MACHINE LEARNING APPLIED TO THE H INDEX OF COLOMBIAN AUTHORS WITH PUBLICATIONS IN SCOPUS

Amelec Viloria, Jenny Paola Lis Gutiérrez, Mercedes Gaitán-Angulo, Carmen Luisa Vásquez Stanescu, Tito Crissien

Abstract.

Our research aims to establish how to predict the H index of Colombian authors with publications in Scopus until 2016. The selection of the date was because, as mentioned earlier, the number of documents indexed per year exceeded 10,000 and they obtained the highest number of documents cited. To accomplish this purpose, a quantitative, nonexperimental, cross-sectional, descriptive, explanatory, and predictive research was designed using supervised learning algorithms. These were applied to information from 8,840 Colombian authors. Among the findings we can highlight that: (i) Colombia is in the fifth position in the scope of countries of South America and the Caribbean, in terms of the number of products and citations; (ii) the largest number of Colombian authors with products in Scopus until 2016, belonged mainly to the area of natural sciences, followed by medical sciences and health; (iii) most of the Colombian authors were men (64.2%, or 5,442) and they have higher H index rates than women; (iv) using random cross validation for 10 iterations, the methods with the best predictive value using R2 and the minimization of mean absolute error (MAE) correspond to: AdaBoost (96.6% and 0.397, respectively); Random Forest (96.8% and 0.431, respectively); KNN (94.4% and 0.525, respectively); Tree (94.9% and 0.53, respectively); and Neural Network (93.3% and 0.7, respectively); and (v) the variables that help predict the H index in the case of the Colombian authors, in addition to the citations, correspond to: the quantity of products, number of products in Q1, and international collaboration.

Keywords

H index, Scopus, Academic publication, Scientific research, Machine learning, Colombia