Análisis del impacto de la virtualización en el rendimiento de algoritmos de data science

Diego Andrés Clavijo Granados
Ingeniería de Sistemas
Universidad Industrial de Santander
2202019

Dilan Alessandro Corredor Diaz
Ingeniería de Sistemas
Universidad Industrial de Santander
2191976

Ronald Alexis Arias Manrique
Ingeniería de Sistemas
Universidad Industrial de Santander
2191927

***Abstract*—En este estudio, analizamos cómo la virtualización afecta el rendimiento de los algoritmos de ciencia de datos, utilizando Python para la programación. Llevamos a cabo experimentos tanto en un entorno no virtualizado como en una máquina virtual con Windows, evaluando el tiempo de ejecución.**

***Index Terms*—Virtualización, Rendimiento, Algoritmos de Ciencia de Datos, Python, Máquina Virtual, Windows.**

I. Objetivos

Objetivo General

Evaluar el impacto de la virtualización en el rendimiento de los algoritmos de ciencia de datos.

Objetivos Específicos

1) Medir el tiempo de ejecución en los entornos planteados
2) Identificar las causas de las diferencias de rendimiento entre los dos entornos.
3) Comparar el rendimiento de los algoritmos de ciencia de datos en entornos no virtualizados y en una máquina virtual con Windows.

II. Introduction

Este artículo examina cómo la virtualización afecta el rendimiento de los algoritmos de ciencia de datos. Con el creciente uso de máquinas virtuales en desarrollo y producción, es crucial comprender su impacto en aplicaciones críticas como estos algoritmos. Usamos Python para llevar a cabo los experimentos en entornos no virtualizados y en una máquina virtual con Windows, evaluando específicamente el tiempo de ejecución y el uso de recursos para ver si la virtualización introduce una sobrecarga significativa.

A. Definición de términos clave

- Virtualización: permite crear versiones virtuales de recursos de hardware, optimizando el uso de recursos y mejorando la eficiencia. Mediante la virtualización, es posible ejecutar múltiples sistemas operativos y aplicaciones en un solo servidor físico, lo que permite una mejor utilización de los recursos y una mayor flexibilidad en la gestión de sistemas y aplicaciones.
- VirtualBox: es una herramienta de virtualización de código abierto que permite la creación y gestión de máquinas virtuales. Con VirtualBox, se pueden ejecutar múltiples sistemas operativos en un solo equipo, lo que facilita la creación de entornos de prueba y desarrollo aislados. En este estudio, utilizamos una máquina virtual de Windows creada con VirtualBox para simular el entorno virtualizado.
- Algoritmo: Un algoritmo es una secuencia finita de instrucciones o pasos que se siguen para realizar una tarea específica o resolver un problema. En el contexto de la ciencia de datos y el aprendizaje automático, los algoritmos son utilizados para procesar datos, identificar patrones y realizar predicciones o clasificaciones basadas en datos de entrada.
- Linear Regression: Un modelo lineal que busca la relación lineal entre las variables de entrada y la variable objetivo.
- Decision Tree Regressor: Un modelo no lineal que divide repetidamente los datos en subconjuntos más pequeños basándose en características específicas para predecir la variable objetivo.
- Random Forest Regressor: Un conjunto de árboles de decisión que promedia múltiples modelos para mejorar la precisión y mitigar el sobreajuste
- Python: es un lenguaje de programación de alto nivel y propósito general que es ampliamente utilizado en el desarrollo de aplicaciones, análisis de datos, aprendizaje automático y más. Es conocido por su sintaxis clara y legible, así como por su amplia gama de bibliotecas y frameworks que facilitan el desarrollo rápido y eficiente de proyectos.

III. Metodología

Fase 1: Preparación y Configuración

Preparación del Entorno de Ejecución: Configuramos dos entornos en Windows: uno no virtualizado y otro virtualizado con VirtualBox. Instalamos Python y las bibliotecas necesarias (`pandas`, `matplotlib`, `scikit-learn`) en ambos entornos.

Fase 2: Implementación de los Scripts Python

Desarrollo de Scripts de Pruebas de Rendimiento:

Creamos un script llamado `test_runner.py` para ejecutar pruebas de rendimiento de los algoritmos

Identify applicable funding agency here. If none, delete this.

en ambos entornos. Este script ejecuta algoritmos de regresión sobre conjuntos de datos de diferentes tamaños y registra los tiempos de ejecución en archivos CSV (`results_no_virtualizado.csv` y `results_virtualizado.csv`).

Desarrollo del Script de Análisis y Comparación:

Implementamos un script llamado `analysis_script.py` para cargar los datos de los archivos CSV, comparar los tiempos de ejecución entre entornos y generar un gráfico de comparación. Fase 3:

Ejecución de las Pruebas

Configuración de Pruebas:

Definimos los tamaños de datos ([1000, 5000, 10000]). Configuramos los parámetros de ejecución de los algoritmos dentro del script de pruebas.

Ejecución de Pruebas:

Ejecutamos el script `test_runner.py` en ambos entornos. Capturamos y almacenamos los resultados en los archivos CSV correspondientes.

Fase 4: Análisis y Visualización

Análisis de Resultados:

Ejecutamos el script `analysis_script.py` para cargar y analizar los datos de rendimiento. Calculamos las diferencias de tiempo y los porcentajes de diferencia entre los entornos no virtualizado y virtualizado. Generamos estadísticas y resúmenes descriptivos de los resultados.

Visualización de Datos:

Utilizamos `matplotlib` para crear un gráfico que muestra la comparación de tiempos de ejecución de los algoritmos para cada tamaño de datos y entorno. Guardamos el gráfico generado como un archivo PNG para su análisis y presentación.

IV. Resultados

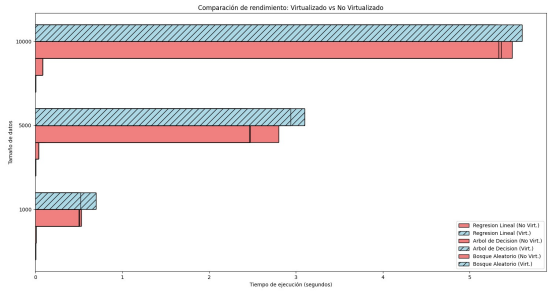

Fig. 1. Tiempo de ejecucion virtualizado vs no virtualizado

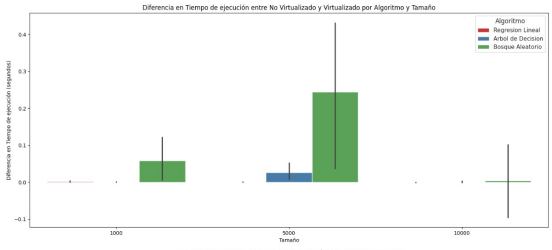

Fig. 2. delta de tiempo ejecución virtualizado vs no virtualizado

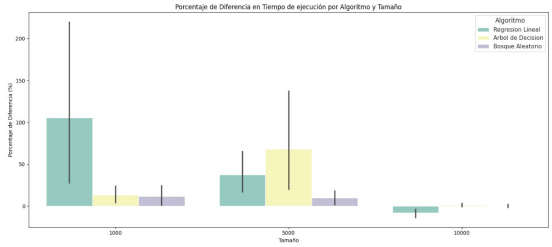

Fig. 3. Porcentaje de diferencia por algoritmo y tamaño de datos

V. Análisis de Resultados

1. Tiempo de ejecución por Algoritmo y Tamaño (No Virtualizado vs Virtualizado) En la primera gráfica, se comparan los tiempos de ejecución de diferentes algoritmos en función del tamaño de los datos tanto en entornos virtualizados como no virtualizados. Los algoritmos considerados son Regresión Lineal, Árbol de Decisión y Bosque Aleatorio.

Observaciones principales:

La Regresión Lineal muestra un rendimiento consistentemente rápido en ambos entornos, con tiempos de ejecución muy bajos independientemente del tamaño de los datos. El Árbol de Decisión presenta tiempos de ejecución ligeramente mayores que la Regresión Lineal, pero sigue siendo bastante eficiente. No obstante, se observa una ligera penalización en el tiempo de ejecución en el entorno virtualizado en comparación con el no virtualizado. El Bosque Aleatorio es el algoritmo que más se ve afectado por la virtualización. Los tiempos de ejecución son significativamente mayores en entornos virtualizados,

especialmente a medida que aumenta el tamaño de los datos.

2. Diferencia en Tiempo de ejecución entre No Virtualizado y Virtualizado por Algoritmo y Tamaño La segunda gráfica muestra la diferencia absoluta en tiempo de ejecución entre los entornos no virtualizado y virtualizado para cada algoritmo y tamaño de datos.

Observaciones principales:

La diferencia en tiempo de ejecución es más pronunciada para el Bosque Aleatorio, especialmente para los tamaños de datos más grandes (5000 y 10000). Esto indica que la virtualización tiene un impacto considerable en el rendimiento de este algoritmo. Para la Regresión Lineal y el Árbol de Decisión, la diferencia en tiempo de ejecución es relativamente menor, pero aún se puede observar que el entorno virtualizado introduce una pequeña penalización en el tiempo de ejecución.

3. Porcentaje de Diferencia en Tiempo de ejecución por Algoritmo y Tamaño La tercera gráfica presenta el porcentaje de diferencia en tiempo de ejecución entre los entornos no virtualizado y virtualizado, normalizando la diferencia absoluta respecto al tiempo de ejecución en el entorno no virtualizado.

Observaciones principales:

La Regresión Lineal tiene un impacto porcentual muy bajo debido a la virtualización, lo que sugiere que este algoritmo es muy robusto frente a los efectos de la virtualización. El Árbol de Decisión también muestra un bajo porcentaje de diferencia, aunque es mayor que el de la Regresión Lineal. El Bosque Aleatorio, sin embargo, presenta un porcentaje de diferencia considerablemente alto, lo que confirma que su rendimiento se ve significativamente afectado en entornos virtualizados.

VI. Conclusiones

Los resultados obtenidos sugieren que, mientras la Regresión Lineal y el Árbol de Decisión son menos sensibles a los efectos de la virtualización, el Bosque Aleatorio experimenta un deterioro significativo en su rendimiento en entornos virtualizados. Esto tiene implicaciones importantes para la selección de algoritmos en aplicaciones donde la eficiencia computacional es crítica, especialmente cuando se consideran entornos virtualizados.

Este estudio resalta la importancia de evaluar el rendimiento de los algoritmos en diferentes configuraciones de infraestructura para tomar decisiones informadas sobre la implementación y optimización de aplicaciones de análisis de datos y aprendizaje automático.

VII. Trabajo Futuro

Para ampliar este estudio y explorar más a fondo cómo la virtualización afecta el rendimiento de los algoritmos de aprendizaje automático, proponemos las siguientes áreas de trabajo futuro:

Mayor Disponibilidad de Hardware:

Ampliar el estudio con un acceso más amplio a hardware que permita una variedad de configuraciones en entornos no virtualizados. Esto incluiría el uso de máquinas con diferentes capacidades de procesamiento, memoria y almacenamiento.

Experimentación con Diferentes Tipos de Carga y Escalabilidad:

Evaluar cómo la virtualización afecta el rendimiento bajo diferentes niveles de demanda y escalabilidad. Analizar el comportamiento de los algoritmos en situaciones de carga variable y cómo la virtualización influye en la capacidad de respuesta en escenarios dinámicos.

Optimización de la Configuración de Virtualización:

Investigar y aplicar técnicas avanzadas para configurar la virtualización de manera que se minimice su impacto en el rendimiento de los algoritmos. Esto podría incluir ajustes en la configuración de la máquina virtual, el uso de tecnologías de contenedores o la exploración de hipervisores optimizados.

Inclusión de Otros Algoritmos y Casos de Uso:

Ampliar el estudio para incluir una gama más amplia de algoritmos de aprendizaje automático y casos de uso específicos. Explorar cómo diferentes tipos de algoritmos, como redes neuronales o algoritmos de agrupamiento, responden a la virtualización en diversos entornos y configuraciones.

Validación y Reproducibilidad:

Realizar validaciones adicionales para asegurar que los resultados sean reproducibles en diferentes entornos y con distintos conjuntos de datos. Documentar detalladamente todos los pasos y configuraciones utilizadas para facilitar la replicación del estudio por parte de otros investigadores.

References

[1] [1] VirtualBox. Oracle VM VirtualBox. Oracle, 2024. [En línea]. Disponible en: https://www.virtualbox.org

[2] [2] Python. Python Programming Language – Official Website. Python Software Foundation, 2024. [En línea]. Disponible en: https://www.python.org

[3] [3] [3] Red Hat. What is a Virtual Machine? Red Hat, 2024. [En línea]. Disponible en: https://www.redhat.com/es/topics/virtualization/what-is-a-virtual-machine

[4] [4] [4] Pandas. pandas - Python Data Analysis Library. PyData, 2024. [En línea]. Disponible en: https://pandas.pydata.org

[5] [5] NumPy. NumPy - The fundamental package for scientific computing with Python. NumPy, 2024. [En línea]. Disponible en: https://numpy.org