

**INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS
E INGENIERÍA
ENTREGA #2**



Presentado por:

David Londoño

Jesús Rodríguez

Presentado a:

Raúl Ramos Poyán

**UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA
2022**

Descripción del progreso alcanzado.

En primer lugar se realizó el preprocesado y exploración de los datos con el fin de identificar las variables que están en el encabezado del archivo, la cantidad de datos en los data frame y los datos faltantes que puedan haber en los mismos.

1. En primera instancia se carga la base de datos y se exploran los encabezados de las columnas para el dataset de prueba y el de entrenamiento. También se explora el tamaño de ambos dataset.
2. Posteriormente se exploran los datos nulos presentes en ambos dataset y se procede a eliminarlos.
3. También se realizó la identificación del tipo de variables en los dataset.
4. Posteriormente se realizó el análisis de las variables categóricas presentes en el conjunto de datos, estas variables categóricas son las posibilidades de estados que pueden tomar cada una de las columnas del conjunto de datos. Para el caso los conjuntos de variables categóricas para cada una de dichas columnas son los siguientes.

```
#Variables categóricas
ccols = [i for i in dtr.columns if not i in dtr._get_numeric_data()]
for c in ccols:
    print ("%10s"%c, np.unique(dtr[c].dropna()))

Hospital_type_code ['a' 'b' 'c' 'd' 'e' 'f' 'g']
Hospital_region_code ['X' 'Y' 'Z']
Department ['TB & Chest disease' 'anesthesia' 'gynecology' 'radiotherapy' 'surgery']
Ward_Type ['P' 'Q' 'R' 'S' 'T' 'U']
Ward_Facility_Code ['A' 'B' 'C' 'D' 'E' 'F']
Type of Admission ['Emergency' 'Trauma' 'Urgent']
Severity of Illness ['Extreme' 'Minor' 'Moderate']
Age ['0-10' '11-20' '21-30' '31-40' '41-50' '51-60' '61-70' '71-80' '81-90'
'91-100']
Stay ['0-10' '11-20' '21-30' '31-40' '41-50' '51-60' '61-70' '71-80' '81-90'
'91-100' 'More than 100 Days']
```

En la imagen anterior se puede observar que variables como el código de hospital toman un conjunto de variables entre las letras a y g, así como el código de región, el tipo de sala y el código de facilidad de sala.

También se puede observar variables como el departamento del hospital en el que se encuentra cada paciente, el tipo de admisión y la severidad de la enfermedad cuyas variables categóricas responden a nombres particulares.

Y por último se tiene un grupo de variables en rangos numéricos como la edad de los pacientes que va de 0 a 100 años, y la variable objetivo que es el tiempo de estadía en el hospital que va de 0 a más de 100 días.

5. Con el análisis anterior se puede tomar la decisión de eliminar algunas columnas que no representan gran cantidad de información o cuya información no es tan relevante para análisis posteriores del modelo, estas columnas eliminadas fueron, el código de región del hospital, el tipo de código, y el código de facilidad de las salas.
6. Conociendo la variable objetivo, que para el caso como se definió en el entregable uno es el tiempo de estadía de los pacientes en el hospital, se realizó la distribución de dicha variable.

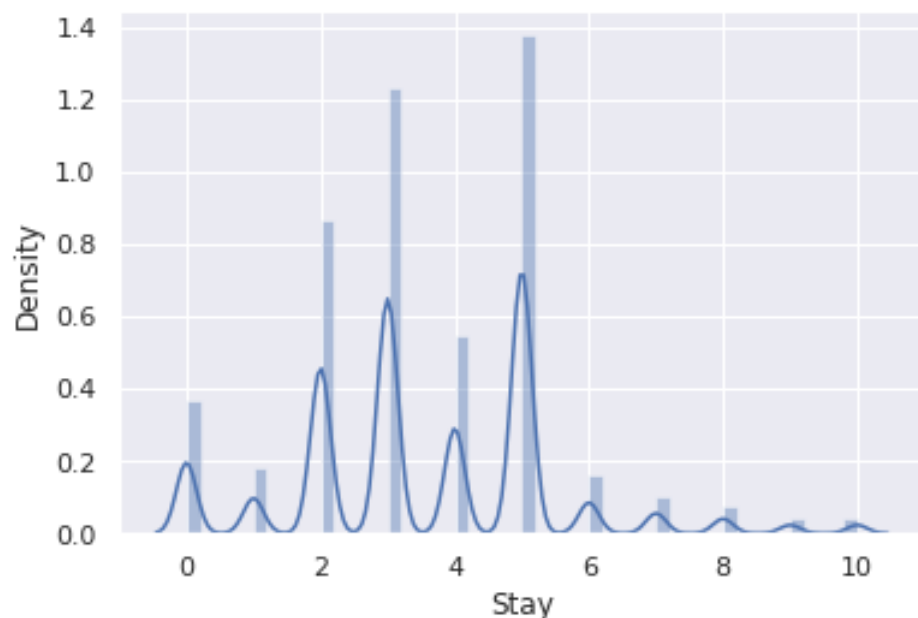


Figura 1. Distribución de la variable objetivo.

7. Por último se realizó una matriz de correlación que muestra cómo están relacionadas entre sí una a una las variables del conjunto de datos. De esta matriz se pueden ver resultados importantes, como la alta interrelación que tiene variables como la severidad de las enfermedades y el grado de cama, la complejidad del hospital y la cantidad de camas extra disponibles, y también la relación entre el tiempo de estadía en el hospital y la cantidad de visitantes con cada paciente.

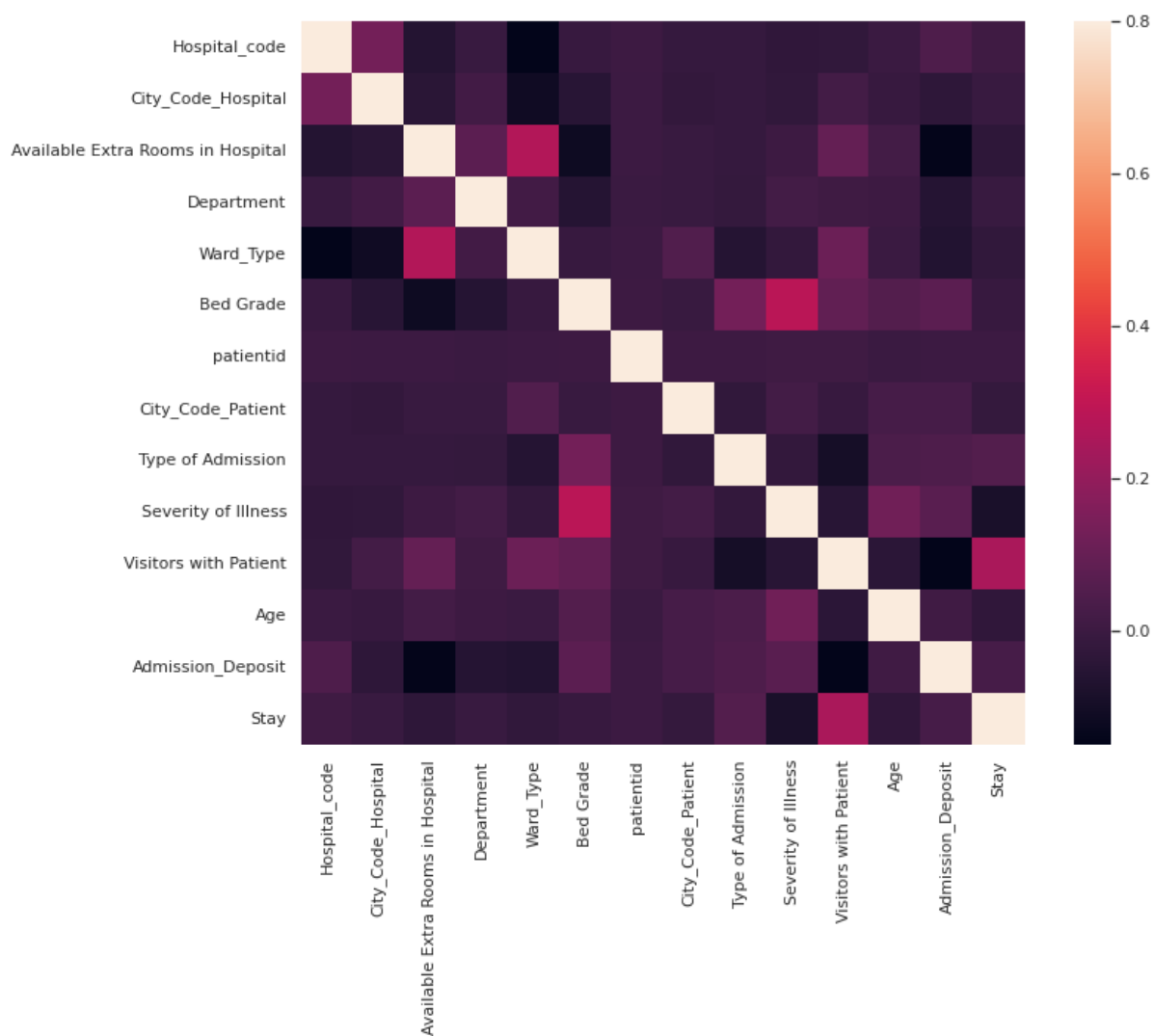


Figura 2. Matriz de correlación.