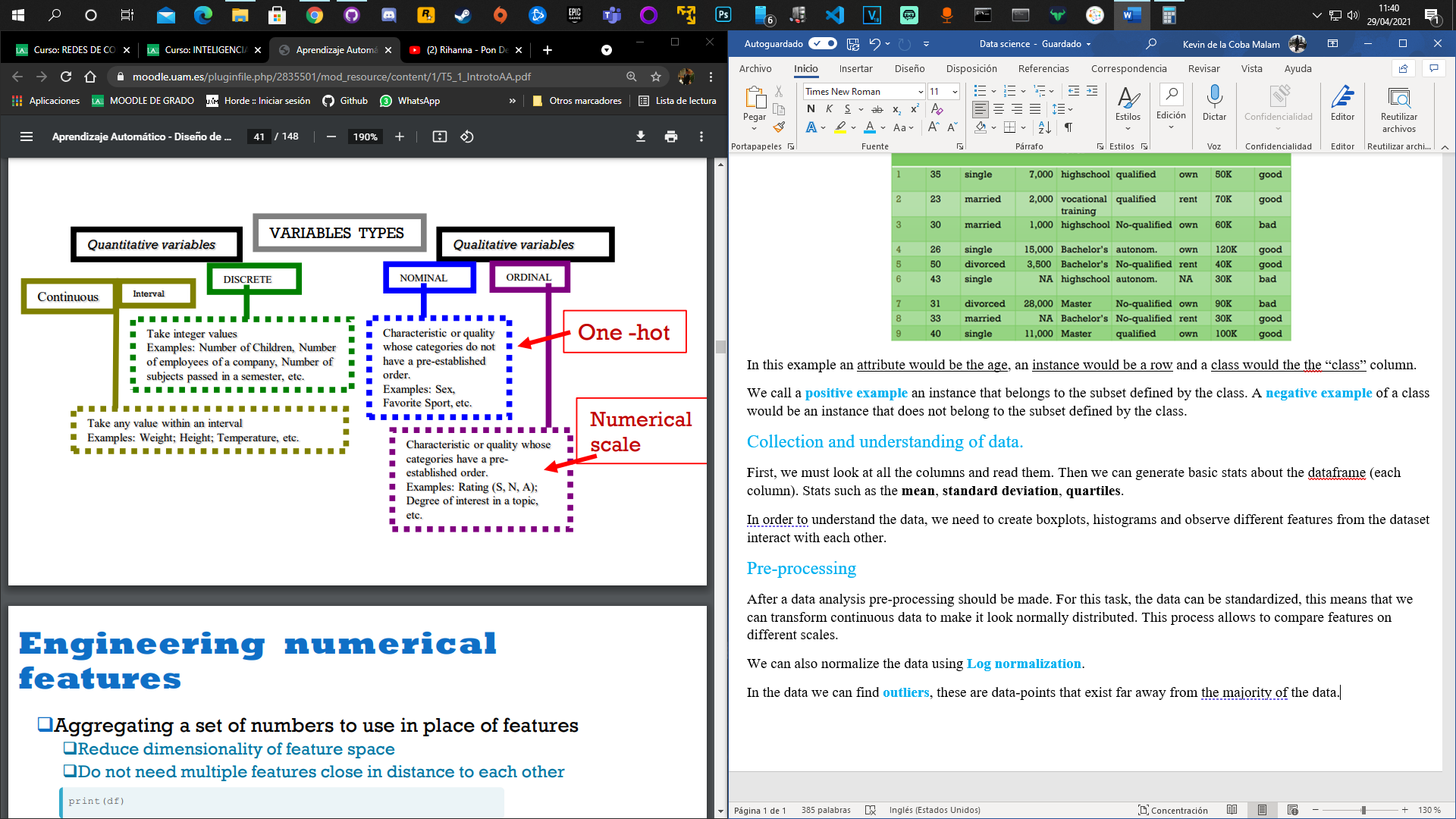
In order to understand the data, we need to create boxplots, histograms and observe different features from the dataset interact with each other.

Pre-processing

After a data analysis pre-processing should be made. For this task, the data can be standardized, this means that we can transform continuous data to make it look normally distributed. This process allows to compare features on different scales.

We can also normalize the data using **Log normalization**.

In the data we can find **outliers**, these are data-points that exist far away from the majority of the data.



We can **select features** to be used for modelling. This does not create new features but improves the model performance. Also, we can remove redundancies.

Modeling.

The methodology would be to first collect the data, then explore a data analysis by the representation of data, choice of variables. After this point we can classify the choice and then train our model. After the training we validate and then evaluate. Finally, we make a decision.

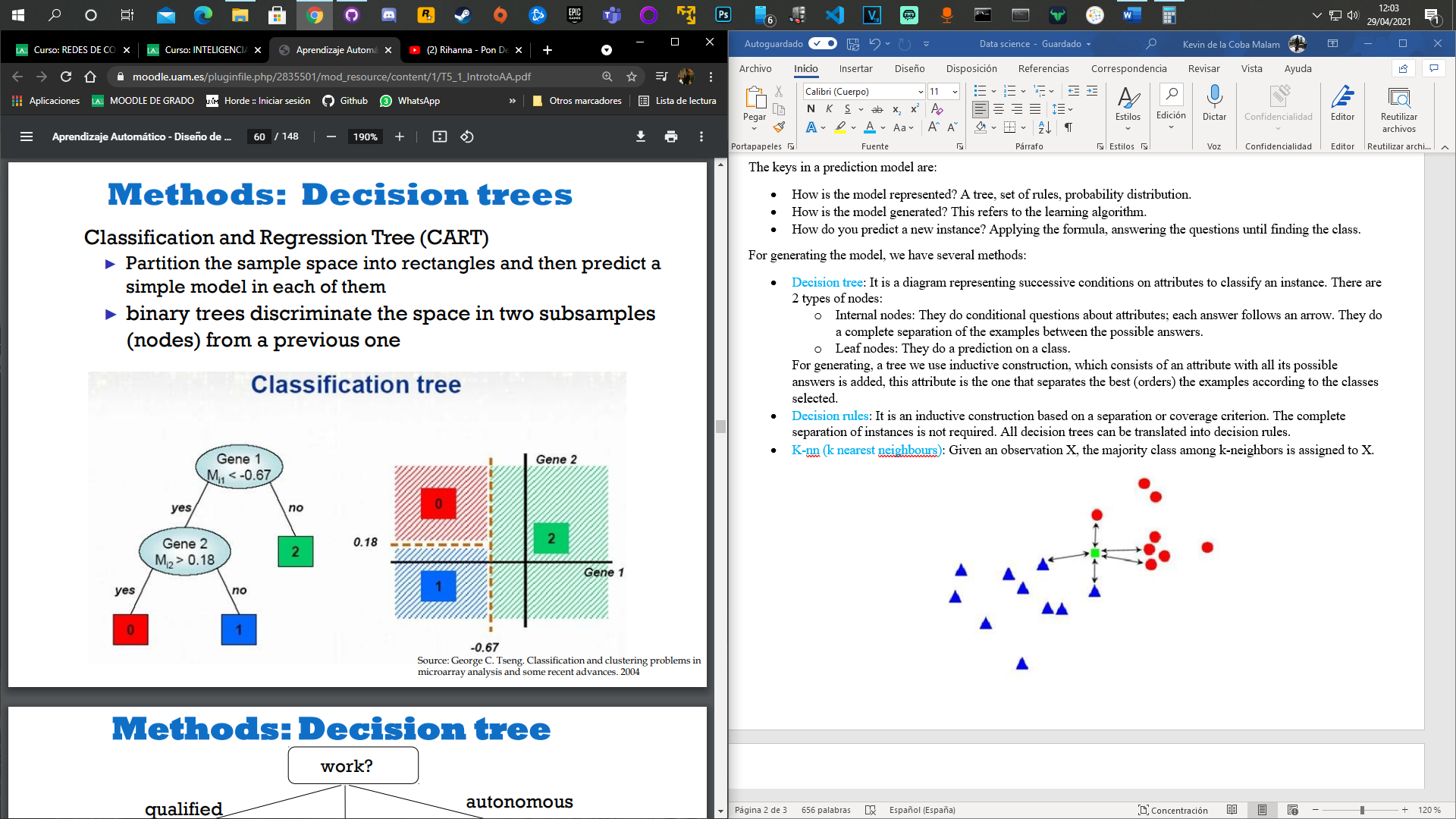
The keys in a prediction model are:

* How is the model represented? A tree, set of rules, probability distribution.
* How is the model generated? This refers to the learning algorithm.
* How do you predict a new instance? Applying the formula, answering the questions until finding the class.

For generating the model, we have several methods:

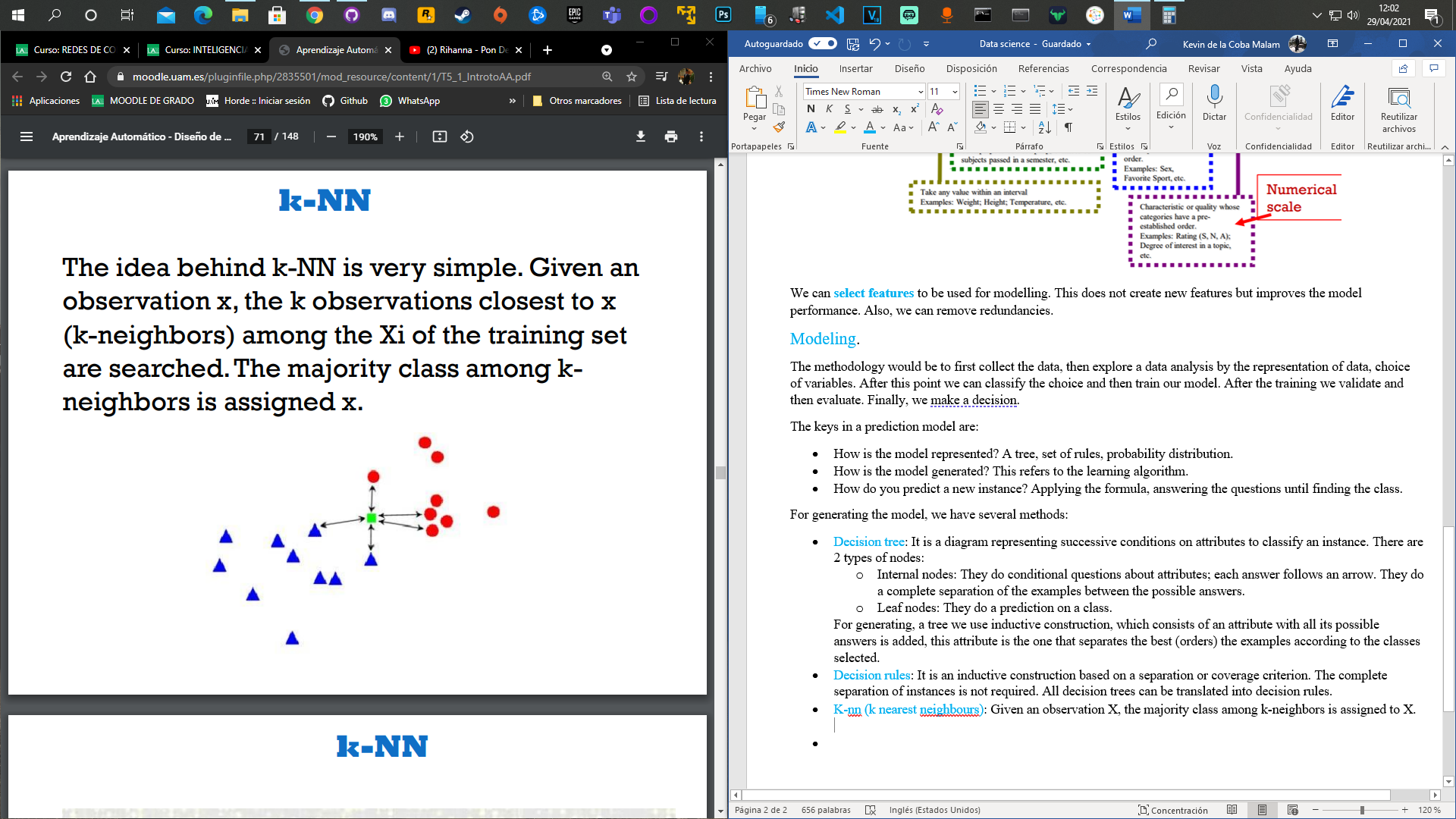
* Decision tree: It is a diagram representing successive conditions on attributes to classify an instance. There are 2 types of nodes:
  + Internal nodes: They do conditional questions about attributes; each answer follows an arrow. They do a complete separation of the examples between the possible answers.
  + Leaf nodes: They do a prediction on a class.

For generating, a tree we use inductive construction, which consists of an attribute with all its possible answers is added, this attribute is the one that separates the best (orders) the examples according to the classes selected.



* Decision rules: It is an inductive construction based on a separation or coverage criterion. The complete separation of instances is not required. All decision trees can be translated into decision rules.
* K-nn (k nearest neighbours): Given an observation X, the majority class among k-neighbors is assigned to X.

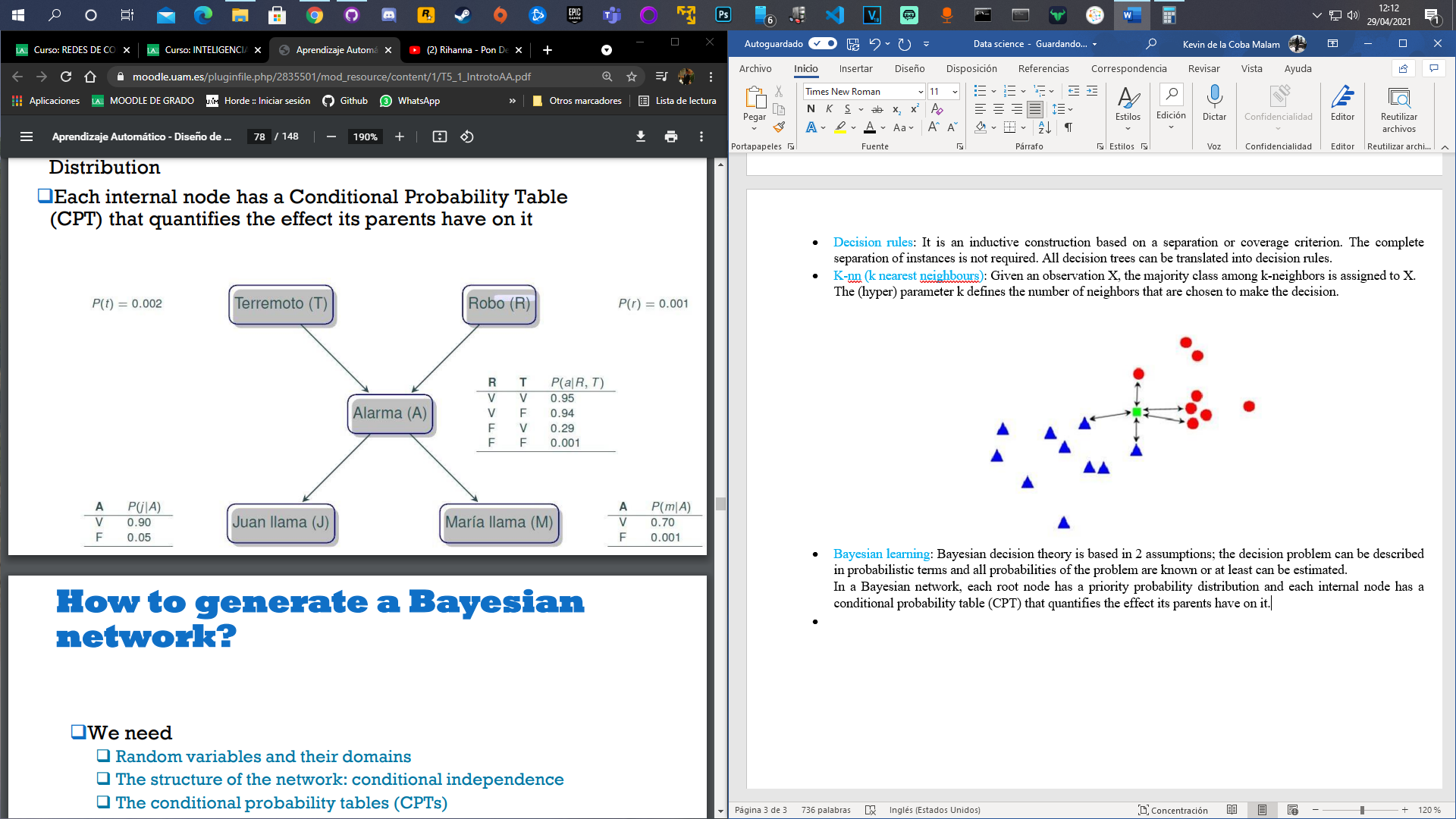
The (hyper) parameter k defines the number of neighbors that are chosen to make the decision.

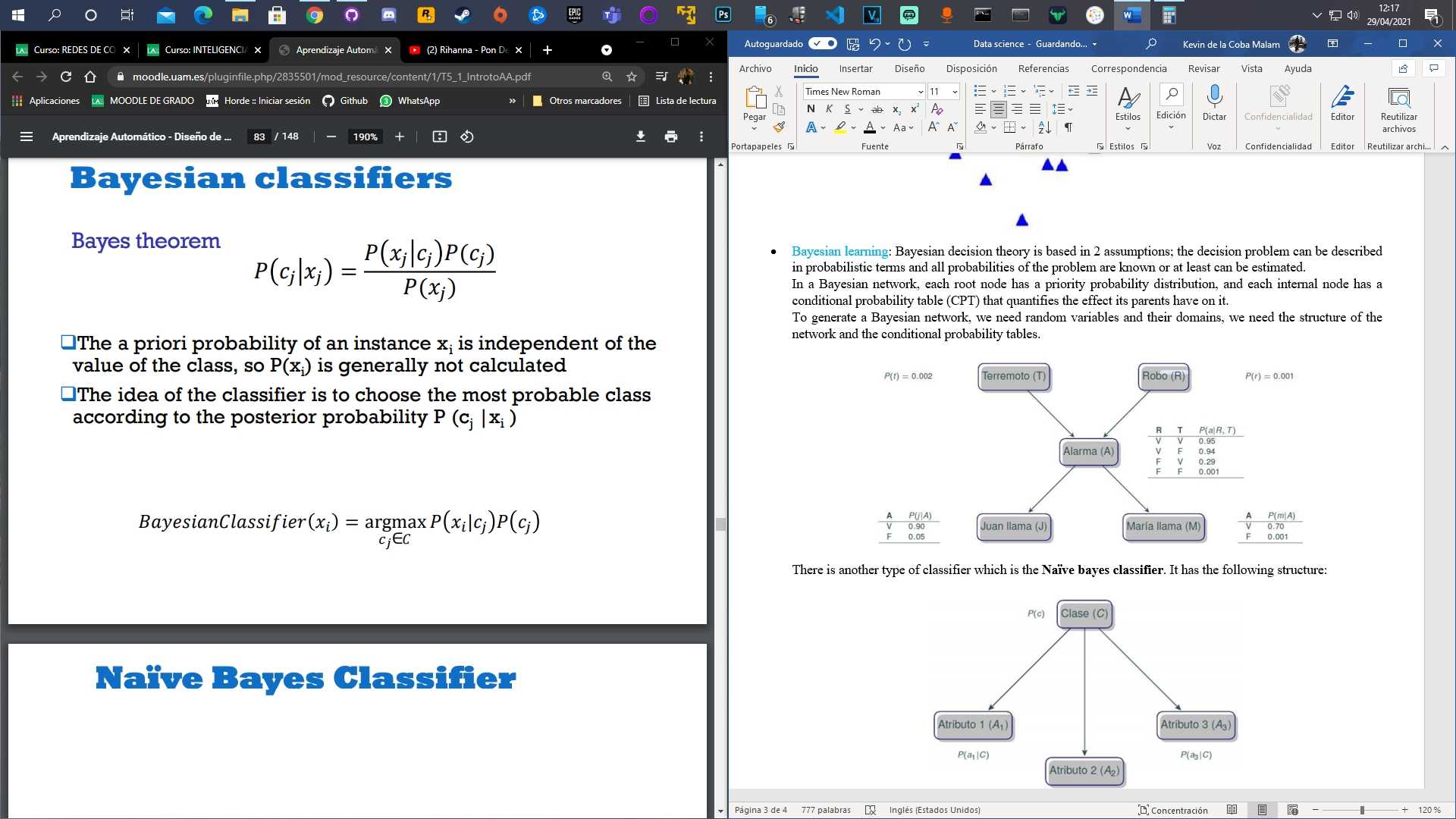


* Bayesian learning: Bayesian decision theory is based in 2 assumptions; the decision problem can be described in probabilistic terms and all probabilities of the problem are known or at least can be estimated.

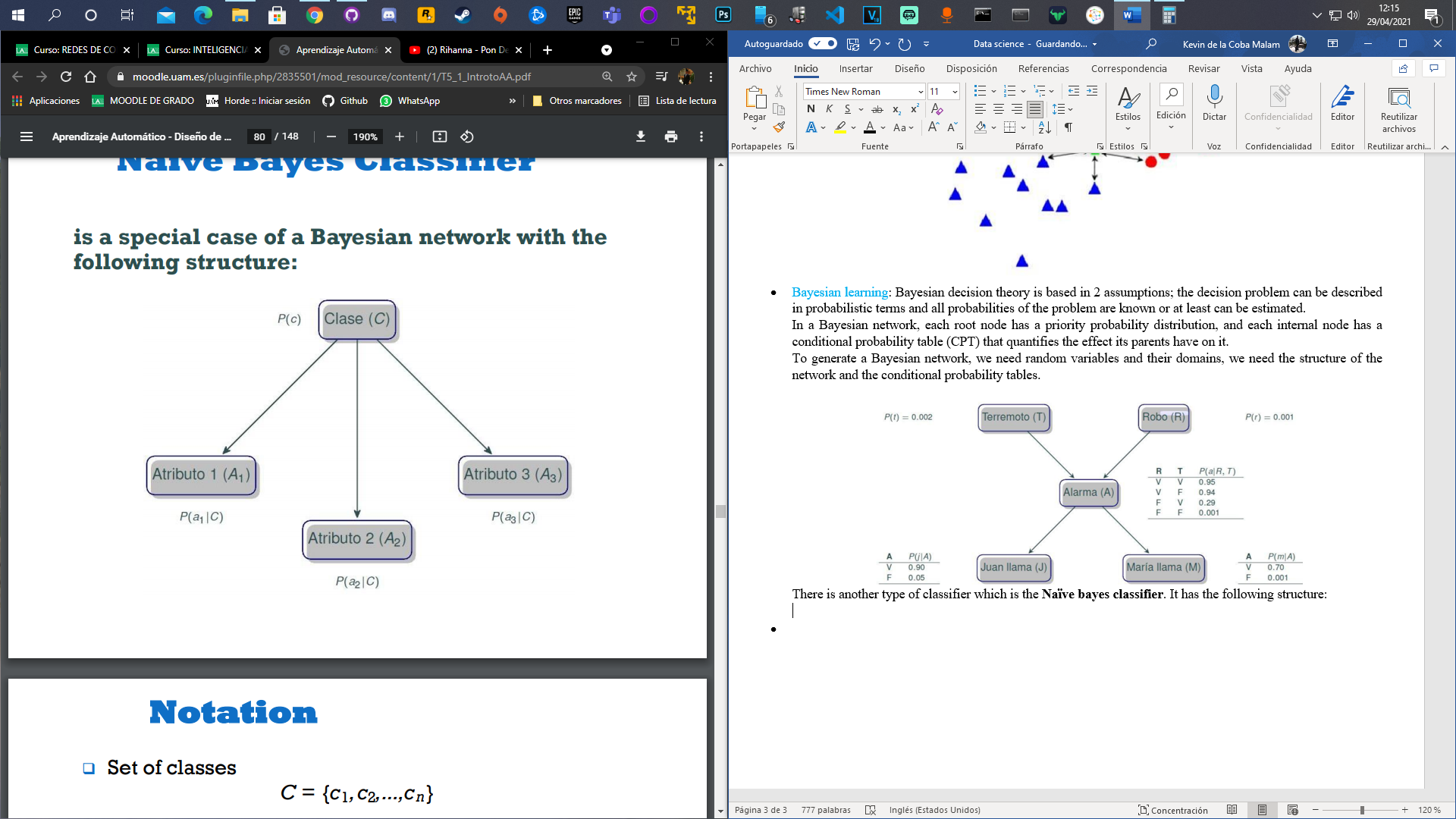
In a Bayesian network, each root node has a priority probability distribution, and each internal node has a conditional probability table (CPT) that quantifies the effect its parents have on it.

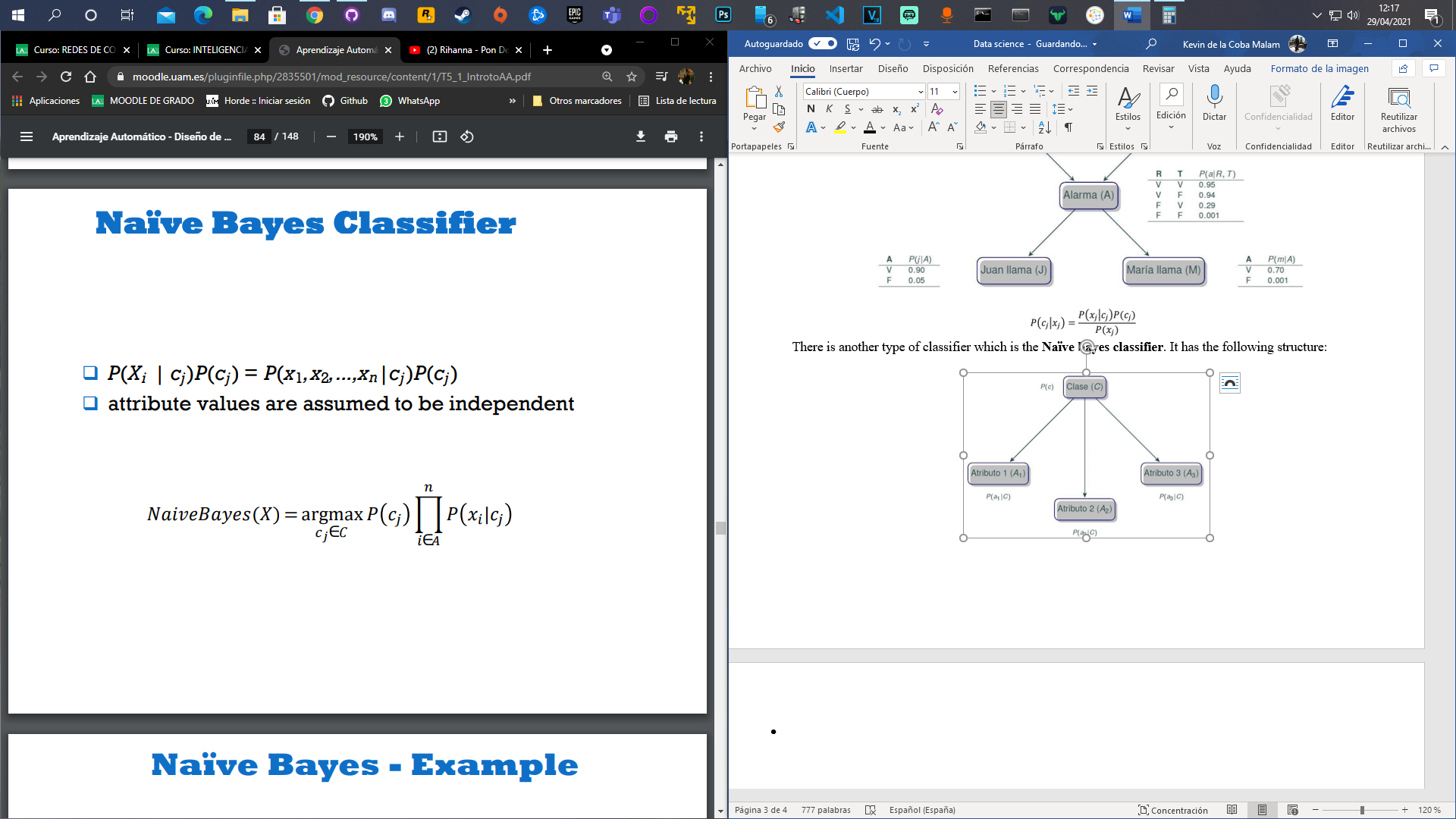
To generate a Bayesian network, we need random variables and their domains, we need the structure of the network and the conditional probability tables.





There is another type of classifier which is the **Naïve bayes classifier**. It has the following structure:





Evaluation and analysis of results.

The questions that we must do are, how good is the model we have generated for predictions of future instances? Will our predictions be better than a simple random or majority ranking?

First, we need to divide the data in a training set, a test set and production set.

The training set is used to generate the model, the test set is used for checking the predictions made of our model, and finally the production set is the one used in the final production phase.

The predictions can have several values which are in the **confusion matrix**.

