Abstract

Manar M. Hafez

Effective Selection of Machine Learning Algorithms for Big Data Analytics Using Apache Spark

Big Data appears with not only the increasing size of data but also complex and different processing and analytical tools. This research aims to compare some Selected machine learning algorithms on datasets of different types and sizes using Apache spark tool in order to make a fair judgment about which one is the best fitting in. The algorithms were compared based on few parameters including mainly accuracy and training time. The algorithms were applied on three datasets of different fields: marketing, packing and statistics, and security datasets. The findings of this experiment show that the decision tree algorithm is the most suitable algorithm for marketing and security datasets. Additionally, logistic regression algorithm had the highest accuracy for packing and statistics dataset.