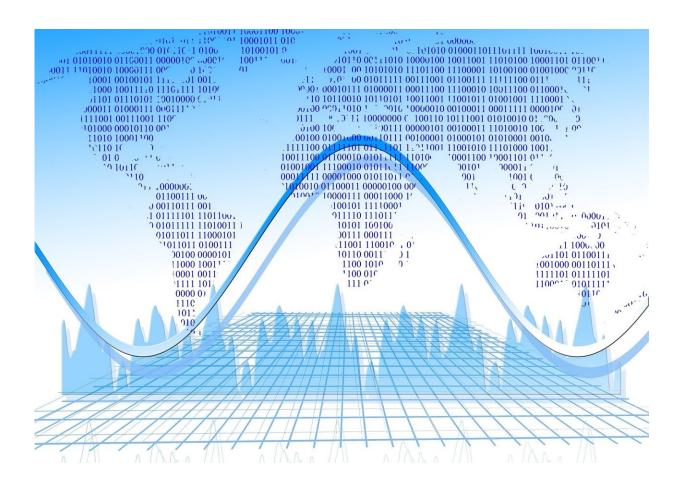# Big data used to predict the future

November 12 2018

Credit: CC0 Public Domain

Technology is moving in giant leaps and bounds, and with it, the information with which society operates daily. Nevertheless, the volume of data needs to be organized, analyzed and correlated to predict certain patterns. This is one of the main functions of what is known as Big Data.

Researchers in the KIDS research group from the University of Cordoba's Department of Computer Science and Numerical Analysis were able to improve the models that predict several variables simultaneously based on the same set of input variables, thus reducing the size of data necessary for an accurate forecast. One example of this is a method that predicts several parameters related to soil quality based on a set of variables such as crops planted, tillage and the use of pesticides.

"When you are dealing with a large volume of data, there are two solutions. You either increase computer performance, which is very expensive, or you reduce the quantity of information needed for the process to be done properly," says researcher Sebastian Ventura, one of the authors of the research article.

When building a predictive model, reliable results depend on two issues: the number of variables that come into play and the number of examples entered into the system. With the idea that less is more, the study has been able to reduce the number of examples by eliminating those that are redundant or "noisy," and that therefore do not contribute any useful information for the creation of a better predictive model.

As Oscar Reyes, the lead author of the research, points out "we have developed a technique that can tell you which set of examples you need so that the forecast is not only reliable but could even be better." In some databases, of the 18 that were analyzed, they were able to reduce the amount of information by 80 percent without affecting the predictive performance, meaning that less than half the original data was used. All of this, says Reyes, "means saving energy and money in the building of a model, as less computing power is required." In addition, it also means saving time, which is interesting for applications that work in real-time, since "it doesn't make sense for a model to take half an hour to run if you need a prediction every five minutes."

Systems that predict several related variables simultaneously, known as multi-output regression models, are gaining more notable importance due to the wide range of applications that could be analyzed under this paradigm of automatic learning, such as those related to healthcare, water quality, cooling systems for buildings and environmental studies.

Provided by University of Córdoba