# Entrega 2 Proyecto de inteligencia artificial

**Presentado por:**

Andrea Sánchez Castrillón (1001420939)
Alejandro Vargas Ocampo (1088298091)

**Profesor:**
Raul Ramos Pollan

Universidad de Antioquia
Facultad de Ingeniería
2023-2

**Entrega 2 proyecto de inteligencia artificial**

## 1. Preprocesamiento de los datos

El archivo con los datos iniciales se encuentra almacenado en el siguiente link de google drive:

https://drive.google.com/uc?id=1AFaagbKMpY07iyvOsPpo8ZPmCMV8J0Pb

En el notebook en collab se encuentra el código que accede directamente a dicho link, para que no se deban descargar los datos.

En los datos obtenidos, denominados sisben_2019.csv, se presentan alrededor de 79 variables, por lo que se estimó conveniente eliminar directamente algunas que se encuentran mejor representadas en otra variable ya existente. Dentro de las variables eliminadas se encuentran:

```
'FORMULARIO', 'SERDOMES', 'EXTRANJERO', 'TIPOESTA', 'EMBARAZA',
'CONYUVIVE', 'TRACTOR', 'ORDEN', 'CUANHORAS', 'CUANDI',  'FECHA',
          'DEPTO', 'BARRIO', 'VEREDA', 'PARED', 'PISO', 'USANITAR',
'USOSANI', 'AGUA', 'LLEGA', 'SUMINIS', 'CUANHORAS', 'PREPARAN'
```

Tras eliminar las variables mencionadas, se obtuvo un dataset con alrededor de 57 variables, incluida la variable de interés llamada PUNTAJE.

Posteriormente, se realizó el reemplazo del 5% de los datos de 3 columnas por el valor de NaN, es decir, como datos nulos. En este sentido, cada columna contenía 447872 registros, lo que en un 5% equivale a 22393 datos a eliminar de cada columna. Se eliminaron datos de las variables denominadas BUSCANDO (Semanas que lleva buscando trabajo), TSANITAR (Cantidad de sanitarios del hogar) y TPERSONA (Cantidad de personas que habitan la vivienda). A continuación se presenta la tabla donde se visualizan los valores nulos para cada variable:

```
Columna 'TCUARTOSVI': 0 valores nulos
Columna 'THOGAR': 0 valores nulos
Columna 'HOGAR': 0 valores nulos
Columna 'TENEVIV': 0 valores nulos
Columna 'TCUARTOS': 0 valores nulos
Columna 'TDORMIR': 0 valores nulos
Columna 'SANITAR': 0 valores nulos
Columna 'TSANITAR': 23377 valores nulos
Columna 'DUCHA': 0 valores nulos
Columna 'COCINA': 0 valores nulos
Columna 'COCINAN': 0 valores nulos
Columna 'ALUMBRA': 0 valores nulos
Columna 'USOTELE': 0 valores nulos
Columna 'NEVERA': 0 valores nulos
Columna 'LAVADORA': 0 valores nulos
Columna 'TVCOLOR': 0 valores nulos
Columna 'TVCABLE': 0 valores nulos
Columna 'CALENTA': 0 valores nulos
Columna 'HORNO': 0 valores nulos
```

```
Columna 'EQUIPO': 0 valores nulos
Columna 'MOTO': 0 valores nulos
Columna 'AUTO1': 0 valores nulos
Columna 'BIERAICES': 0 valores nulos
Columna 'TPERSONA': 23369 valores nulos
Columna 'TIPODOC': 0 valores nulos
Columna 'PARENTES': 0 valores nulos
Columna 'ESTCIVIL': 0 valores nulos
Columna 'HIJOSDE': 0 valores nulos
Columna 'SEXO': 0 valores nulos
Columna 'FECHANTO': 0 valores nulos
Columna 'DISCAPA': 0 valores nulos
Columna 'CARNET': 0 valores nulos
Columna 'ASISTE': 0 valores nulos
Columna 'GRADO': 0 valores nulos
Columna 'NIVEL': 0 valores nulos
Columna 'ACTIVI': 0 valores nulos
Columna 'BUSCANDO': 23349 valores nulos
Columna 'PERCIBE': 0 valores nulos
Columna 'INGRESOS': 0 valores nulos
Columna 'PAGAPOR': 0 valores nulos
Columna 'PUNTAJE': 0 valores nulos
```

Como se puede observar, las variables TSANITAR, BUSCANDO y TPERSONA contienen valores nulos equivalentes a, al menos, el 5% de los datos de cada una de ellas.

Posteriormente se procedió a calcular el promedio de los datos de las variables TSANITAR, BUSCANDO y TPERSONA y con ello se reemplazaron los valores nulos por dichos promedios. El valor de los promedios se encuentra a continuación:

```
Promedio de BUSCANDO: 0.7932545309783712
Promedio de TSANITAR: 1.0431172172044838
Promedio de TPERSONA: 5.166114647280815
```

Con los anteriores promedios se realizó imputación de los datos que contenían valores nulos.

```
Cantidad de nulos: 0
```

## 2. Correlación de Pearson

Tras llevar a cabo el preprocesamiento de los datos, se llevó a cabo la determinación de la correlación de Pearson para evidenciar qué variables tenían una mayor influencia sobre la variable de respuesta PUNTAJE. A continuación se presentan los resultados de dicha correlación, teniendo en cuenta que se eligieron las 30 variables con mayor influencia sobre la variable PUNTAJE.

| | | | |
|---|---|---|---|
| PUNTAJE | 1.000000 | INGRESOS | 0.195066 |
| ESTRATO | 0.352149 | DUCHA | 0.188893 |
| TPERSONA | 0.335194 | ACUEDUC | 0.186207 |
| LAVADORA | 0.329305 | THOGAR | 0.165095 |
| ZONA | 0.318070 | TCUARTOSVI | 0.162389 |
| TCUARTOS | 0.307183 | TIPODOC | 0.159893 |
| VIVIENDA | 0.299863 | TSANITAR | 0.158085 |
| NIVEL | 0.299811 | MOTO | 0.157553 |
| ALCANTA | 0.285275 | TVCOLOR | 0.151427 |
| TVCABLE | 0.282250 | HIJOSDE | 0.142096 |
| COMPUTADOR | 0.271075 | PERCIBE | 0.139172 |
| SANITAR | 0.270883 | AUTO1 | 0.132105 |
| NEVERA | 0.261723 | USOTELE | 0.130389 |
| GRADO | 0.257657 | ESTCIVIL | 0.127800 |
| GAS | 0.256582 | HOGAR | 0.126723 |
| TELEFONO | 0.248966 | COCINAN | 0.120431 |
| HORNO | 0.229138 | PARENTES | 0.113926 |
| EQUIPO | 0.224590 | ACTIVI | 0.079633 |
| TDORMIR | 0.211229 | DISCAPA | 0.077668 |
| COCINA | 0.209659 | BASURA | 0.053970 |
| CALENTA | 0.199708 | CARNET | 0.051045 |
| | | ELIMBASURA | 0.050493 |
| | | AIRE | 0.042774 |
| | | TENEVIV | 0.042205 |

Se filtraron los datos, formando un nuevo dataframe que contuviera las 30 variables con la correlación más alta con la variable puntaje.

### 3. Modelos de Machine Learning implementados hasta ahora.

Inicialmente se emplearon los modelos de regresión lineal, árbol de decisión y bosque aleatorio. Para tal fin, se dividieron los datos en dos conjuntos, un conjunto de entrenamiento equivalente al 75% de los datos y un conjunto de prueba equivalente al 25% de los datos. Los resultados para dichos modelos se presentan a continuación:

| | Modelo | Mean Squared Error (MSE) | R-squared (R^2) |
|---|---|---|---|
| 0 | Regresión Lineal | 113.592857 | 0.538082 |
| 1 | Árbol de Decisión | 146.713552 | 0.403399 |
| 2 | Bosque Aleatorio | 98.581483 | 0.599125 |

Regresión Lineal:

Mean Squared Error (MSE): 113.592857

R-squared (R^2): 0.538082

MSE mide el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales. Un valor de MSE más bajo indica un mejor ajuste del modelo a los datos.

R-squared (R^2) es una medida de cuánta varianza en la variable objetivo es explicada por el modelo. Un valor de R^2 más cercano a 1 indica que el modelo explica una mayor proporción de la variabilidad en los datos. En este caso, R^2 es 0.538082, lo que significa que el modelo de regresión lineal explica alrededor del 53.81% de la variabilidad en la variable PUNTAJE.

Árbol de Decisión:

Mean Squared Error (MSE): 146.713552

R-squared (R^2): 0.403399

El MSE es más alto en comparación con la Regresión Lineal, lo que indica que el modelo de Árbol de Decisión tiene un ajuste menos preciso a los datos en comparación con la Regresión Lineal.

El R^2 es más bajo (0.403399), lo que sugiere que el modelo de Árbol de Decisión explica alrededor del 40.34% de la variabilidad en la variable PUNTAJE. Esto es menor que el R^2 de la Regresión Lineal, lo que indica que el modelo de Árbol de Decisión se ajusta menos a los datos.

Bosque Aleatorio:

Mean Squared Error (MSE): 98.581483

R-squared (R^2): 0.599125

El MSE es más bajo en comparación con los dos modelos anteriores, lo que indica que el modelo de Bosque Aleatorio tiene un mejor ajuste a los datos en términos de precisión de las predicciones.

El R^2 es más alto (0.599125), lo que sugiere que el modelo de Bosque Aleatorio explica alrededor del 59.91% de la variabilidad en la variable objetivo. Esto es superior a los modelos de Regresión Lineal y Árbol de Decisión, lo que indica que el modelo de Bosque Aleatorio se ajusta mejor a los datos.

En resumen, el modelo de Bosque Aleatorio parece tener el mejor rendimiento de los tres modelos evaluados en términos de MSE y R^2. Esto significa que es capaz de hacer predicciones más precisas y explicar una mayor proporción de la variabilidad en la variable PUNTAJE en comparación con los otros dos modelos, sin embargo, resulta conveniente realizar evaluaciones con otros modelos, pues los porcentajes de variabilidad explicada por los modelos abordados hasta ahora aun resultan ser relativamente bajos.

Posteriormente, se espera evaluar los modelos Suport Vector Machines y Redes Neuronales Artificiales, para evaluar si se logra explicar una mayor variabilidad de la variable PUNTAJE en relación con las demás variables.