MAESTRÍA EN SISTEMAS DE INFORMACIÓN, MENCIÓN EN INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

INFORME DE INVESTIGACIÓN

TÍTULO: “ANÁLISIS DEL COMPORTAMIENTO DE CLIENTES EN EMPRESAS E-COMMERCE, MENDIANTE EL DESARROLLO DE UN MODELO DISTRIBUIDO DE CLUSTERING UTILIZANDO LA PLATAFORMA DATABRICKS COMMUNITY EDITION CON PYSPARK”

Alumno: K. Irina Avalos Serrano
Tutor: Msc. J. Carlos Cisneros Cevallos

Quito, junio de 2021

ABSTRACT

El crecimiento constante del comercio electrónico ha hecho que muchas empresas se enfrenten al manejo y análisis de una gran cantidad de datos, que les permita tomar decisiones oportunas para ser más competitivas, aplicando estrategias basadas en las preferencias del cliente. Este proyecto se enfoca en la segmentación de productos de una empresa minorista de e-commerce, donde se identifican las características de los productos preferidos por los usuarios. Se ha utilizado la metodología CRISP-DM, como guía para el desarrollo del modelo, la misma que consta de seis etapas iterativas que son: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación y Despliegue. Se desarrollaron tres modelos distribuidos de segmentación, utilizando el componente MLlib de PySpark, para los algoritmos: K-means, Bisecting k-means y Gaussian Mixture. Como resultado se obtuvieron cuatro segmentos: Diamante, Oro, Plata y Bronce de acuerdo con la cantidad de likes y precio del producto.

The continued increase of e-commerce has made many companies face the management and analysis of big data, which allows them to make timely decisions to be more competitive, applying strategies based on customers preferences. This project focuses on the segmentation of products of an e-commerce retail company, for identified the best characteristics of the products preferred. CRISP-DM methodology has been used as a guide for the development of the model, which consists of six iterative steps: Understanding Business, Understanding Data, Data Preparation, Modeling, Evaluation and Deployment. Three distributed segmentation models were developed, using the MLlib component of PySpark, for the algorithms: K-means, Bisecting k-means and Gaussian Mixture. As a result, they were found four segments: Diamond, Gold, Silver and Bronze according to the number of likes and price of the product.

PALABRAS CLAVE

Aprendizaje automático, segmentación, pyspark, Databricks, comercio electrónico

Machine learning, clustering, pyspark, Databricks, e-commerce