Integrantes

- Silvio Otero Guzman - C.c. 1071357464 - Ingeniería de Sistemas
- Daniela Gonzales Estrada - C.c. 1000559195 - Ingeniería Civil

Progreso alcanzado

Dos procesos fueron llevados a cabo y están distribuidos en distintos notebooks, debido a que a través de los notebooks se van creando nuevos archivos que se emplean en los siguientes notebooks, los datasets empleados fueron almacenados con el fin de que no sea necesario descargar los datos en la máquina local y subirlos al Colab, de esta forma los notebooks podrán ser reproducibles. Los distintos procesos son los que se presentan a continuación:

Simulación de datos

Con el objetivo de analizar la calidad de los datos y su influencia en la precisión de los modelos predictivos, se llevó a cabo una simulación a partir de un dataset que contenía información sobre la cantidad de especies de árboles presentes en un área de bosque. Este dataset contaba con un total de 5 mil datos y presentaba algunas columnas con información faltante y otras con datos categóricos.

Para la simulación, se procedió a eliminar de forma aleatoria la información faltante de tres columnas y la información categórica de diez columnas, con el objetivo de analizar el impacto de la falta de datos en el rendimiento de los modelos.

Los resultados obtenidos de esta simulación son importantes, ya que nos permiten conocer la influencia que tiene la calidad de los datos en los modelos predictivos. En primer lugar, se pudo observar que la eliminación de los datos faltantes tuvo un impacto significativo en la precisión de los modelos, disminuyendo la calidad de las predicciones. Por otro lado, la eliminación de la información categórica no tuvo un

impacto tan significativo en la precisión de los modelos, aunque se observó que en algunos casos sí influyó en la calidad de las predicciones.

Limpieza de datos

La limpieza de datos es un proceso esencial en la ciencia de datos, ya que puede afectar significativamente el análisis y las conclusiones que se obtengan a partir de los datos. En este caso, se trabajó en la limpieza de un dataset con el objetivo de eliminar los datos faltantes.

Para lograr esto, se utilizó un preprocesado en el cual se calculó el promedio de todos los datos en cada columna y se reemplazaron los datos faltantes por este valor. Este enfoque es comúnmente utilizado en la limpieza de datos, ya que puede ayudar a reducir la influencia de valores atípicos y mejorar la precisión de los análisis.

El dataset en cuestión se sometió a un riguroso proceso de limpieza, en el cual se identificaron y reemplazaron los datos faltantes en todas las columnas. Este proceso puede ser tedioso y requiere tiempo, pero es fundamental para garantizar la calidad de los datos y la precisión de los análisis que se realicen.

Después de la limpieza de datos, el dataset fue sometido a un análisis estadístico detallado para determinar las tendencias y patrones que se presentan en los datos. Gracias a la limpieza rigurosa que se realizó, se pudo obtener información valiosa y precisa que puede ser utilizada para tomar decisiones importantes en diversas áreas.

En conclusión, la limpieza de datos es un proceso crítico en la ciencia de datos, que puede ayudar a garantizar la calidad y la precisión de los análisis que se realicen. En este caso, se utilizó un enfoque de preprocesado para reemplazar los datos faltantes con el promedio de los datos en cada columna, lo que permitió obtener resultados valiosos y precisos a partir del dataset limpio.

Dificultades

Las dificultades presentes durante este proceso se debieron a la poca experiencia en este tipo de procesos y a la manipulación de datos experimental, al tratar de cumplir con el objetivo aún no se ha llegado a una visión ideal de la forma correcta para abordar esta solución, asunto que tendrá que ser abordado y afianzado en este periodo de tiempo futuro.