**ANALISIS DESCRIPTIVO DEL CONJUNTO DE DATOS ADULT**

1. **Realice una descripción para contextualizar acerca del conjunto de datos tomado. (Fuente, área, dominio, tiempo en que se tomó, método de recolección).**

El conjunto de datos tomado lleva como nombre Adult, es un conjunto de datos públicamente disponible en el repositorio de datos para Machine Learning de la Universidad de California en Irvine (UCI). Fue extraído por Barry Becker de la base de datos de ingresos de los habitantes de Estados Unidos tomada por Census en 1994, se extrajo conjunto de datos con las siguientes condiciones: Edad mayor a 16 años (AAGE > 16), puntaje demográfico mayor a 1 (AFNLWGT>1) y horas a la semana mayores a 0 (HRSWK>0).

1. **Realice un análisis descriptivo básico para cada una de las variables.**

El conjunto de datos cuenta con 48,842 observaciones y 15 variables consistentes en 9 variables categóricas y 6 variables numéricas. Se realizó un filtrado del conjunto de datos eliminando las observaciones que contenían variables con valores “NA” dejando como resultado un conjunto de datos de 30162 observaciones.

La variable **workclass** describe el tipo de empleador y **ocupation** describe el tipo de empleo. La variable **education** contiene el nivel de educación más alto alcanzado. La variable **relationship** indica el tipo de relación en la que se encuentra y **marital\_status** el estado de la relación marital como puede ser casado o separado. Las demás variables categóricas son **native\_country** que indica el país de origen, **gender** que contiene el género tomando como valores masculino o femenino, **ethnicity** que indica la raza de la persona e **income** que denota si los ingresos son mayores o menores a $50.000 USD.

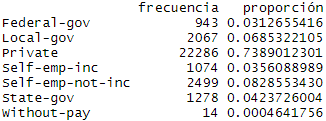
Las variables numéricas son **age** (edad), **fnlwgt** que representa un puntaje demográfico asignado a un individuo basado en información como estado de residencia y tipo de empleo. La variable **education\_num** es una representación numérica de la variable **education**. Las variables **capital\_ gain** y **capital\_loss** indican las ganancias y pérdidas en el capital. Y **hours\_per\_week** contiene la cantidad de horas trabajadas a la semana.

**Resumen de las variables numéricas:**

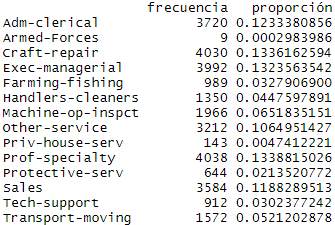
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Media | Mediana | Min | Max | Desviación  Estándar. |
| age | 38.44 | 37 | 17 | 90 | 13.13488 |
| fnlwgt | 189798 | 178429 | 13769 | 1484705 | 105652.7 |
| education\_num | 10.12 | 10 | 1 | 16 | 2.549983 |
| capital\_gain | 1092 | 0 | 0 | 99999 | 7406.467 |
| capital\_loss | 88.38 | 0 | 0 | 4356 | 404.3048 |
| hours\_per\_week | 40.93 | 40 | 1 | 99 | 11.98018 |

**Resumen de las variables categóricas:**

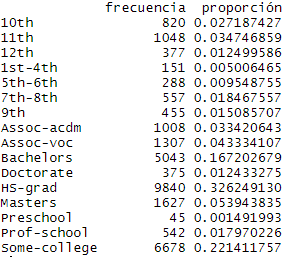
* **Workclass:**



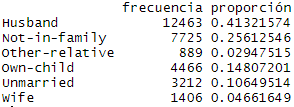
* **Ocupation:**



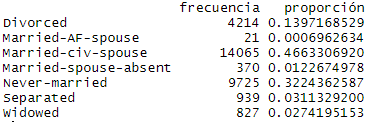
* **Education:**



* **Relationship:**

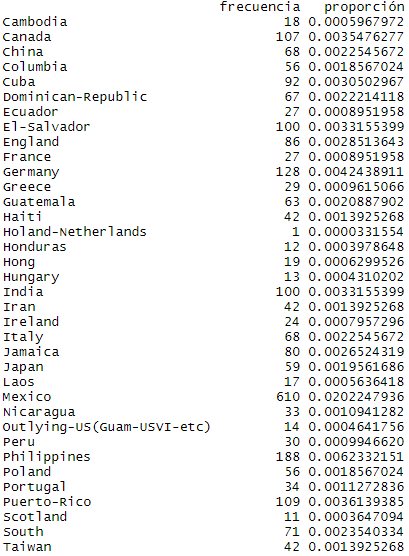


* **Marital\_status:**



Se agrupan las observaciones "Married-spouse-absent" ,“Married-AF-spouse” y “Married-civ-spouse” como “Married”.

* **Native\_country:**



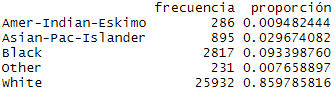


Dada la poca cantidad de las observaciones de países diferentes a “United-States” o “Mexico”, se decide agruparlos en “Otros”, quedando con 3 posibles valores en **native\_country**.

* **Gender:**



* **Ethnicity:**



Se destaca que la gran mayoría de los encuestados 85.9% son personas con etnia blanca, seguidos con los de raza negra con un 9.3% y luego con las demás etnias como minorías.

* **Income:**



**Tratamiento de los datos:**

Como se explicó anteriormente, se eliminaron las observaciones que contenían variables con valores “NA”. También se encontró una gran asociación entre las variables **relationship** y **marital\_status** al realizar un test Chi-cuadrado para comparar la significancia entre las variables categóricas.



Las hipótesis del test son las siguientes:

Del resultamos obtenemos que el valor p es mucho menor al nivel de significancia por lo que se rechaza la hipótesis nula y se concluye que las variables **relationship** y **marital\_status** están relacionadas y por tanto se excluye la variable **relationship**.

1. **Indique cual es la variable independiente y de interés para predecir.**

La variable independiente es **income** y se busca predecir si una persona gana más de $50.000 USD al año.