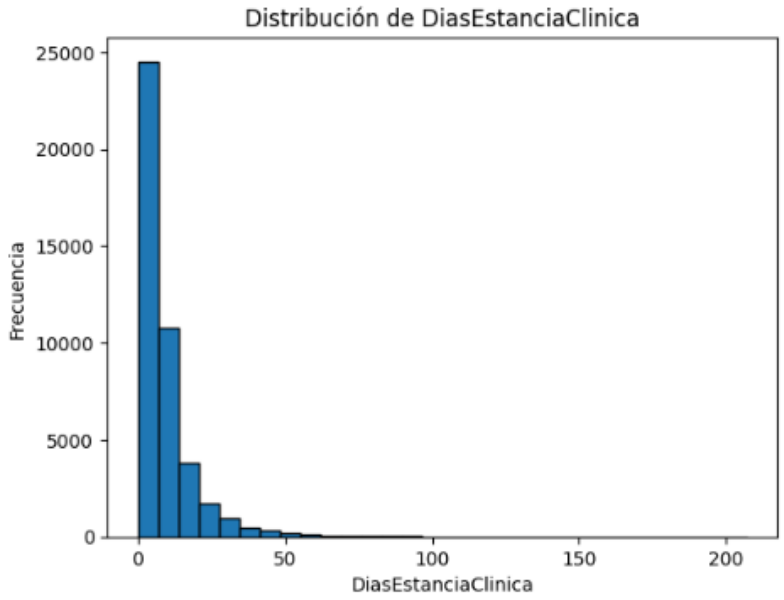
Análisis exploratorio de datos

Inicialmente se comenzó por hacer un análisis con respecto a la cantidad y tipo de datos que se tiene en el dataset, dicho procedimiento se hizo mediante la función .info a partir de esta se encontró que se cuenta con un dataset de 73 columnas y 43154 filas. En el dataset se cuentan variables numéricas y categóricas que describen datos de pacientes ingresados a determinada institución hospitalaria.

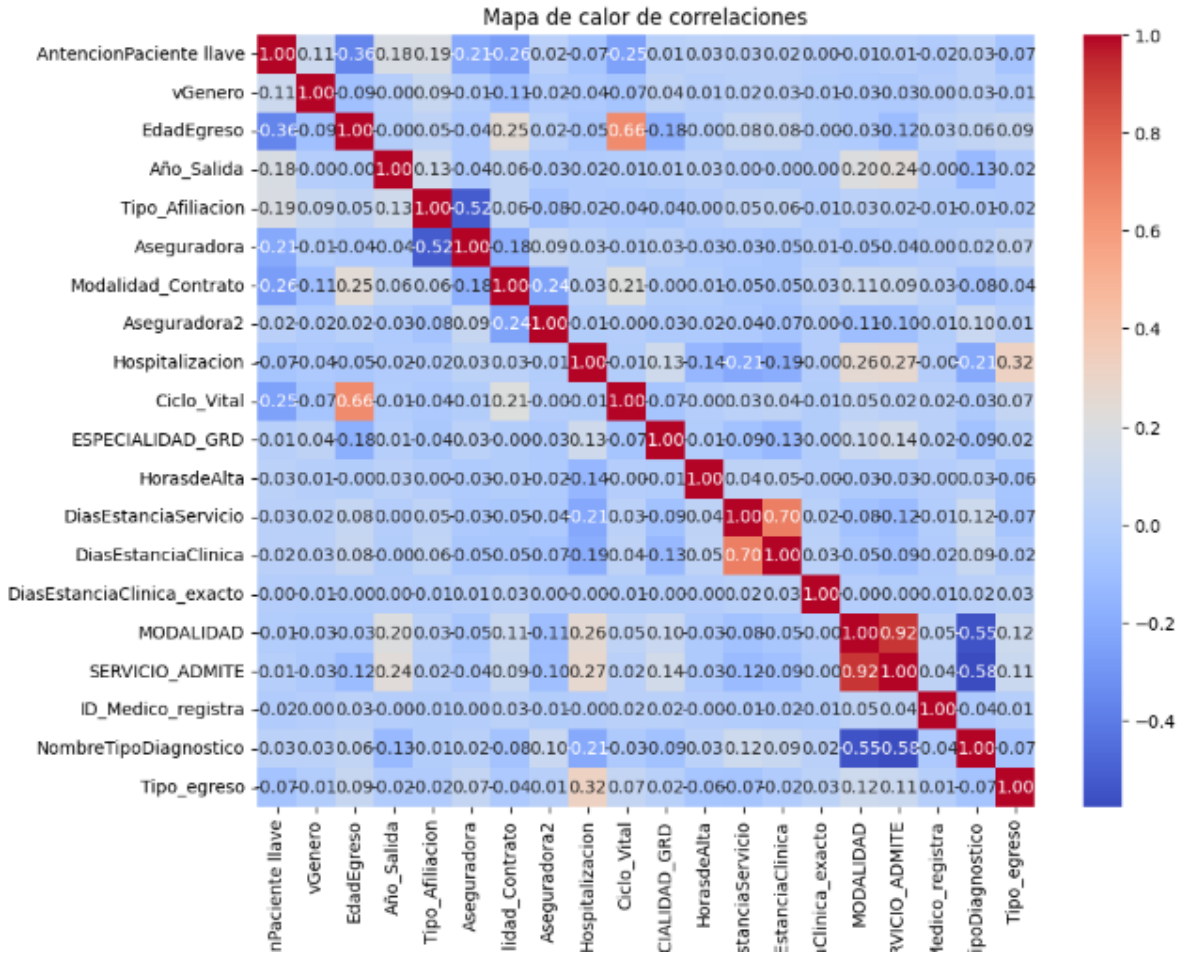
A partir del data set presentado se busca tratar de predecir el tiempo de estancia de un paciente en la unidad hospitalaria por lo que dicha variable será nuestro propósito de predicción. Por lo que se procede a realizar un diagrama de distribución además de calcular la mediana, la media, el rango, la desviación estándar y los cuartiles para asi tener una mejor comprensión del comportamiento de la variable. Se encuentran los siguientes resultados:




```
Media: 8.822449830838393
Mediana: 5.0
Rango: 207
Desviación estándar: 10.894774835119057
Cuartiles: 0.25     3.0
0.50     5.0
0.75    11.0
Name: DiasEstanciaClinica, dtype: float64
```

A partir de la información podemos determinar que la variable días de estancia de servicio cuanta con una distribución de asimetría positiva en la cual la mayor cantidad de valores se concentran en los valores más bajos y a medida que aumenta el valor de la variable disminuye su frecuencia. Además sabemos que el tiempo de estancia media es de 8.82 días aproximadamente y la mediana es 5.0 aproximadamente, dichos valores nos dan un indicativo de la magnitud que debemos aspirar encontrar una vez se apliquen modelos para predecir dicha variable.

Una vez analizada nuestra variable de salida se procede a hacer un mapa de calor con las correlaciones entre cada variable numérica entregada en el data set, obteniéndose el siguiente resultado:



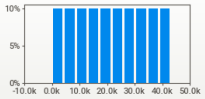
Se puede decir que dos variables tienen una correlación positiva perfecta si el coeficiente es 1 esto es que si una variable incrementa la otra también lo hace,, si el coeficiente es -1 indica que las variables tienen una correlación negativa, esto implica que cuando una incrementa la otra disminuye y viceversa. Cuando el coeficiente es 0 implica que no existe ninguna correlación entre las variables.

Lo que nos interesa principalmente es encontrar los coeficientes de correlación entre cada variable numérica con la variable días de estancia clínica para asi comenzar a descartar variables que no nos aportan información relevante a la predicción de nuestra variable de interés.

Luego, haciendo uso del software sweetviz se analizaron cada una de las variables presentadas en el data set. De esta forma se encontró cuales variables presentaban gran cantidad de datos faltantes o información poco aportante a nuestra variable de saida. Por lo que se eliminaron todas las variables con que tuvieran mas del 80% de los datos como información faltante:

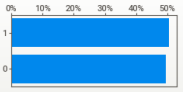
1 **AntencionPaciente llave**

VALUES:	43,154 (100%)	MAX	43,154	RANGE	43,153
MISSING:	---	95%	40,996	IQR	21,576
DISTINCT:	43,154 (100%)	Q3	32,366	STD	12,458
ZEROES:	---	MEDIAN	21,578	VAR	155.2M
		AVG	21,578	KURT.	-1.20
		Q1	10,789	SKEW	0.00
		5%	2,159	SUM	931.2M
		MIN	1		



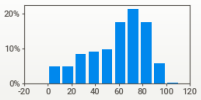
2 **vGenero**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	2 (<1%)



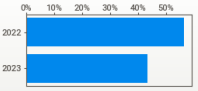
3 **EdadEgreso**

VALUES:	43,154 (100%)	MAX	111	RANGE	111
MISSING:	---	95%	89	IQR	35.0
DISTINCT:	109 (<1%)	Q3	77	STD	23.8
ZEROES:	199 (<1%)	MEDIAN	64	VAR	565
		AVG	58	KURT.	-0.441
		Q1	42	SKEW	-0.649
		5%	12	SUM	2.5M
		MIN	0		



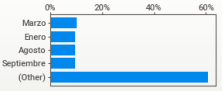
4 **Año_Salida**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	2 (<1%)



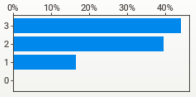
5 **MES**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	12 (<1%)



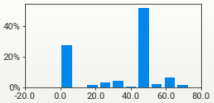
6 **Tipo_Afiliacion**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	4 (<1%)



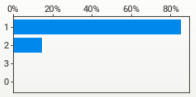
7 **Aseguradora**

VALUES:	43,154 (100%)	MAX	73.0	RANGE	73.0
MISSING:	---	95%	63.0	IQR	45.0
DISTINCT:	74 (<1%)	Q3	48.0	STD	21.4
ZEROES:	1 (<1%)	MEDIAN	48.0	VAR	458
		AVG	35.1	KURT.	-1.23
		Q1	3.0	SKEW	-0.596
		5%	3.0	SUM	1.5M
		MIN	0.0		



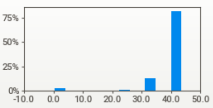
8 **Modalidad_Contrato**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	4 (<1%)



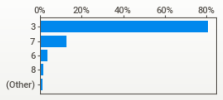
9 **Aseguradora2**

VALUES:	43,154 (100%)	MAX	44.0	RANGE	44.0
MISSING:	---	95%	43.0	IQR	0.00
DISTINCT:	45 (<1%)	Q3	43.0	STD	8.38
ZEROES:	119 (<1%)	MEDIAN	43.0	VAR	70.2
		AVG	39.9	KURT.	12.0
		Q1	43.0	SKEW	-3.41
		5%	31.0	SUM	1.7M
		MIN	0.0		



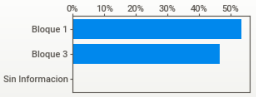
10 **Hospitalizacion**

VALUES:	43,154 (100%)
MISSING:	---
DISTINCT:	10 (<1%)



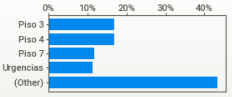
11 **bloque**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	3	(<1%)



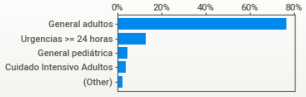
12 **piso**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	12	(<1%)



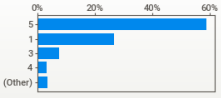
13 **Tipo_Internacion**

VALUES:	42,945 (>99%)	
MISSING:	209	(<1%)
DISTINCT:	7	(<1%)



14 **Ciclo_Vital**

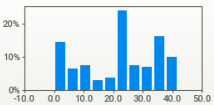
VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	6	(<1%)



15 **ESPECIALIDAD_GRD**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	43	(<1%)
ZEROES:	1	(<1%)

MAX	42.0		RANGE	42.0
95%	41.0		IQR	24.0
Q3	34.0		STD	12.1
MEDIAN	23.0		VAR	145
AVG	22.4			
Q1	10.0		KURT	-1.17
5%	4.0		SKEW	-0.194
MIN	0.0		SUM	966k



-10.0 0.0 10.0 20.0 30.0 40.0 50.0

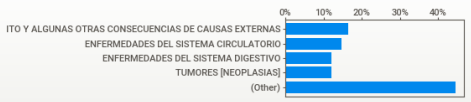
16 **DX_PRINCIPAL**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	2,850	(7%)

1,420	3%	N390
997	2%	J441
672	2%	K800
637	1%	U071
590	1%	I679
575	1%	I110
490	1%	E105
37,773	88%	(Other)

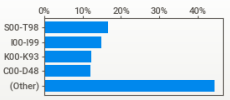
17 **D_DX_PPAL**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	22	(<1%)



18 **D_Capitulo_CIE10**

VALUES:	43,013 (>99%)	
MISSING:	141	(<1%)
DISTINCT:	21	(<1%)



19 **RangoCapitulo**

VALUES:	43,013 (>99%)	
MISSING:	141	(<1%)
DISTINCT:	957	(2%)

1,432	3%	N39
1,326	3%	K80
1,313	3%	J44
1,298	3%	E10
911	2%	N20
832	2%	J18
811	2%	I67
35,090	82%	(Other)

20 **Categoria**

VALUES:	43,154 (100%)	
MISSING:	---	
DISTINCT:	958	(2%)

1,432	3%	OTROS TRASTORNOS DEL SISTEMA URINARIO
1,326	3%	COLELITIASIS
1,313	3%	OTRAS ENFERMEDADES PULMONARES OBSTRUCTIVAS CRONICAS
1,298	3%	DIABETES MELLITUS INSULINODEPENDIENTE
911	2%	CALCULO DEL RINON Y DEL URETER
832	2%	NEUMONIA, ORGANISMO NO ESPECIFICADO
811	2%	OTRAS ENFERMEDADES CEREBROVASCULARES
35,231	82%	(Other)

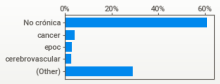
26 **D_DX_MEDICO**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 3,127 (7%)

1,340	3%	ATENCION PALIATIVA
1,205	3%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
872	2%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON EXACERBACION AGUDA NO ESPECIFICADA
822	2%	INSUFICIENCIA CARDIACA CONGESTIVA
743	2%	OTROS DOLORES ABDOMINALES Y LOS NO ESPECIFICADOS
653	2%	HEMORRAGIA GASTROINTESTINAL NO ESPECIFICADA
639	1%	ENFERMEDAD CEREBROVASCULAR NO ESPECIFICADA
36,889	85%	(Other)

27 **ClasificacionCronico_DX_Medico**

VALUES: 43,062 (>99%)
MISSING: 92 (<1%)
DISTINCT: 61 (<1%)



28 **DX_R1**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,836 (7%)

3,166	7%	I10X
723	2%	-1
716	2%	J22X
640	1%	J440
561	1%	I120
554	1%	N390
541	1%	I110
36,253	84%	(Other)

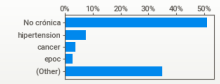
29 **D_DX_R1**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,820 (7%)

3,166	7%	HIPERTENSION ESENCIAL (PRIMARIA)
723	2%	Sin informacion en base de datos
716	2%	INFECCION AGUDA NO ESPECIFICADA DE LAS VIAS RESPIRATORIAS INFERIORES
640	1%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON INFECCION AGUDA DE LAS VIAS RESPIRAT...
561	1%	ENFERMEDAD RENAL HIPERTENSIVA CON INSUFICIENCIA RENAL
554	1%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
541	1%	ENFERMEDAD CARDIACA HIPERTENSIVA CON INSUFICIENCIA CARDIACA (CONGESTIVA)
36,253	84%	(Other)

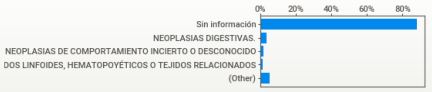
30 **ClasificacionCronico_DX_R1**

VALUES: 43,023 (>99%)
MISSING: 131 (<1%)
DISTINCT: 60 (<1%)



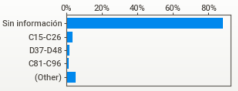
21 **D_subcapitulo_CIE10**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 19 (<1%)



22 **DescSubgrupoOncologia**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 19 (<1%)



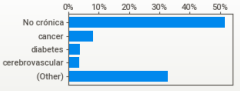
23 **SubgrupoOncologia**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,849 (7%)

1,420	3%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
997	2%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON EXACERBACION AGUDA NO ESPECIFICADA
672	2%	CALCULO DE LA VESICULA BILIAR CON COLECISTITIS AGUDA
637	1%	COVID-19 VIRUS IDENTIFICADO
590	1%	ENFERMEDAD CEREBROVASCULAR NO ESPECIFICADA
575	1%	ENFERMEDAD CARDIACA HIPERTENSIVA CON INSUFICIENCIA CARDIACA (CONGESTIVA)
490	1%	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLICACIONES CIRCULATORIAS PERIFERICAS
37,773	88%	(Other)

24 **ClasificacionCronico_DX_Principal**

VALUES: 43,042 (>99%)
MISSING: 112 (<1%)
DISTINCT: 57 (<1%)



25 **DX_MEDICO**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 3,129 (7%)

1,340	3%	Z515
1,205	3%	N390
872	2%	J441
822	2%	I500
743	2%	R104
653	2%	K922
639	1%	I679
36,889	85%	(Other)

31 **DX_R2**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,626 (6%)

4,120	10%	I10X
3,375	8%	-1
983	2%	I500
816	2%	I120
727	2%	J22X
494	1%	I110
446	1%	J960
32,193	75%	(Other)

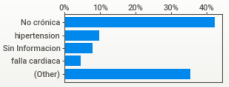
32 **D_DX_R2**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,623 (6%)

4,120	10%	HIPERTENSION ESENCIAL (PRIMARIA)
3,375	8%	Sin informacion en base de datos
983	2%	INSUFICIENCIA CARDIACA CONGESTIVA
816	2%	ENFERMEDAD RENAL HIPERTENSIVA CON INSUFICIENCIA RENAL
727	2%	INFECCION AGUDA NO ESPECIFICADA DE LAS VIAS RESPIRATORIAS INFERIORES
494	1%	ENFERMEDAD CARDIACA HIPERTENSIVA CON INSUFICIENCIA CARDIACA (CONGESTIVA)
446	1%	INSUFICIENCIA RESPIRATORIA AGUDA
32,193	75%	(Other)

33 **ClasificacionCronico_DX_R2**

VALUES: 43,064 (>99%)
MISSING: 90 (<1%)
DISTINCT: 61 (<1%)



34 **DX_R3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,371 (5%)

6,905	16%	-1
4,006	9%	I10X
912	2%	I500
706	2%	I120
544	1%	E756
514	1%	E119
410	<1%	I110
29,157	68%	(Other)

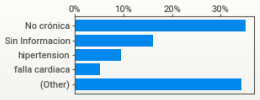
35 **D_DX_R3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,368 (5%)

6,905	16%	Sin informacion en base de datos
4,006	9%	HIPERTENSION ESENCIAL (PRIMARIA)
912	2%	INSUFICIENCIA CARDIACA CONGESTIVA
706	2%	ENFERMEDAD RENAL HIPERTENSIVA CON INSUFICIENCIA RENAL
544	1%	TRASTORNO DEL ALMACENAMIENTO DE LIPIDOS NO ESPECIFICADO
514	1%	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION
410	<1%	ENFERMEDAD CARDIACA HIPERTENSIVA CON INSUFICIENCIA CARDIACA (CONGESTIVA)
29,157	68%	(Other)

36 **ClasificacionCronico_DX_R3**

VALUES: 43,087 (>99%)
MISSING: 67 (<1%)
DISTINCT: 61 (<1%)



37 **DX_F3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,371 (5%)

11,661	27%	-1
1,616	4%	R104
906	2%	N390
834	2%	J441
815	2%	R060
672	2%	K922
602	1%	I679
26,048	60%	(Other)

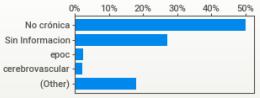
38 **D_DX_F3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 2,371 (5%)

11,661	27%	Sin informacion en base de datos
1,616	4%	OTROS DOLORES ABDOMINALES Y LOS NO ESPECIFICADOS
906	2%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
834	2%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON EXACERBACION AGUDA NO ESPECIFICADA
815	2%	DISNEA
672	2%	HEMORRAGIA GASTROINTESTINAL NO ESPECIFICADA
602	1%	ENFERMEDAD CEREBROVASCULAR NO ESPECIFICADA
26,048	60%	(Other)

39 **ClasificacionCronico_DX_F3**

VALUES: 43,097 (>99%)
MISSING: 57 (<1%)
DISTINCT: 56 (<1%)



40 **DX_F3_R1**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,012 (2%)

38,250	89%	-1
198	<1%	I10X
180	<1%	N390
127	<1%	R104
114	<1%	I500
101	<1%	J441
100	<1%	R509
4,084	9%	(Other)

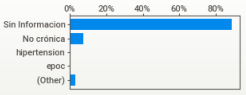
41 **D_DX_F3_R1**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,012 (2%)

38,250	89%	Sin informacion en base de datos
198	<1%	HIPERTENSION ESENCIAL (PRIMARIA)
180	<1%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
127	<1%	OTROS DOLORES ABDOMINALES Y LOS NO ESPECIFICADOS
114	<1%	INSUFICIENCIA CARDIACA CONGESTIVA
101	<1%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON EXACERBACION AGUDA NO ESPECIFICADA
100	<1%	FIEBRE NO ESPECIFICADA
4,084	9%	(Other)

42 **ClasificacionCronico_DX_F3_R1**

VALUES: 43,147 (>99%)
MISSING: 7 (<1%)
DISTINCT: 54 (<1%)



43 **DX_F3_R2**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 470 (1%)

41,843	97%	-1
83	<1%	I10X
56	<1%	R509
30	<1%	Z038
27	<1%	N390
23	<1%	R104
22	<1%	I500
1,070	2%	(Other)

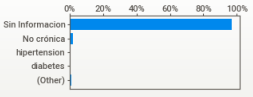
44 **D_DX_F3_R2**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 470 (1%)

41,843	97%	Sin informacion en base de datos
83	<1%	HIPERTENSION ESENCIAL (PRIMARIA)
56	<1%	FIEBRE NO ESPECIFICADA
30	<1%	OBSERVACION POR SOSPECHA DE OTRAS ENFERMEDADES Y AFECCIONES
27	<1%	INFECCION DE VIAS URINARIAS SITIO NO ESPECIFICADO
23	<1%	OTROS DOLORES ABDOMINALES Y LOS NO ESPECIFICADOS
22	<1%	INSUFICIENCIA CARDIACA CONGESTIVA
1,070	2%	(Other)

45 **ClasificacionCronico_DX_F3_R2**

VALUES: 43,153 (>99%)
MISSING: 1 (<1%)
DISTINCT: 45 (<1%)



46 **DX_F3_R3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 401 (<1%)

42,242	98%	-1
96	<1%	Z038
48	<1%	I10X
19	<1%	R060
15	<1%	R509
15	<1%	R104
12	<1%	I500
707	2%	(Other)

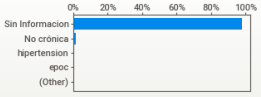
47 **D_DX_F3_R3**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 401 (<1%)

42,242	98%	Sin informacion en base de datos
96	<1%	OBSERVACION POR SOSPECHA DE OTRAS ENFERMEDADES Y AFECCIONES
48	<1%	HIPERTENSION ESENCIAL (PRIMARIA)
19	<1%	DISNEA
15	<1%	FIEBRE NO ESPECIFICADA
15	<1%	OTROS DOLORES ABDOMINALES Y LOS NO ESPECIFICADOS
12	<1%	INSUFICIENCIA CARDIACA CONGESTIVA
707	2%	(Other)

48 **ClasificacionCronico_DX_F3_R3**

VALUES: 43,152 (>99%)
MISSING: 2 (<1%)
DISTINCT: 44 (<1%)



49 **DX_pre_cx**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,461 (3%)

29,123	67%	-1
482	1%	N201
449	1%	K800
295	<1%	S721
247	<1%	N40X
239	<1%	I702
204	<1%	N200
12,115	28%	(Other)

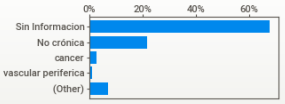
50 **D_DX_pre_cx**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,461 (3%)

29,123	67%	Sin informacion en base de datos
482	1%	CALCULO DEL URETER
449	1%	CALCULO DE LA VESICULA BILIAR CON COLECISTITIS AGUDA
295	<1%	FRACTURA PERTROCANTERIANA
247	<1%	HIPERPLASIA DE LA PROSTATA
239	<1%	ATEROSCLEROSIS DE LAS ARTERIAS DE LOS MIEMBROS
204	<1%	CALCULO DEL RINON
12,115	28%	(Other)

51 **ClasificacionCronico_DX_pre_cx**

VALUES: 43,123 (>99%)
MISSING: 31 (<1%)
DISTINCT: 44 (<1%)



52 **DX_pos_cx**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,492 (3%)

29,123	67%	-1
442	1%	N201
411	<1%	K800
298	<1%	S721
252	<1%	N40X
241	<1%	I702
226	<1%	N200
12,161	28%	(Other)

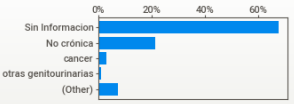
53 **D_DX_pos_cx**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 1,492 (3%)

29,123	67%	Sin informacion en base de datos
442	1%	CALCULO DEL URETER
411	<1%	CALCULO DE LA VESICULA BILIAR CON COLECISTITIS AGUDA
298	<1%	FRACTURA PERTROCANTERIANA
252	<1%	HIPERPLASIA DE LA PROSTATA
241	<1%	ATEROSCLEROSIS DE LAS ARTERIAS DE LOS MIEMBROS
226	<1%	CALCULO DEL RINON
12,161	28%	(Other)

54 **ClasificacionCronico_DX_pos_cx**

VALUES: 43,126 (>99%)
MISSING: 28 (<1%)
DISTINCT: 43 (<1%)



55 **DX_MUERTE**

VALUES: 43,154 (100%)
MISSING: ---
DISTINCT: 576 (1%)

39,736	92%	-1
303	<1%	J440
192	<1%	U071
174	<1%	I132
129	<1%	I120
128	<1%	C349
104	<1%	I219
2,388	6%	(Other)

56 **D_DX_MUERTE**

VALUES:	43,154 (100%)	39,736	92%	Sin informacion en base de datos
MISSING:	---	303	<1%	ENFERMEDAD PULMONAR OBSTRUCTIVA CRONICA CON INFECCION AGUDA DE LAS VIAS RESPIRAT...
		192	<1%	COVID-19 VIRUS IDENTIFICADO
DISTINCT:	575 (1%)	174	<1%	ENFERMEDAD CARDIORRENAL HIPERTENSIVA CON INSUFICIENCIA CARDIACA Y RENAL (CONGESTIV...
		129	<1%	ENFERMEDAD RENAL HIPERTENSIVA CON INSUFICIENCIA RENAL
		128	<1%	TUMOR MALIGNO DE LOS BRONQUIOS O DEL PULMON PARTE NO ESPECIFICADA
		104	<1%	INFARTO AGUDO DEL MIOCARDIO SIN OTRA ESPECIFICACION
		2,388	6%	(Other)

57 **D_Subcapitulo_DX_Muerte**

VALUES:	43,154 (100%)	39,736	92%	Sin informacion en base de datos
MISSING:	---	306	<1%	OTRAS ENFERMEDADES PULMONARES OBSTRUCTIVAS CRONICAS
		194	<1%	USO DE EMERGENCIA DE U07
DISTINCT:	296 (<1%)	181	<1%	ENFERMEDAD CARDIORRENAL HIPERTENSIVA
		146	<1%	TUMOR MALIGNO DE LOS BRONQUIOS Y DEL PULMON
		129	<1%	ENFERMEDAD RENAL HIPERTENSIVA
		118	<1%	INFARTO AGUDO DEL MIOCARDIO
		2,344	5%	(Other)

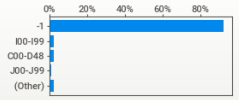
58 **Subcapitulo_DX_Muerte**

VALUES:	43,144 (>99%)	39,736	92%	-1
MISSING:	10 (<1%)	306	<1%	J44
		194	<1%	U07
DISTINCT:	296 (<1%)	181	<1%	I13
		146	<1%	C34
		129	<1%	I12
		118	<1%	I21
		2,334	5%	(Other)

59 **capitulo_DX_Muerte**

VALUES: 43,144 (>99%)
MISSING: 10 (<1%)

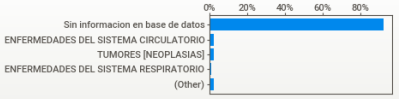
DISTINCT: 16 (<1%)



60 **D_capitulo_DX_Muerte**

VALUES: 43,154 (100%)
MISSING: ---

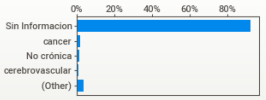
DISTINCT: 17 (<1%)



61 **ClasificacionCronico_DX_MUERTE**

VALUES: 43,143 (>99%)
MISSING: 11 (<1%)

DISTINCT: 38 (<1%)



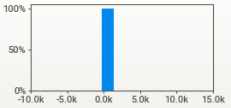
62 **HorasdeAlta**

VALUES: 43,154 (100%)
MISSING: ---

DISTINCT: 41 (<1%)

ZEROES: 3,154 (7%)

MAX	11,448	RANGE	19,824
95%	120	IQR	48.0
Q3	72	STD	78.5
AVG	50	VAR	6,168
MEDIAN	48		
Q1	24	KURT.	13,394
5%	0	SKEW	43.8
MIN	-8,376	SUM	2.1M



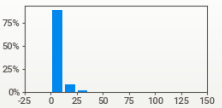
63 **DiasEstanciaServicio**

VALUES: 43,154 (100%)
MISSING: ---

DISTINCT: 84 (<1%)

ZEROES: 1,485 (3%)

MAX	122	RANGE	122
95%	18	IQR	5.00
Q3	7	STD	6.52
AVG	6	VAR	42.6
MEDIAN	4		
Q1	2	KURT.	22.7
5%	1	SKEW	3.53
MIN	0	SUM	245k



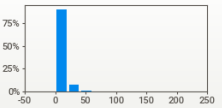
64 **DiasEstanciaClinica**

VALUES: 43,154 (100%)
MISSING: ---

DISTINCT: 134 (<1%)

ZEROES: 331 (<1%)

MAX	207	RANGE	207
95%	28	IQR	8.00
Q3	11	STD	10.9
AVG	9	VAR	119
MEDIAN	5		
Q1	3	KURT.	34.2
5%	1	SKEW	4.30
MIN	0	SUM	381k



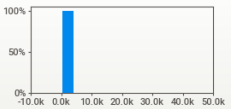
65 **DiasEstanciaClinica_exacto**

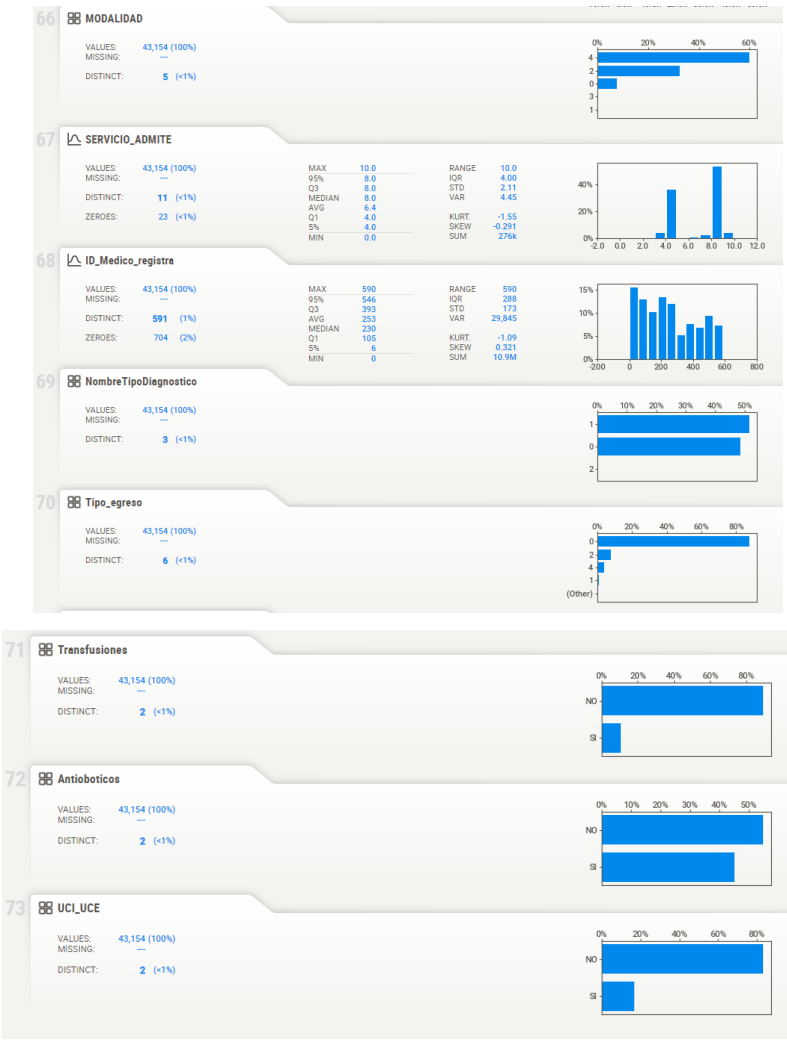
VALUES: 43,154 (100%)
MISSING: ---

DISTINCT: 43,122 (>99%)

ZEROES: ---

MAX	45,069	RANGE	45,069
95%	28	IQR	7.94
Q3	11	STD	306
AVG	11	VAR	93,627
MEDIAN	6		
Q1	3	KURT.	21,521
5%	1	SKEW	147
MIN	0	SUM	472k





A partir del mapa de calor con correlaciones y los diagramas de frecuencia se opta por eliminar las siguientes variables puesto que no aportan información valiosa a la predicción de nuestra variable de salida ya sea por datos faltantes o falta de correlación:

'AntencionPaciente llave', 'Año_Salida', 'MES','Aseguradora', 'Aseguradora2', 'Modalidad_Contrato', 'DX_F3_R3', 'DX_F3_R3', 'bloque', 'piso', 'RangoCapitulo', 'Categoria', 'D_subcapitulo_CIE10','DescSubgrupoOncologia', 'SubgrupoOncologia', 'DX_MEDICO', 'D_DX_MEDICO','D_Subcapitulo_DX_Muerte','Subcapitulo_DX_Muerte','ClasificacionCronico_DX_F3_R2','capitulo_DX_Muerte','D_capitulo_DX_Muerte','ClasificacionCronico_DX_F3_R1','ClasificacionCronico_DX_MUERTE', 'DX_R1', 'D_DX_R2', 'D_DX_R3', 'DX_F3_R3', 'D_DX_F3_R2', 'DiasEstanciaClinica_exacto', 'D_DX_F3', 'DX_F3_R1','D_DX_F3_R1', 'SERVICIO_ADMITE', 'ID_Medico_registra', 'NombreTipoDiagnostico', 'D_DX_pre_cx', 'DX_pre_cx', 'vGenero', 'Tipo_egreso'

Modificaciones a la base de datos

Enfocandonos en el diagnostico principal del paciente como una de las variables determinantes para la predicción se procede a eliminar todas las variables relacionada con diagnosticos secundarios y se obtiene el siguiente data set, con las siguientes variables:

```
#  Column            Non-Null Count  Dtype
---  ------           --------------  -----
 0  EdadEgreso        43154 non-null  float64
 1  Tipo_Afiliacion   43154 non-null  int64
 2  Hospitalizacion   43154 non-null  int64
 3  Tipo_Internacion  42945 non-null  object
 4  Ciclo_Vital       43154 non-null  int64
 5  ESPECIALIDAD_GRD  43154 non-null  int64
 6  DX_PRINCIPAL      43154 non-null  object
 7  HorasdeAlta       43154 non-null  int64
 8  DiasEstanciaServicio  43154 non-null  int64
 9  DiasEstanciaClinica   43154 non-null  int64
 10 MODALIDAD         43154 non-null  int64
 11 Transfusiones     43154 non-null  int64
 12 Antioboticos      43154 non-null  int64
 13 UCI_UCE           43154 non-null  int64
```

Luego se hizo una codificación de las variables categóricas y normalización de las variables numéricas para que asi la base de datos se adapte mejor a un modelo predictivo. Luego de dichas modificaciones se obtuvo una base de datos con 2869 columnas y el mismo numero de filas. Se plantearon los siguientes modelos:

Random forest para la variable días de estancia clínica:

```
Error Cuadrático Medio: 56.572474062210226
Coeficiente de Determinación (R^2): 0.5642520704092626
Mean Absolute Error (MAE): 2.747564384932129
```

:

Random forest para la variable días de estancia servicio:

```
Error Cuadrático Medio: 16.457387860388785
Coeficiente de Determinación (R^2): 0.6284483985817577

Mean Absolute Error (MAE): 1.8653939599580893
```

Según los resultados obtenidos mediante el random forest se encontró una predicción para los días de estancia clínica con un error de mas o menos 2.75 días y para los días de estancia de servicio se encontró un error de mas o menos 1.87 días. .

Árbol de decisiones para la variable días de estancia clínica:

```
Error Cuadrático Medio: 77.53201959345512
Mean Absolute Error (MAE): 3.2090333294712856
```

Árbol de decisiones para la variable días de estancia servicio:

```
Error Cuadrático Medio: 28.399808828640946
Mean Absolute Error (MAE): 2.205711968485691
```

Regresión lineal para la variable días de estancia clínica:

```
Error Cuadrático Medio: 2.761431616846068e+18
Coeficiente de Determinación (R^2): -2.1269851278271376e+16
Mean Absolute Error (MAE): 2.6900858113605524
```

Regresión lineal para la variable días de estancia servicio:

```
Error Cuadrático Medio: 22.4502364330495
Coeficiente de Determinación (R^2): 0.4931503486653164
Mean Absolute Error (MAE): 1.8653939599580893
```

sdads

Red Neuronal

```
270/270 [==============================] - 1s 5ms/step - loss: 25.7859 - mae: 2.4953
Error Cuadrático Medio: 25.78586769104004
Error absoluto medio: 2.495349407196045
```

Conclusiones

Días de Estancia Clínica:

El modelo de Árbol de Decisiones tiene un Error Cuadrático Medio (MSE) de aproximadamente 77.53 y un MAE de alrededor de 3.21, lo que indica un rendimiento moderado.
El modelo de Random Forest muestra un MSE de alrededor de 56.57 y un MAE de alrededor de 2.75, lo que indica un mejor rendimiento en comparación con el Árbol de Decisiones.
El modelo de Regresión Lineal tiene un MSE extremadamente alto, lo que sugiere un mal ajuste del modelo a los datos en este caso.

Días de Estancia Servicio:

El modelo de Árbol de Decisiones tiene un MSE de aproximadamente 28.40 y un MAE de alrededor de 2.21, lo que muestra un rendimiento moderado.
El modelo de Random Forest muestra un MSE de alrededor de 16.46 y un MAE de aproximadamente 1.87, lo que indica un mejor rendimiento en comparación con el Árbol de Decisiones.

El modelo de Regresión Lineal tiene un MSE de alrededor de 22.45 y un MAE de aproximadamente 1.87, lo que indica un rendimiento razonable para la Regresión Lineal en este caso.
Red Neuronal:

La red neuronal tiene un MSE de aproximadamente 25.79 y un MAE de alrededor de 2.50. Si bien no es el mejor rendimiento en términos de MAE, la red neuronal es capaz de capturar relaciones más complejas en los datos.

En general, el modelo de Random Forest parece funcionar bien en ambos casos (Días de Estancia Clínica y Días de Estancia Servicio), proporcionando predicciones precisas con MAE bajos. La Regresión Lineal y el Árbol de Decisiones también dan resultados aceptables en algunos casos, pero el modelo de Red Neuronal puede ser una opción interesante para capturar relaciones más complejas si estás dispuesto a invertir en su entrenamiento y ajuste de hiperparámetros.

Se podrían seguir ajustando los hiperparámetros para mejorar aún más las predicciones.