Leticia Castineyra Coutiño – A01173881
Aremy Díaz Cárcamo – A01552123
Edgar Díaz Ibáñez – A01552600
Carlos Morales Díaz – A01731899

# Actividad 7: regresión logística.

**Preprocesamiento de nulos y outliers.**

Los modelos se generaron utilizando la base de datos registrada en el archivo 'BD_Socio formador (TrainingDataComplete).csv'. Este cuenta desde el principio con los datos completos, por lo que no fue necesario limpiar datos nulos y con la finalidad de respetar los registros originales no se limpiaron datos atípicos.

**Conversiones a variables dicotómicas.**

Dentro de las variables categóricas se analizaron las variables ‘Profession’, ‘CITY’ y ‘STATE’, esto debido a que en las otras variables categóricas disponibles se obtienen resultados inefectivos. Para transformar estas variables en dicotómicas, se obtiene la moda de cada uno de los datos únicos existentes en sus respectivas columnas, los cuatro datos más recurrentes de las variables son seleccionados para filtrar los primeros dos datos de la variable CITY y luego los siguientes dos datos (el tercero y el cuarto) para realizar su conversión a variables dicotómicas. Este proceso también se aplicó para las variables ‘Profession’ y ‘STATE’.

En cuanto a las variables numéricas, la primer columna transformada fue ‘Age’, para su transformación se analizaron los valores únicos registrados y se utilizó la mediana como criterio de conversión a dicotómica y donde a las edades mayores a la mediana se les asignó el valor 1 y al resto el valor 0.

La segunda columna numérica transformada fue ‘Experience’, para su transformación se calculó la media de los registros siendo esta igual a 10 y se creó una nueva columna donde si la experiencia era mayor al promedio se asignaba el valor 1 y al resto el valor 0.

La tercera columna numérica transformada fue ‘Income’, para su transformación se calculó la media redondeada de los registros siendo esta igual a 4,997,117 y se creó una nueva columna donde si el ingreso era mayor al promedio se asignaba el valor 1 y al resto el valor 0.

**Modelos y evaluación.**

| No. | Dicotómica dependiente | Independientes | Precisión | Exactitud | Sensibilidad | F1 |
|---|---|---|---|---|---|---|
| 1 | Profession (Comedian, | Income,Age,Experience,CURRE | 0.5468 | 0.5441 | 0.7654 | 0.6379 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Surgeon) | NT_JOB_YRS,CURRENT_HOUSE_YRS | | | | |
| 2 | Profession (Computer_hardware_engineer, Software_Developer) | Income,Age,Experience,CURRENT_JOB_YRS,CURRENT_HOUSE_YRS | 0.5453 | 0.5458 | 0.4141 | 0.4712 |
| 3 | CITY | Income,Experience,CURRENT_JOB_YRS,CURRENT_HOUSE_YRS | 0.5927 | 0.6002 | 0.6046 | 0.5986 |
| 4 | CITY | Income,Experience,CURRENT_JOB_YRS,CURRENT_HOUSE_YRS | 0.5782 | 0.6113 | 0.6017 | 0.5897 |
| 5 | STATE | Income, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS | 0.5579 | 0.5195 | 0.0259 | 0.0495 |
| 6 | STATE | Income, Age, Experience, CURRENT_JOB_YRS, CURRENT_HOUSE_YRS | 0.4940 | 0.5249 | 0.1886 | 0.2730 |
| 7 | Age | CURRENT_HOUSE_YRS | 0.5138 | 0.5111 | 0.8155 | 0.6304 |
| 8 | Experience | Income, CURRENT_HOUSE_YRS | 0.5198 | 0.5187 | 0.9660 | 0.67587 |
| 9 | Income | Age | 0.5049 | 0.5049 | 0.4825 | 0.4934 |
| 10 | Income | Experience | 0.4928 | 0.4943 | 0.6572 | 0.5632 |

Para determinar cuál es el mejor modelo de regresión logística, se compararon las métricas presentadas en la tabla anterior. Basándose en las métricas de precisión, exactitud, sensibilidad y puntuación F1 el Modelo 3 se destaca como el mejor, ya que tiene una precisión de 0.5927, siendo la más alta, una exactitud de 0.6002, lo que indica una buena tasa de clasificación correcta, una sensibilidad de 0.6046, lo que sugiere una buena capacidad para detectar la clase positiva y la puntuación F1 es de 0.5986, lo que refleja una medida equilibrada de precisión y sensibilidad.

En resumen, el Modelo 3 supera a los demás modelos en términos de precisión, exactitud, sensibilidad y puntuación F1, lo que lo convierte en el mejor modelo de regresión logística dentro de los generados. Este modelo tiene un mejor rendimiento en la clasificación de la variable dependiente en comparación con los otros modelos.