Universidad Distrital Francisco José De Caldas

Facultad De Ingeniería

Estudiantes:

Gómez Rodríguez Brayan Camilo – 20231020167

Josep Emanuel León Joya - 20231020160

Profesor:

Alberto Acosta Lopez

Asignatura:

Probabilidad y Estadística

Grupo: 020 - 84

Bogotá, D.C.

2025

Informe de Procesamiento y Filtrado de Datos - Ingeniería de Sistemas

1. Objetivo del procesamiento

El propósito principal de este procedimiento fue filtrar y preparar un subconjunto de datos del archivo original de resultados del ICFES (en formato Excel) para enfocarse únicamente en los estudiantes de programas relacionados con Ingeniería de Sistemas, Software o Computación. Esta limpieza y transformación de datos es un paso clave previo al desarrollo de modelos predictivos, como los realizados anteriormente para predecir el desempeño en las pruebas Saber Pro.

2. Descripción de las acciones realizadas

- Carga de datos: Se utilizó la biblioteca pandas para cargar el archivo 'datos_saber.xlsx', el cual contiene información detallada de los estudiantes.

- Identificación de la columna del programa académico: Se detectó automáticamente la columna que contiene el nombre del programa académico del estudiante, verificando si su nombre contenía la cadena 'ESTU_PRGM_ACADEMICO'.

```
# Primero detectamos en cuál columna esta el programa academico
columna_programa = None
for col in df.columns:
    if  "ESTU_PRGM_ACADEMICO"  in col:
        columna_programa = col
        break
```

- Filtrado por palabras clave: Se definieron palabras clave ‘Ingenieria’) que representan áreas afines a la Ingeniería. Se filtraron los registros que contenían estas palabras, independientemente de si estaban en mayúsculas o minúsculas.

```
# Lista de palabras clave a buscar
palabras_clave = ["ingenieria"]
```

- Transformación de respuestas binarias: Las respuestas tipo 'Sí'/'No' fueron transformadas a formato numérico (1 para afirmativo, 0 para negativo). Esta transformación es fundamental para realizar análisis estadísticos o alimentar algoritmos de Machine Learning, como se hizo en el documento previo donde se entrenaron modelos para predecir resultados del ICFES con algoritmos como regresión lineal y árboles de decisión.

```
# Reemplazar respuestas de sí y no por 1 y 0
df_filtrado.replace({'Sí': 1, 'Si': 1, 'No': 0, 'si': 1, 'no': 0, 'SI': 1, 'NO': 0}, inplace=True)
```

- Manejo de valores faltantes: Se reemplazaron todos los valores faltantes (NaN) por 99, actuando como un valor marcador. Aunque no es una práctica recomendada para modelado predictivo final (donde es preferible imputar con media, mediana o técnicas más complejas), esta estrategia permite seguir explorando los datos sin errores por valores nulos.

```
# Reemplazar valores faltantes con 99
df_filtrado.fillna(99, inplace=True)
```

- Exportación de los datos procesados: Finalmente, los datos filtrados y transformados se guardaron en un nuevo archivo Excel llamado 'datos_filtrados_sistemas.xlsx', que servirá como base limpia para futuros análisis descriptivos o predictivos.

```
# Guardamos el resultado
df_filtrado.to_excel("datos_filtrados_sistemas.xlsx", index=False)
```

3. Justificación del proceso

Este tipo de filtrado es necesario cuando se busca centrar el análisis en poblaciones específicas, como estudiantes de Ingeniería de Sistemas, quienes podrían tener patrones distintos de rendimiento en las pruebas Saber. Además, el preprocesamiento garantiza que los datos estén en condiciones óptimas para ser usados en herramientas de predicción, como se evidenció en el análisis anterior donde se intentó predecir el puntaje global del ICFES a partir de variables académicas, socioeconómicas y personales.

4. Conclusión

Se filtraron los registros asociados a programas de Ingeniería de Sistemas y áreas afines, lo cual representa un conjunto de datos valioso para estudios focalizados, análisis de desempeño, o la creación de modelos predictivos que puedan aportar a estrategias de mejoramiento educativo en este campo. Esta limpieza forma parte de un flujo más amplio de análisis y ciencia de datos aplicado a la educación.