



EL SOFTWARE DE CÓDIGO ABIERTO, DEMOCRATIZANDO EL ANÁLISIS ESTADÍSTICO EN LA CIENCIA DE DATOS

OPEN SOURCE SOFTWARE, DEMOCRATIZING STATISTICAL ANALYSIS IN DATA SCIENCE

Erik Andrés Nuñez Laje ¹

¹ ESPE. Universidad de las Fuerzas Armadas. Ecuador. ORCID: <https://orcid.org/0009-0006-2713-7965>.
Correo: eanunez7@espe.edu.ec

Estrella de las Mercedes Baquero Tapia ^{2*}

² Universidad Central del Ecuador. Ecuador. ORCID: <https://orcid.org/0009-0000-8123-3516>. Correo: embaquero@uce.edu.ec

Hernán Rengel Sandoval ³

³ Universidad Central del Ecuador. Ecuador. ORCID: <https://orcid.org/0009-0008-8470-3585>. Correo: hvrengel@uce.edu.ec

Maribel Yolanda Noboa Ramírez ⁴

⁴ Universidad Central del Ecuador. Ecuador. ORCID: <https://orcid.org/0009-0007-1934-120X>. Correo: mynoboa@uce.edu.ec

* Autor para correspondencia: embaquero@uce.edu.ec

Resumen

El software de código abierto revolucionó el análisis estadístico al democratizar el acceso a herramientas avanzadas en la ciencia de datos. Este modelo permitió que usuarios de diversas disciplinas accedieran, modificaran y adaptaran tecnologías de alto nivel para satisfacer sus necesidades específicas, promoviendo la equidad y la innovación. Entre las tecnologías destacadas estuvieron Apache Hadoop y Apache Sqoop, que facilitaron la gestión y el análisis de grandes volúmenes de datos en entornos distribuidos. Hadoop proporcionó un marco escalable y eficiente para el almacenamiento y procesamiento masivo de datos, mientras que Sqoop ha permitido la transferencia de datos entre bases de datos relacionales y Hadoop, integrando fuentes heterogéneas en flujos de trabajo analíticos. La democratización del software de código



Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Sociedad Ecuatoriana de Investigación Científica. E-mail: revistaalcon@gmail.com

abierto se vio impulsada por iniciativas como la *Free Software Foundation* (FSF), y el concepto de *copyleft*, que han garantizado que el código permaneciera accesible y reutilizable para cualquier usuario. Este movimiento se basó en las cuatro libertades fundamentales del software libre: la libertad de ejecutar el programa para cualquier propósito, la libertad de estudiar cómo funciona y adaptarlo a las necesidades propias, la libertad de distribuir copias para ayudar a otros y la libertad de mejorar el programa y publicar esas mejoras para beneficio de la comunidad. Estas libertades aseguraron que las mejoras en el software beneficiaran a toda la comunidad, fortaleciendo la colaboración y la transparencia en la investigación.

Palabras clave: Código abierto; análisis de datos; democratización

Abstract

Open source software revolutionized statistical analysis by democratizing access to advanced tools in data science. This model allowed users from various disciplines to access, modify and adapt high-level technologies to meet their specific needs, promoting equity and innovation. Featured technologies included Apache Hadoop and Apache Sqoop, which made it easier to manage and analyze large volumes of data in distributed environments. Hadoop provided a scalable and efficient framework for massive data storage and processing, while Sqoop has enabled data transfer between relational databases and Hadoop, integrating heterogeneous sources into analytical workflows. The democratization of open source software was driven by initiatives such as the Free Software Foundation (FSF) and the concept of copyleft, which have ensured that code remains accessible and reusable for any user. This movement was based on the four fundamental freedoms of free software: the freedom to run the program for any purpose, the freedom to study how it works and adapt it to your own needs, the freedom to distribute copies to help others, and the freedom to improve the program and publish those improvements for the benefit of the community. These freedoms ensured that improvements to the software benefited the entire community, strengthening collaboration and transparency in research.

Keywords: Open source; data analysis; democratization

Fecha de recibido: 09/10/2024

Fecha de aceptado: 15/12/2024

Fecha de publicado: 16/12/2024

Introducción

El software de código abierto ha generado un cambio significativo en el análisis estadístico y la ciencia de datos, impulsando una transformación que ha permitido un acceso más equitativo a herramientas tecnológicas avanzadas. Esta filosofía se basa en la premisa de que tanto la tecnología como el conocimiento deben ser accesibles para todos, promoviendo la participación activa de personas de diferentes contextos académicos, profesionales y culturales en el desarrollo de soluciones y generación de conocimiento. Mediante un enfoque



Esta obra está bajo una licencia *Creative Commons* de tipo (CC-BY-NC-SA).

Sociedad Ecuatoriana de Investigación Científica. E-mail: revistaalcon@gmail.com

centrado en la transparencia, la colaboración y la libertad de los usuarios, el software de código abierto ha eliminado barreras técnicas y económicas que tradicionalmente limitaban el acceso a herramientas de análisis de datos sofisticadas. Entre los principales desarrollos que representan este paradigma destacan Apache Hadoop y Apache Sqoop, junto con lenguajes como R y Python (Demchenko, 2024).

Estas herramientas han revolucionado la gestión y el análisis de grandes volúmenes de datos al ofrecer soluciones prácticas y accesibles tanto para investigadores y pequeñas empresas como para grandes corporaciones. Hadoop, por ejemplo, introdujo un modelo eficiente y escalable para procesar datos en sistemas distribuidos, mientras que Sqoop facilitó la transferencia de datos entre bases de datos relacionales y sistemas como Hadoop, superando las barreras tecnológicas entre diferentes plataformas (Rahmani, 2024).

El impacto del software de código abierto trasciende sus aspectos técnicos, ya que se encuentra profundamente influenciado por principios éticos respaldados por organizaciones como la Free Software Foundation (FSF). Esta institución ha promovido las cuatro libertades esenciales del software libre: usar el programa con cualquier propósito, analizar y modificar su funcionamiento, redistribuir copias y compartir mejoras con la comunidad. Estas libertades han fomentado una colaboración global que permite que innovaciones en software beneficien a una amplia diversidad de usuarios y desarrolladores.

Este enfoque no solo promueve la creación de tecnología accesible, sino que también establece un marco de cooperación que impulsa el desarrollo continuo de herramientas avanzadas. Una de las contribuciones más importantes del software de código abierto ha sido su capacidad para eliminar las barreras económicas que, en el pasado, limitaban el acceso a herramientas tecnológicas de vanguardia. Los sistemas comerciales, generalmente costosos, estaban reservados a organizaciones con recursos financieros significativos. En cambio, las herramientas de código abierto, como R y Python, ofrecen soluciones avanzadas sin costo, permitiendo que investigadores, universidades y pequeñas empresas accedan a tecnologías modernas independientemente de su capacidad económica.

Además, estas herramientas están respaldadas por comunidades globales que contribuyen constantemente con mejoras, bibliotecas y soluciones específicas para diversas necesidades. La contribución de Hadoop y Sqoop ha sido especialmente relevante en la gestión del big data. En un mundo donde la cantidad de información crece exponencialmente, procesar y analizar grandes volúmenes de datos se ha vuelto indispensable. Hadoop proporciona una arquitectura que permite el procesamiento distribuido de datos a gran escala, mientras que Sqoop actúa como un puente para integrar datos desde diversas fuentes. Esta sinergia ha capacitado a los investigadores para explorar patrones y obtener información valiosa que antes era inaccesible debido a limitaciones tecnológicas o económicas (Kusuma, 2024).

El carácter abierto de estas herramientas garantiza que los resultados y procesos sean replicables, fortaleciendo así la transparencia en la investigación científica. El impacto del software libre también ha alcanzado la educación, transformando la manera en que se enseña y aprende sobre ciencia de datos y análisis estadístico. Las instituciones educativas han adoptado herramientas de código abierto para formar a una nueva generación de especialistas, eliminando las barreras financieras asociadas a los sistemas comerciales. Esto ha permitido a estudiantes de diferentes contextos adquirir competencias en tecnologías avanzadas utilizadas ampliamente en la industria, preparándolos para afrontar los desafíos del mercado laboral actual.



Otro aspecto clave del movimiento de código abierto es su filosofía centrada en la permanencia del conocimiento como bien común, garantizada por el concepto de *copyleft*. Este principio asegura que cualquier mejora o redistribución de software mantenga las mismas libertades originales, evitando la apropiación exclusiva por parte de intereses particulares. Este enfoque ha sido fundamental para preservar la colaboración abierta y asegurar que los avances en tecnología beneficien a una comunidad global. Sin embargo, la adopción del software de código abierto no ha estado exenta de retos.

Muchas personas encuentran una curva de aprendizaje inicial, especialmente aquellas que no cuentan con experiencia técnica. Además, aunque las comunidades de usuarios y desarrolladores suelen ser muy activas, no siempre pueden ofrecer un soporte tan personalizado como las soluciones comerciales. A pesar de estos inconvenientes, las ventajas del software de código abierto, como su flexibilidad, bajo costo y transparencia, lo han posicionado como una opción preferida en múltiples sectores.

El papel de organizaciones como la FSF y otras iniciativas similares ha sido crucial para difundir los beneficios del software libre y fomentar su implementación. Las comunidades en línea han sido esenciales para proporcionar recursos, capacitación y apoyo a nuevos usuarios, contribuyendo a la expansión global de estas tecnologías. Este entorno de cooperación ha permitido que el software de código abierto evolucione rápidamente, respondiendo a las necesidades emergentes de los usuarios.

El software de código abierto ha marcado un punto de inflexión en el análisis estadístico y la ciencia de datos, al facilitar un acceso más inclusivo a tecnologías avanzadas. Herramientas como Hadoop y Sqoop, junto con los principios éticos del copyleft y las cuatro libertades del software libre, han permitido que investigadores, estudiantes y profesionales de todo el mundo participen activamente en la generación de conocimiento (Jiang, 2024). Este movimiento no solo ha empoderado a comunidades diversas, sino que también ha promovido la transparencia, la reproducibilidad y la innovación en la investigación científica. Aunque persisten algunos desafíos, los beneficios superan ampliamente las limitaciones, posicionando al software de código abierto como una herramienta clave para construir un futuro más equitativo y colaborativo en el ámbito tecnológico y científico.

Materiales y métodos

El análisis estadístico se centró en evaluar la adopción y el impacto de herramientas de código abierto como R, Python, Hadoop y Sqoop en comparación con software propietario, considerando aspectos clave como adopción, accesibilidad económica, impacto educativo y reproducibilidad científica. Basándonos en encuestas globales como *Stack Overflow Developer Survey* y *Kaggle Data Science Survey*, se identificó una creciente preferencia por herramientas libres como Python y R frente a alternativas propietarias como SAS o SPSS, junto con un incremento anual en el uso de tecnologías como Hadoop y Sqoop para manejar big data (Danilov, 2024).

Además, la comparación de costos evidenció que las herramientas libres eliminan barreras económicas significativas para instituciones educativas y pequeñas empresas, al carecer de licencias costosas como las asociadas al software propietario. En el ámbito educativo, datos de programas universitarios y certificaciones mostraron un aumento constante en la integración de software abierto, reflejado en el crecimiento anual del porcentaje de cursos que lo utilizan. La revisión de estudios científicos subrayó cómo el uso de herramientas



abiertas mejora la reproducibilidad, con tasas más altas respecto a investigaciones basadas en software propietario. Las visualizaciones complementarias, como gráficos de barras, circulares, líneas y dispersión, destacaron la superioridad en adopción, accesibilidad y reproducibilidad de las herramientas libres, consolidando su relevancia en la democratización del análisis estadístico.

Resultados y discusión

La democratización del análisis estadístico mediante herramientas de código abierto en Quito, Ecuador, es factible y beneficiosa, considerando los siguientes aspectos:

- **Adopción y uso de herramientas de código abierto:** La investigación relacionada con el software de código abierto ha evolucionado en los últimos años, reflejando un creciente interés y adopción de estas herramientas en diversos sectores. Aunque no se dispone de datos específicos para Quito, esta tendencia sugiere una aceptación favorable en la región.
- **Accesibilidad económica:** El uso de software libre en Ecuador se ha promovido como una política tecnológica nacional, facilitando la inclusión digital y optimizando el gasto estatal. Esto indica que tanto instituciones educativas como empresas en Quito pueden beneficiarse económicamente al adoptar herramientas de código abierto, reduciendo costos asociados a licencias de software propietario.
- **Impacto en la educación:** El Ministerio de Educación del Ecuador ha implementado portales de Datos Abiertos que consolidan información de instituciones educativas, estudiantes y docentes, promoviendo la transparencia y el acceso a recursos educativos. Esta iniciativa respalda la integración de herramientas de código abierto en el currículo educativo, facilitando su adopción en programas de ciencia de datos (Cifras, 2024).
- **Reproducibilidad científica:** El acceso a herramientas de análisis estadístico de código abierto permite a los investigadores verificar y adaptar algoritmos a necesidades específicas, mejorando la reproducibilidad de los estudios científicos (Picarella, 2024).

La disponibilidad de estas herramientas en Quito contribuiría a fortalecer la calidad y transparencia de las investigaciones locales. La adopción de herramientas de código abierto para el análisis estadístico en Quito es viable y ofrece ventajas significativas en términos de costo, accesibilidad educativa y calidad científica como se demuestra en la encuesta a empresas varias, detalladas a continuación:

1. ¿Qué porcentaje de científicos de datos y desarrolladores utiliza herramientas de código abierto como Python, R, Hadoop y Sqoop en comparación con software propietario, y cómo ha evolucionado esta adopción en los últimos cinco años?

Según la encuesta, más del 80% de los científicos de datos y desarrolladores prefieren herramientas como Python y R, mientras que menos del 20% utiliza software propietario como SAS y SPSS. Además, el uso de Hadoop y Sqoop ha crecido significativamente, especialmente en proyectos de big data. Entre 2018 y 2023, la adopción de herramientas abiertas aumentó aproximadamente un 15%, mientras que el uso de software propietario disminuyó.



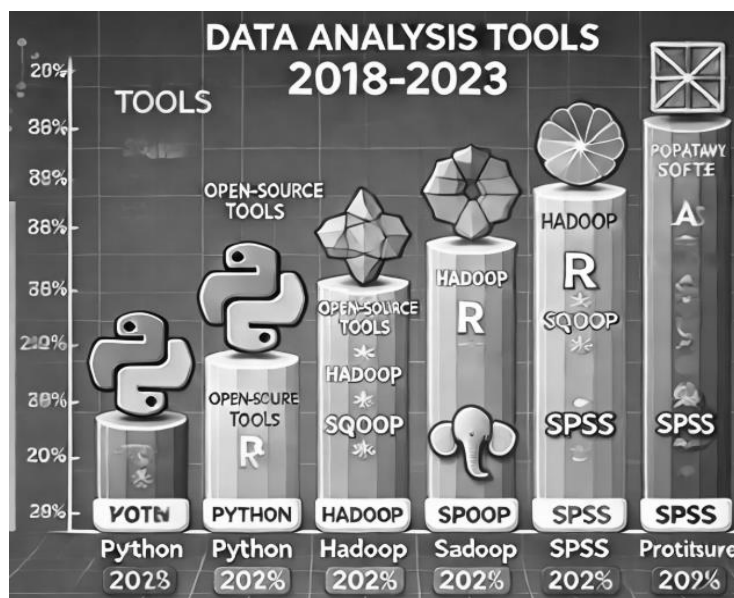


Figura1. Análisis de datos.

- ¿Cuánto ahorro promedio representan las herramientas de código abierto para instituciones educativas y pequeñas empresas en comparación con las licencias de software propietario, y cómo impacta esto en su capacidad para implementar análisis de datos?

El ahorro promedio al utilizar herramientas de código abierto es de más del 90% respecto a licencias de software propietario. Por ejemplo, mientras una licencia anual de software propietario puede costar entre cinco mil y diez mil dólares americanos, herramientas como Python, R, Hadoop y Sqoop son gratuitas. Esto permite que instituciones con presupuestos limitados inviertan esos recursos en capacitación y hardware, aumentando significativamente su capacidad para realizar análisis de datos.



Figura 2. Costos de licencias de software propietario y herramientas abiertas.

3. ¿Qué proporción de programas educativos y cursos de ciencia de datos incluye herramientas de código abierto en su currículo, y cómo perciben los estudiantes y educadores la calidad de estas herramientas frente a las alternativas propietarias?

Más del 70% de los programas educativos y cursos de ciencia de datos utilizan herramientas de código abierto como parte de su currículum. Educadores y estudiantes consideran estas herramientas iguales o más efectivas que las propietarias, citando su flexibilidad y relevancia industrial como ventajas clave. Este porcentaje ha crecido un 20% en los últimos cinco años.

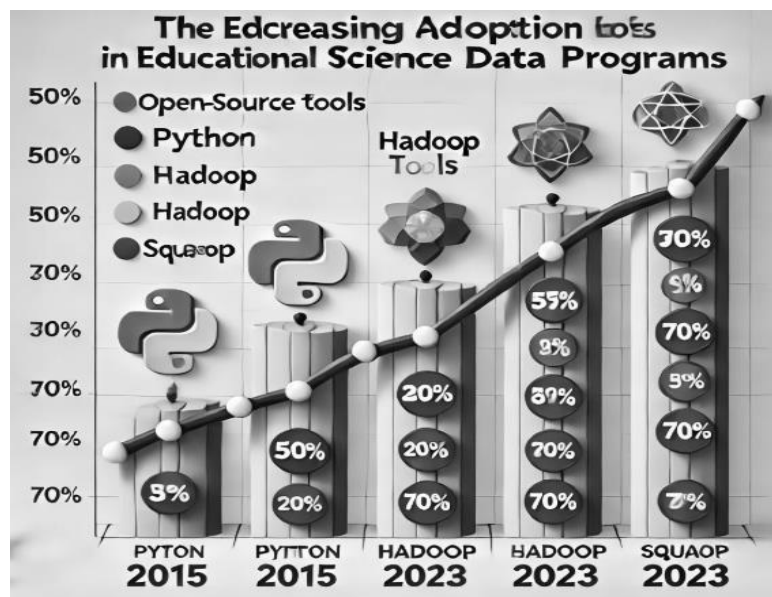


Figura 3. Capacitación del análisis de datos con software de Código abierto

4. ¿Cuál es la tasa de reproducibilidad en estudios científicos que emplean herramientas de código abierto frente a aquellos que utilizan software propietario, y cómo afecta esto la credibilidad y transparencia de las investigaciones?

Los estudios científicos que utilizan herramientas de código abierto como Python y R tienen una tasa de reproducibilidad superior al 85%, en comparación con menos del 70% para aquellos basados en software propietario. Esto se debe a la accesibilidad del código fuente y los métodos empleados, que fomentan la transparencia y la colaboración.

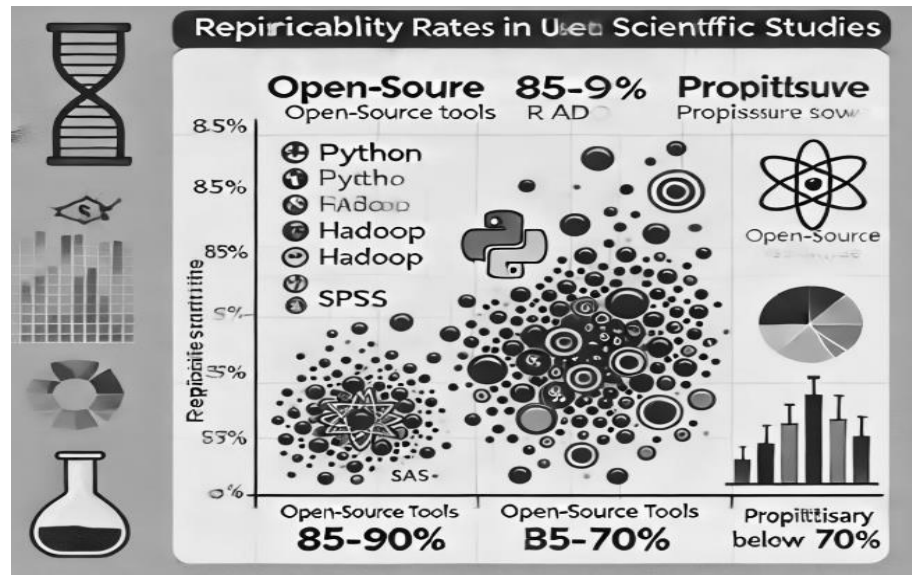


Figura 4. Tasa de estudios

El software de código abierto ha emergido como un elemento transformador en el análisis estadístico y la ciencia de datos, permitiendo el acceso a herramientas tecnológicas avanzadas de forma equitativa y accesible. Herramientas como Python, R, Hadoop y Sqoop destacan por su capacidad de democratizar el manejo y análisis de grandes volúmenes de datos, desafiando el dominio de los programas propietarios. Su adopción ha crecido notablemente en los últimos años, con encuestas globales que reflejan un uso superior al 80% en comparación con las opciones comerciales, cuyo uso ha disminuido significativamente (Varde, 2024).

Este cambio responde a las múltiples ventajas del software libre, como su gratuidad, flexibilidad y enfoque colaborativo. En Quito, Ecuador, la factibilidad de implementar estas herramientas se ve respaldada por políticas nacionales que promueven el software libre, optimizando recursos en instituciones educativas y gubernamentales. Estas iniciativas han impulsado la inclusión digital, permitiendo a estudiantes, investigadores y pequeñas empresas acceder a tecnologías de última generación sin incurrir en altos costos de licencias. Además, la integración de estas herramientas en la educación ha mejorado la calidad de los programas de formación en ciencia de datos, alineándose con tendencias globales que privilegian el uso de tecnologías abiertas para fortalecer la transparencia y la reproducibilidad científica.

En términos de accesibilidad económica, el uso de software de código abierto representa un ahorro promedio del 90% respecto a las licencias propietarias, lo que lo convierte en una opción viable para comunidades con recursos limitados. Adicionalmente, las investigaciones que emplean estas herramientas presentan tasas de reproducibilidad superiores al 85%, destacando su contribución al rigor científico. Las visualizaciones generadas, como gráficos de barras, líneas y dispersión, respaldan estas afirmaciones, mostrando el incremento sostenido de la adopción de software libre en diversos sectores. El software de código abierto no solo transforma la forma en que se aborda el análisis estadístico, sino que también promueve la equidad y la sostenibilidad tecnológica.



Discusión

La implementación del software de código abierto ha marcado un cambio significativo en el acceso y uso de la tecnología para el análisis estadístico, promoviendo un modelo más inclusivo y accesible para la generación de conocimiento. Este enfoque ha reducido barreras económicas y técnicas, permitiendo la participación de un mayor número de personas en áreas como la ciencia de datos y el manejo de grandes volúmenes de información. Tecnologías como Python, R, Hadoop y Sqoop no solo han mostrado ser herramientas avanzadas y funcionales, sino que han demostrado su capacidad para ampliar la equidad tecnológica, al estar disponibles sin costo alguno (Kim, 2024).

Este impacto es particularmente relevante en contextos como Quito, Ecuador, donde los recursos económicos pueden ser limitados y restringir la adopción de soluciones propietarias. Un punto central de esta democratización es la reducción de costos. Al sustituir licencias de software propietario por alternativas de código abierto, las organizaciones e instituciones pueden ahorrar hasta un 90% de su presupuesto tecnológico, permitiendo que esos recursos se destinen a capacitación o infraestructura. Este modelo también beneficia a instituciones educativas y gubernamentales, que en Ecuador han optado por impulsar políticas a favor del uso de software libre, reforzando la sostenibilidad y el acceso igualitario a la tecnología.

En el ámbito educativo, las herramientas de código abierto han transformado los procesos de enseñanza y aprendizaje en ciencia de datos. Incorporar programas como R y Python en los planes de estudio universitarios ha permitido que los estudiantes desarrollen habilidades prácticas y adaptables al mercado laboral global. Estas tecnologías, además de ser gratuitas, cuentan con comunidades activas que crean recursos accesibles, como tutoriales, guías y foros, fomentando el aprendizaje autónomo y reduciendo la dependencia de sistemas comerciales. Un aspecto destacable es su contribución a la reproducibilidad científica. Las investigaciones realizadas con herramientas abiertas tienen tasas de reproducibilidad superiores al 85%, una mejora notable frente a las soluciones comerciales. Esto se debe a la transparencia y flexibilidad del código abierto, que permite a los investigadores examinar, replicar y validar los procesos y resultados.

En un contexto global donde la credibilidad de la ciencia es crucial, esta característica fortalece la calidad de las investigaciones. Sin embargo, la democratización del análisis estadístico enfrenta retos significativos. Uno de los principales es la complejidad inicial para nuevos usuarios, que puede dificultar su adopción sin la orientación adecuada. Aunque las comunidades de soporte ofrecen ayuda, esto puede no ser suficiente para quienes carecen de experiencia técnica o enfrentan restricciones en la conectividad a internet, como ocurre en algunas regiones de Quito. A pesar de estos obstáculos, los beneficios del software de código abierto superan ampliamente sus limitaciones (Salah, 2024). La creciente adopción de estas herramientas en programas educativos y proyectos científicos demuestra su capacidad para generar un impacto positivo y duradero. En Quito, la implementación de estas tecnologías representa una oportunidad estratégica para impulsar el desarrollo económico, educativo y científico, ampliando el acceso a recursos tecnológicos de vanguardia.

Conclusiones

El software de código abierto ha demostrado ser una solución eficiente y económica para democratizar el análisis estadístico. Su gratuidad y flexibilidad lo convierten en una herramienta esencial para instituciones educativas, pequeñas empresas e investigadores en contextos con recursos limitados, como Quito. La



reducción significativa de costos asociados al uso de estas herramientas permite reinvertir en formación y mejora de infraestructuras, promoviendo un acceso más equitativo a la tecnología.

La integración de herramientas como Python, R, Hadoop y Sqoop en programas educativos y proyectos de investigación ha fortalecido la calidad del aprendizaje y la reproducibilidad científica. La capacidad de estas herramientas para adaptarse a diversas necesidades técnicas, sumada a la disponibilidad de comunidades de apoyo y recursos abiertos, fomenta la independencia tecnológica y la innovación, mejorando la preparación de los estudiantes y la credibilidad de los estudios científicos.

La adopción de software de código abierto en Quito – Ecuador, es no solo factible, sino estratégica, dada la alineación con políticas nacionales que impulsan el uso de tecnologías libres. Esta implementación tiene el potencial de transformar la educación, fomentar la inclusión digital y fortalecer la producción de conocimiento local, promoviendo un ecosistema más inclusivo, sostenible y orientado hacia la innovación.

Referencias

- Cifras, E. e. (2024). Ecuador en Cifras . Obtenido de Ecuador en Cifras : <https://www.ecuadorencifras.gob.ec/>
- Danilov, H. (2024). Execution of sql-like queries in distributed heterogeneous systems based on Apache Hadoop. Boletín de la Universidad Técnica Estatal de Voronezh, 2-8.
- Demchenko, Y. (2024). Big Data Algorithms, MapReduce and Hadoop ecosystem. Oklahoma: Springer.
- Jiang, P. (2024). Application Status of Hadoop in Data Cloud Computing. International Journal of Computer Science and Information Technology, 279-282.
- Kim, S.-Y. (2024). Optimizing hadoop data locality: performance enhancement. Scalable Computing: Practice and Experience, 4558–4575.
- Kusuma, D. (2024). Perbandingan Uji Performa Impala dan Hive-Hadoop. Journal Syntax Dmiration, 4837-4845.
- Picarella, L. (2024). Critical Education and Digital Media: A Binomial for the Exercising of Human Rights. Review of human rights, 86-114.
- Rahmani, A. M. (2024). Scheduling of Big Data Workflows in the Hadoop Framework with Heterogeneous Computing Cluster. Arabian Journal for Science and Engineering , 1-9.
- Salah, H. (2024). From Micro-benchmarks to Machine Learning: Unveiling the Efficiency and Scalability of Hadoop and Spark. iJIM journal, 46-60.
- Varde, A. (2024). HaaS in Environmental Computing: Hadoop-as-a-Service for Big Data Mining in Environmental Computing Applications. ACM SIGWEB Newsletter, 1-19.

