INTRODUCCIÓN AL MACHINE LEARNING

Diego Alejandro Bermúdez

Universidad Sergio Arboleda
Calle 74 # 14-14
Bogotá, Colombia
+57 601 3258181
diego.bermudez01@correo.usa.edu.co

Abstract: *En una era en la que la informática hace parte estructural de la toma de decisiones en diversos ámbitos del quehacer humano, resulta imperativa la optimización de los procesos computacionales que se llevan a cabo en las organizaciones. Es por lo anterior que este proyecto de investigación tiene como propósito realizar un análisis, sobre dos lenguajes de programación (C++ y Python), a través de la implementación de un modelo clásico de regresión lineal. Esto con el fin de comparar los resultados obtenidos mediante un modelo implementado mediante clases de tipo artesanal contra las clases ya estandarizadas de python.*

Keywords: *C + +, comparativa, Sklearn, modelo lineal, regresión, datos de entrenamiento, datos de prueba, normalización.*

1. Introducción:

Con el avance tecnológico de las últimas décadas se ha logrado reducir el tiempo de ejecución de los programas de cómputo hasta el punto en el que la mayoría de usuarios han dejado a un lado el hecho de tener en consideración la complejidad temporal de los mismos. Sin embargo, así como los tiempos de ejecución se reducen de manera significativa con el pasar de los años, también aumenta el volumen de datos generados de manera exponencial. Según Shah (2020), para el año 2020 se han generado 44 zettabytes de datos (equivalente a 4.4×10^{13} terabytes), de los cuales el 90% se han generado en los últimos dos años. Esto, por consiguiente, significa un enorme reto para quienes procesan y analizan grandes cantidades de información, ya que el tiempo de ejecución de un programa aumenta por lo general de manera potencial con respecto a la cantidad de datos ingresados.
Por lo tanto, resulta de gran interés realizar validaciones a los datos, tales como estandarización y normalización de los mismos. Esto con el fin de lograr los mejores resultados al momento de la estimación de valores gracias a los diferentes modelos diseñados para este fin.

Es por este motivo que la presente práctica pretende establecer un parangón entre los resultados obtenidos entre las clases y funciones estandarizadas de python referente a sus pares de C + +, estas últimas de tipo artesanal.

2. Metodología:

Resulta importante resaltar que la precisión de cada modelo será evaluada mediante métricas preestablecidas. La métrica empleada en esta ocasión será la de mínimos cuadrados ordinarios. La cual consiste en dividir por la cantidad de datos, la sumatoria del cuadrado de la diferencia entre los valores predichos con los valores reales.

Es por lo anteriormente mencionado que, para el desarrollo de esta práctica experimental se escribió, en los dos lenguajes de programación, un programa que toma una pequeña base de datos con tres variables independientes y una dependiente, esta última es el objetivo o valor a estimar.

Cabe resaltar que, debido a las características específicas para cada lenguaje de programación y las bibliotecas disponibles para cada uno de ellos, el código varía drásticamente, pues en python se dispone de librerías prediseñadas, mientras que en c + + no.

La variable establecida para efectuar la comparación de rendimiento de los modelos fue el valor obtenido de la métrica R2_score.
En cuanto a diferencias de implementación del código de los dos lenguajes, para C + + fue necesario implementar clases para la extracción, ejecución y manejo de la información. Sin embargo, para Python se utilizaron las bondades de las múltiples bibliotecas para realizar la misma tarea.

Adicionalmente para cada lenguaje se tomaron diferentes medidas frente al dataSet pues su manejo

se hacía más viable de este modo, en cada uno de los códigos fuentes se encuentra más detallado y explicado este apartado.

Acto seguido, para cada uno de los lenguajes de programación se escribió el programa que representará todo el modelo de regresión lineal según corresponda.

3. Resultados obtenidos.

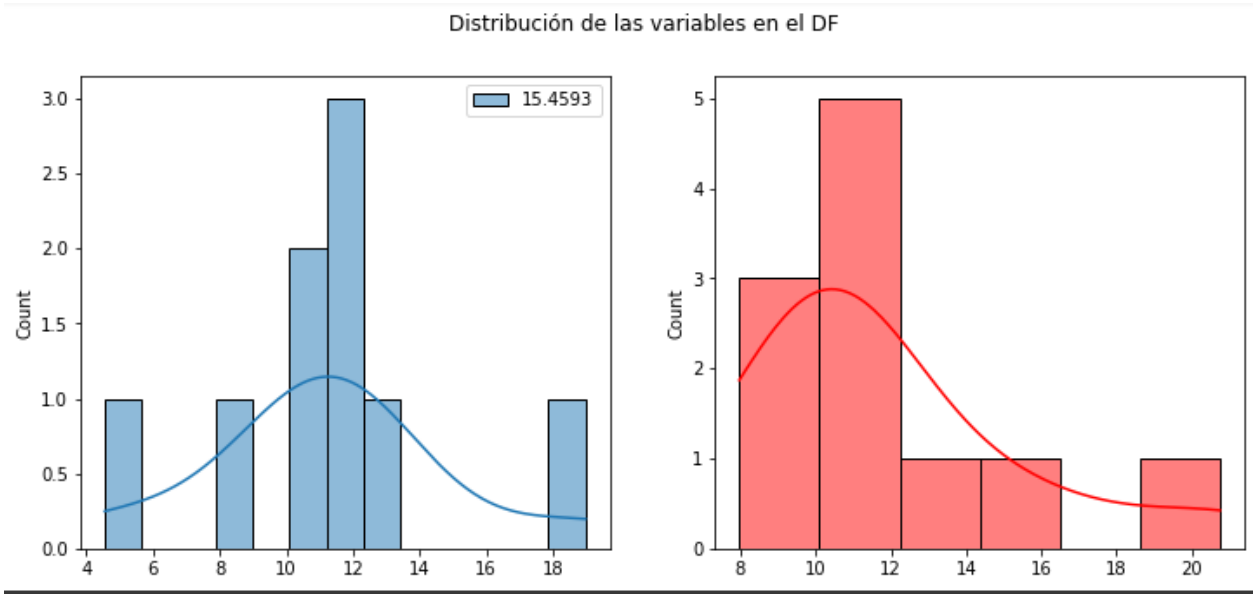

Gráfico 1:Predicciones de C++ vs Predicciones de python.

Este gráfico evidencia diferencias significativas entre los dos lenguajes de programación en cuanto a la estimación de valores.

- Discusión de los resultados obtenidos.

En la mayoría de los ítems de comparación obtenidos tanto en las gráficas como internamente en los diferentes apartados de cada código fuente se encontró una similitud constante en los resultados encontrados.
Estas discusiones más específicas se encuentran en cada parte de los códigos fuente y adicionalmente en un anexo en el cual se capturaron y analizaron datos de normalización y promedios.

- Conclusiones: del trabajo realizado:

En primer lugar, se evidencia que el lenguaje de programación en el que se escribe un programa tiene diferentes "ayudas" dependiendo los requerimientos que se tengan o los lineamientos ya preestablecidos
Yendo un poco más a fondo con lo previamente mencionado, cabe resaltar también que la complejidad detrás de cada modelo se puede mejorar y gracias a la implementación de nuevas maneras de pensar los mismos problemas podemos obtener resultados acorde a los esperados o simplemente dejar de lado este proceso de reaprendizaje y emplear los conocimientos ya disponibles a nuestra merced.

Por último se evidenció que es muy importante tener valores de referencia para validar los resultados obtenidos en el desarrollo de nuevas clases y/o funciones de origen artesanal, de aquí la importancia de tener medios como las librerías ya evaluadas un sinfín de veces (como las de python) para poder contrastar y validar nuestros modelos. En cuanto a los resultados de las predicciones de cada modelo

podemos decir que gracias a que la cantidad de datos es pequeña se evidencio una poca fiabilidad en los mismos, en cada modelo se explica el porqué de esta formación con sus respectivos argumentos.

Referencias bibliográficas:

- Shah, K. (2020, Septiembre 17). How much data is created every day in 2020? LinkedIn. Recuperado de https://www.linkedin.com/pulse/how-much-data-created-every-day-2020-kesha-shah.
-

Documentación y Anexos:

Todo el código utilizado, los resultados y la documentación con respecto a este laboratorio está disponible en el siguiente repositorio:

INTRODUCCIÓN AL ML

Biografía:

Diego Alejandro Bermúdez González

Estudiante becado de pregrado en Ingeniería en Ciencias de la Computación e Inteligencia Artificial de la Universidad Sergio Arboleda en Bogotá, Colombia**.**

Oriundo de la ciudad de Bogotá, Colombia nacido en el año 2001. Después de hacer una carrera deportiva en el fútbol-fútbol sala por más de 15 años, Diego Bermúdez decidió buscar una perfecta armonía en su área atlética e intelectual, por lo que se incursionó en un nuevo reto, uno en el campo de las Ciencias de la Computación y, más específicamente, en la ciberseguridad. Actualmente cursa cuarto semestre gracias una beca obtenida por ser parte de la selección de fútbol sala, recibida en el año 2019.