SEGUNDA ENTREGA DE PROYECTO

Predicción del tiempo de permanencia de pacientes en un hospital

POR:

Paula Andrea Gil Vargas

Kevin Manuel Jaimes Ojeda

Alejandro Rivera Pérez

PROFESOR:

Raul Ramos Pollan

Universidad de Antioquia

Facultad de Ingeniería

2023

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 dataframe de prueba y el de entrenamiento. También se explora el tamaño de ambos dataframe.
2. Posteriormente se exploran los datos nulos presentes en ambos dataframes y se procede a eliminarlos.

3. También se realizó la identificación del tipo de variables en los dataframes.

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 data_train.columns if not i in data_train._get_numeric_data()]
for c in ccols:
    print ("%10s"%c, np.unique(data_train[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.

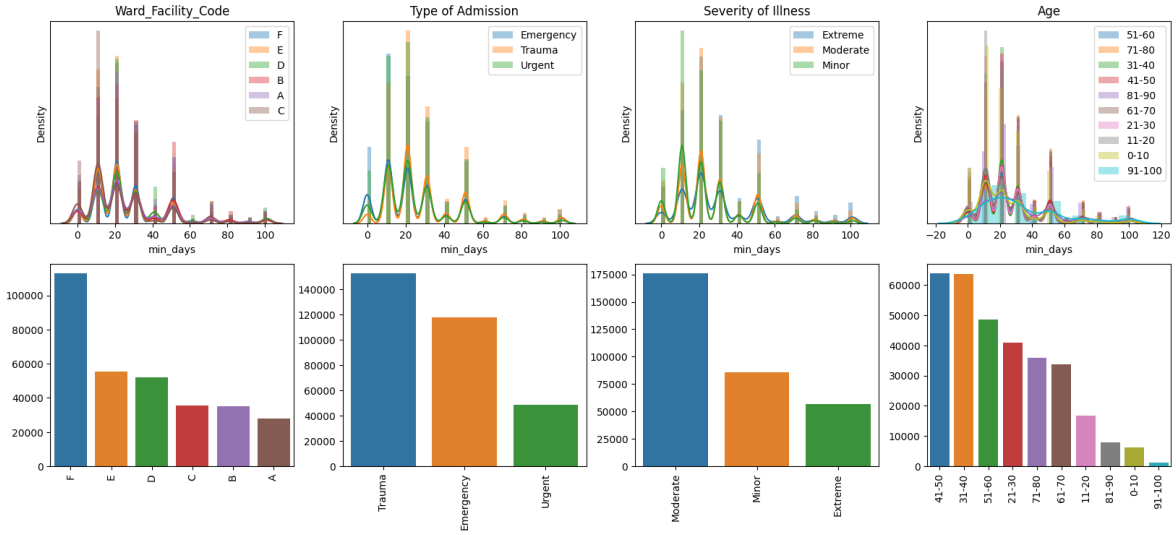
También se puede observar variables como el departamento del hospital en el que se encuentra cada paciente, el tipo de admisión y la severidad de la enfermedad cuyas variables categóricas responden a nombres particulares.

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.

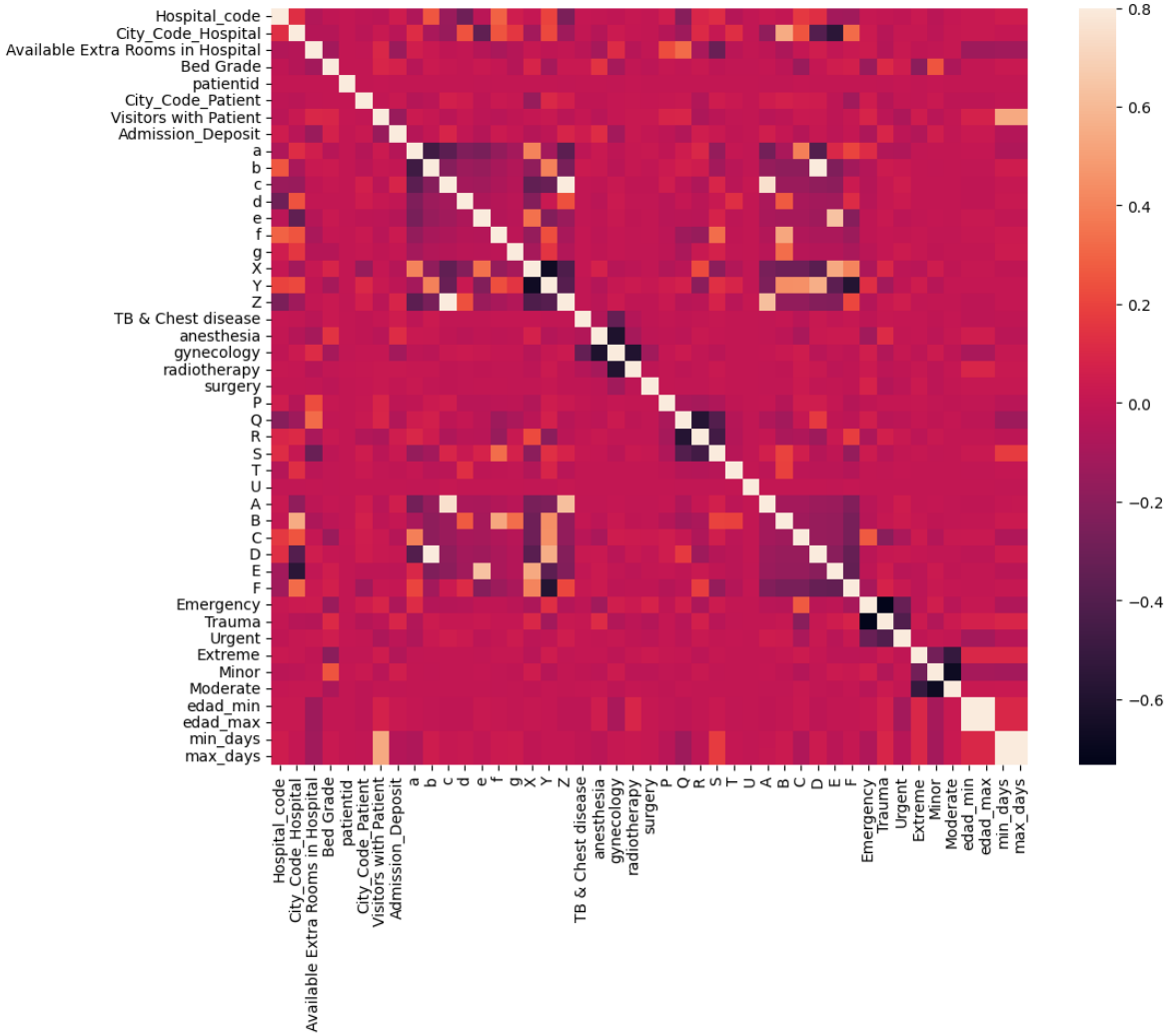
5. Con el análisis anterior se puede tomar la decisión de eliminar algunas columnas que no representan gran cantidad de información o cuya información no es tan relevante para análisis posteriores del modelo, estas columnas eliminadas fueron, el código de región del hospital, el tipo de código, y el código de facilidad de las salas.

6. También se le asignó un valor numérico a cada variable categórica.

7. 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 densidades de dicha variable.



8. 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 Visitantes con pacientes y el tiempo de estadía del paciente.


Hospital_code
City_Code_Hospital
Available Extra Rooms in Hospital
Bed Grade
patientid
City_Code_Patient
Visitors with Patient
Admission_Deposit
a
b
c
d
e
f
g
X
Y
Z
TB & Chest disease
anesthesia
gynecology
radiotherapy
surgery
P
Q
R
S
T
U
A
B
C
D
E
F
Emergency
Trauma
Urgent
Extreme
Minor
Moderate
edad_min
edad_max
min_days
max_days
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6