

Modelo predictivo de días de estancia hospitalaria como herramienta para la optimización de recursos

Proyecto integrador

Inspira Crea Transforma

**Gustavo Rubio
Juan Pablo Bertel
Gustavo Jerez**

Planteamiento del problema

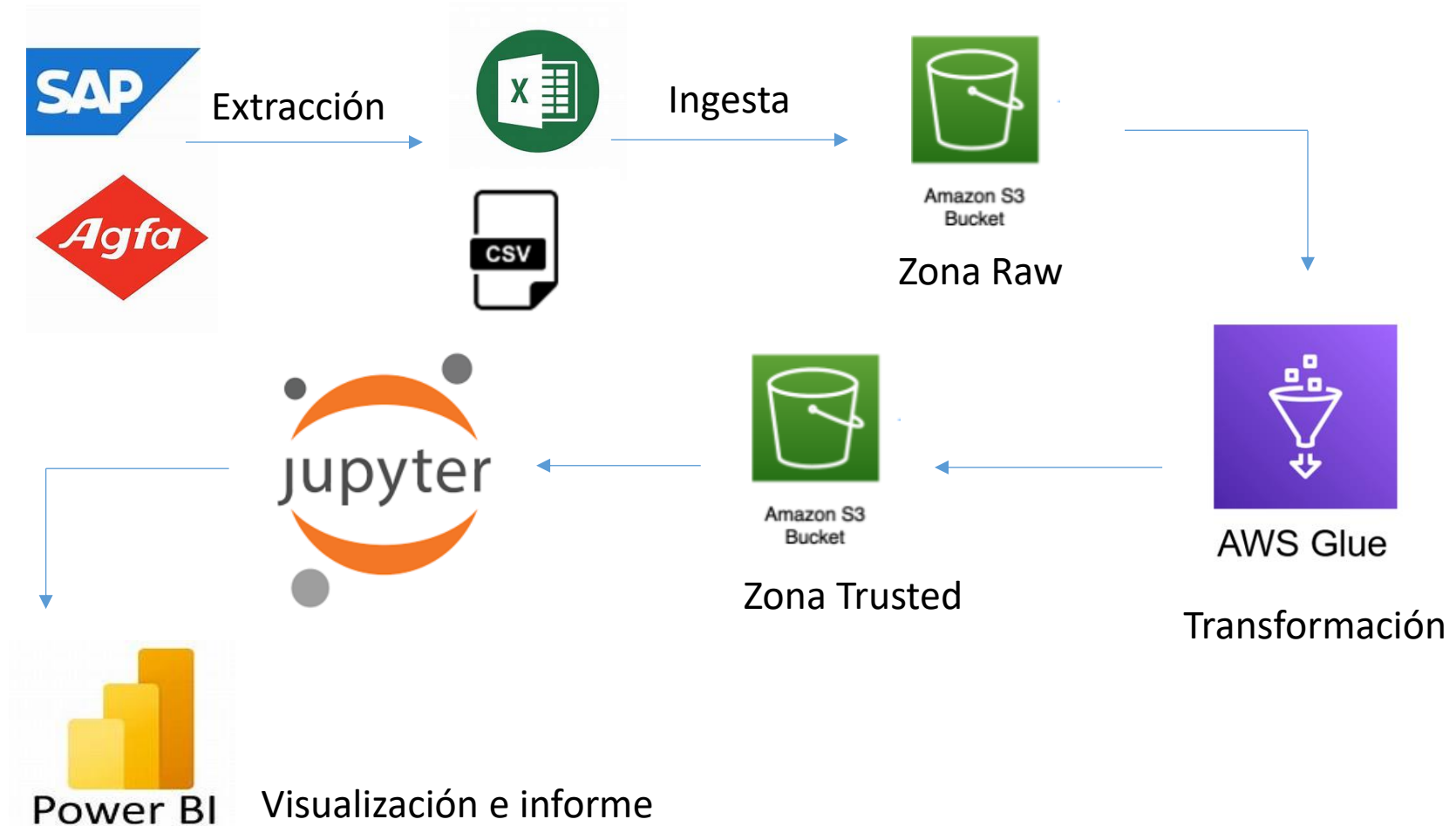
El objetivo principal del proyecto es estimar la duración de la estancia hospitalaria de los pacientes en función de variables clínicas y operativas. La estancia hospitalaria es un indicador clave de la eficiencia hospitalaria y está asociada a costos, uso de recursos y calidad de atención. Este análisis tiene el potencial de optimizar la gestión hospitalaria y prever necesidades operativas

¿Cómo podemos predecir la duración total de la estancia hospitalaria de un paciente utilizando información clínica y operativa disponible durante las primeras 24 horas de la admisión del paciente?

Metodología CRISP-DM



Arquitectura



Desarrollo del modelo

Etapa 1- Asegurar la calidad del dataset

- Definición de variables de modelación y variable de respuesta
- Separación de variables numéricas y categóricas para análisis
- Limpieza de texto en variables categóricas
- Eliminación de registros nulos y duplicados

	estancia_en_uci	edad	costo_operativo_estimado	peso_ir_estimado
count	78052.000000	78052.000000	7.805200e+04	78052.000000
mean	1.458335	41.395711	1.743718e+07	1.820014
std	5.720490	27.512572	2.576580e+07	1.777316
min	0.000000	0.000000	2.296900e+02	0.000000
25%	0.000000	15.000000	4.607076e+06	0.636500
50%	0.000000	41.000000	9.537359e+06	1.360200
75%	0.000000	65.000000	1.942217e+07	2.404025
max	295.000000	128.000000	6.484755e+08	22.457400

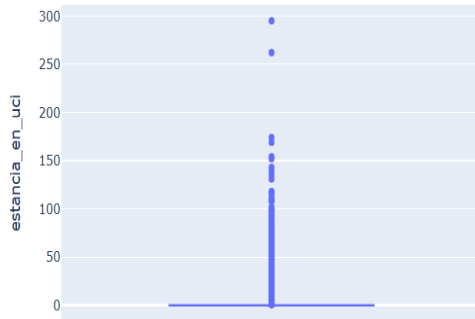
	estancia_en_uci	edad	costo_operativo_estimado	peso_ir_estimado	estancia_total
estancia_en_uci	1.000000	0.000672	0.674788	0.527176	0.594789
edad	0.000672	1.000000	0.072235	0.069794	0.050194
costo_operativo_estimado	0.674788	0.072235	1.000000	0.497779	0.816013
peso_ir_estimado	0.527176	0.069794	0.497779	1.000000	0.424316
estancia_total	0.594789	0.050194	0.816013	0.424316	1.000000

Desarrollo del modelo

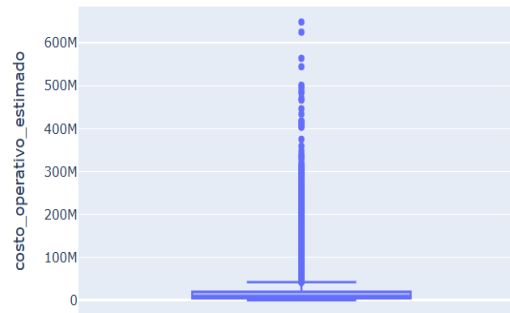
Etapa 2 - Análisis exploratorio de datos

Feats numéricas

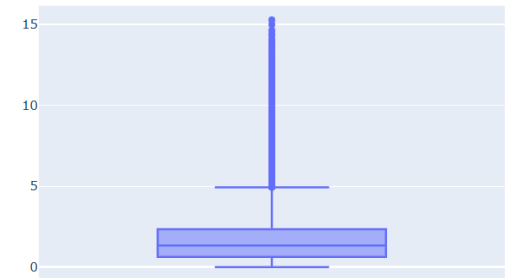
estancia_en_uci



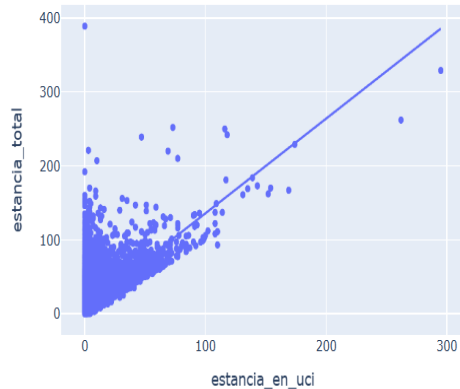
costo_operativo_estimado



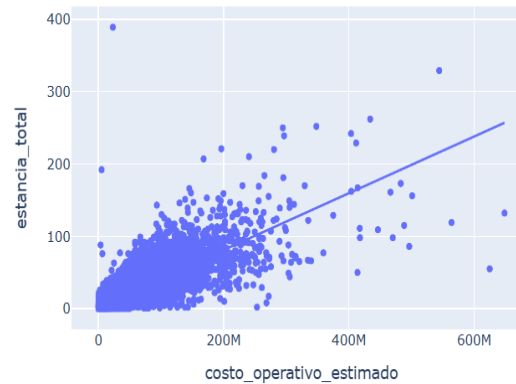
peso_ir_estimado



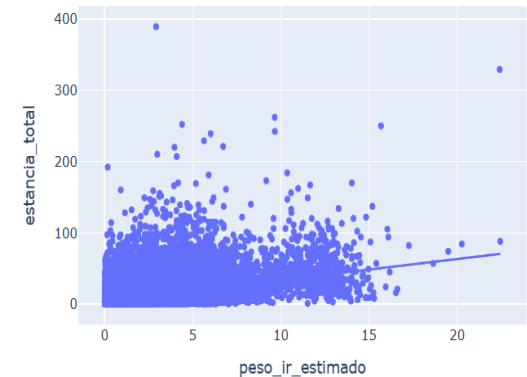
estancia_en_uci



costo_operativo_estimado



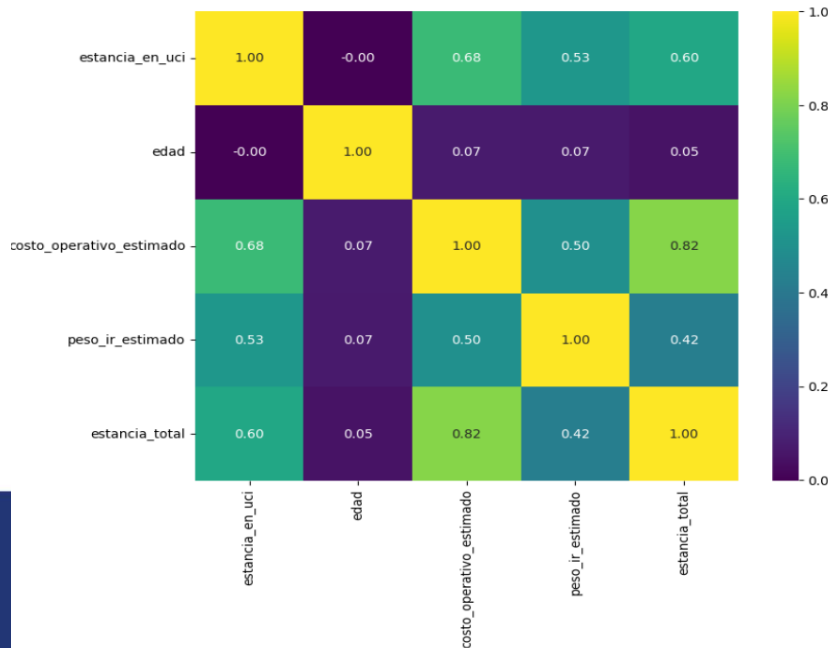
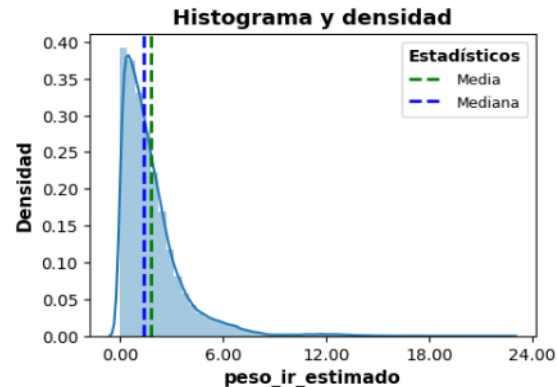
peso_ir_estimado



Desarrollo del modelo

Etapa 2 - Análisis exploratorio de datos

Feats numéricas



Frecuencia		TC y Posición		Dispersión y Forma		Normalidad	
Tendencia central				Posición			
		Resultado				Resultado	
Medida				Medida			
Moda		0.23		Mínimo		0.00	
Media		1.83		Percentil 1		0.03	
Media Armónica		0.00		Percentil 5		0.13	
Media Geométrica		0.00		Percentil 10		0.25	
Media Cuadrática		2.56		Percentil 25		0.64	
Media Trunc.(5%)		1.62		Percentil 50		1.37	
Media IQ		1.42		Percentil 75		2.42	
Media Wins.(5%)		1.73		Percentil 90		3.86	
Trimedia		1.45		Percentil 95		5.23	
Mediana		1.37		Percentil 99		8.56	
Mid Range		11.23		Máximo		22.46	
Mid Hinge		1.53					

Desarrollo del modelo

Etapa 2 - Análisis exploratorio de datos

Feats categóricas

Estadísticos - ir_cdm

Frecuencia

Conteo de registros

Conteo de frecuencias

Frec.Abs. Frec.Rel. Frec.Rel.Acum.
Con dato 76,258 100.00% 100.00%

Categorías

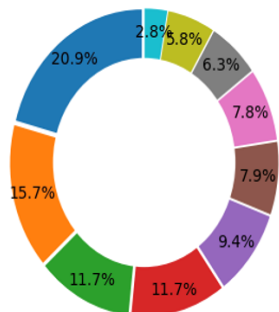
enfermedades y trastornos del aparato respiratorio	13,408	17.58%	17.58%
enfermedades y trastornos del aparato circulatorio	10,094	13.24%	30.82%
enfermedades y trastornos sistema musculoesqueletico y tejido conectivo	7,528	9.87%	40.69%
enfermedades y trastornos del aparato digestivo	7,522	9.86%	50.55%
enfermedades y trastornos del sistema nervioso	6,049	7.93%	58.49%
enfermedades y trastornos del aparato urinario	5,047	6.62%	65.11%

Gráficos descriptivos - ir_cdm

Circular

Pareto

Diagrama circular - Top 10 categorías



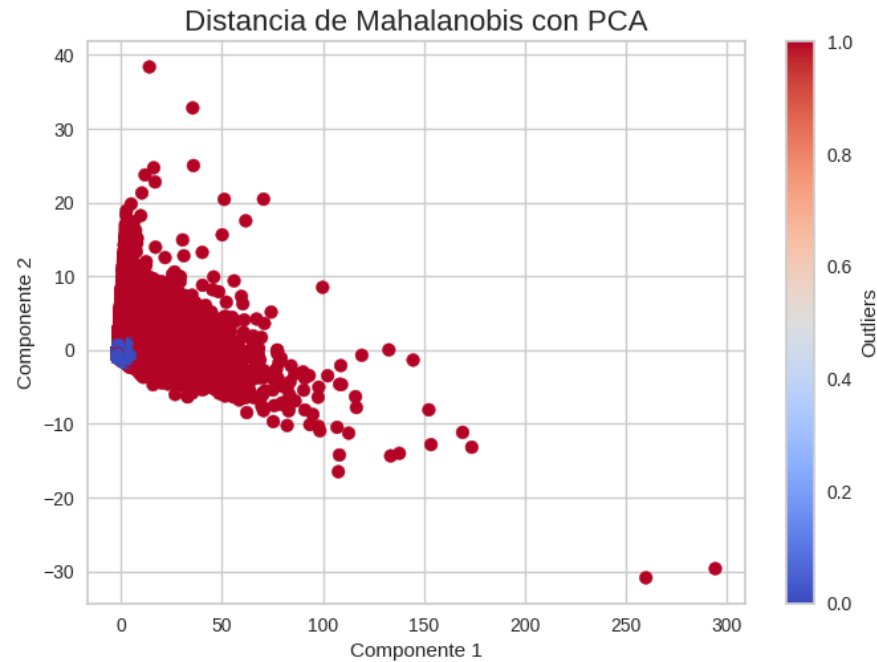
ir_cdm

- enfermedades y trastornos del aparato respiratorio
- enfermedades y trastornos del aparato circulatorio
- enfermedades y trastornos sistema musculoesqueletico y tejido conectivo
- enfermedades y trastornos del aparato digestivo
- enfermedades y trastornos del sistema nervioso
- enfermedades y trastornos del aparato urinario
- enfermedades y trastornos de piel, tejido subcutaneo y mama
- lesiones, envenenamiento y efecto toxico de drogas
- enfermedades y trastornos de higado, sistema biliar y pancreas
- enfermedades y trastornos mieloproliferativos y neoplasias mal diferenciadas

Desarrollo del modelo

Etapa 3 - Ingeniería de características

Remove outliers de feats numéricas



¿Por qué no utilizar la distancia de Mahalanobis?

Desarrollo del modelo

Etapa 3 - Ingeniería de características

Remove outliers de feats numéricas

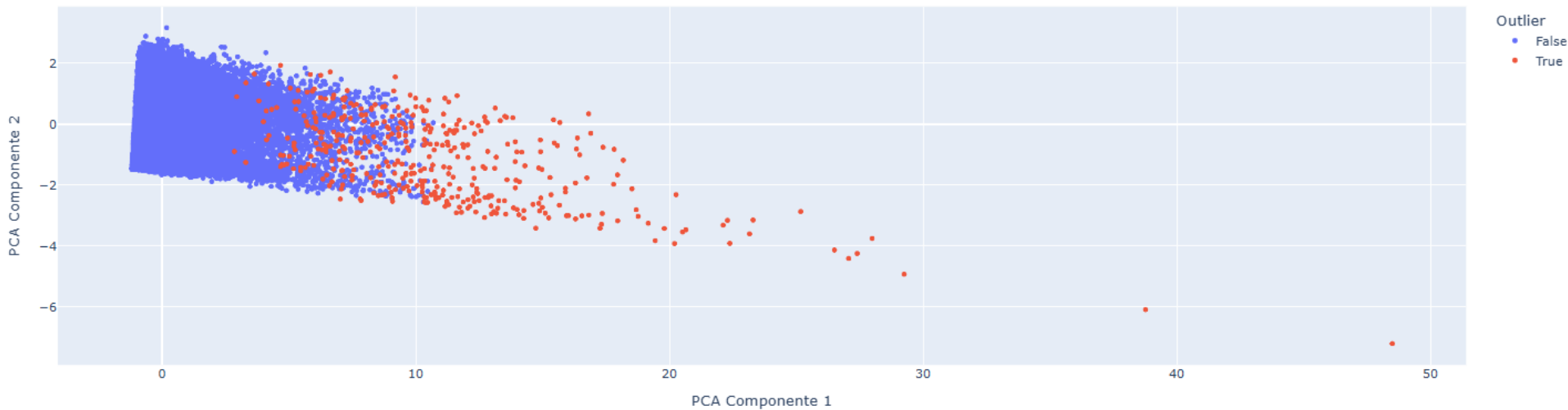
DBSCAN

Resultados de DBSCAN:

Coeficiente de Silueta: 0.8264040136567306

Índice de Davies-Bouldin: 0.5889918342636833

DBSCAN Clustering (Proyección PCA)



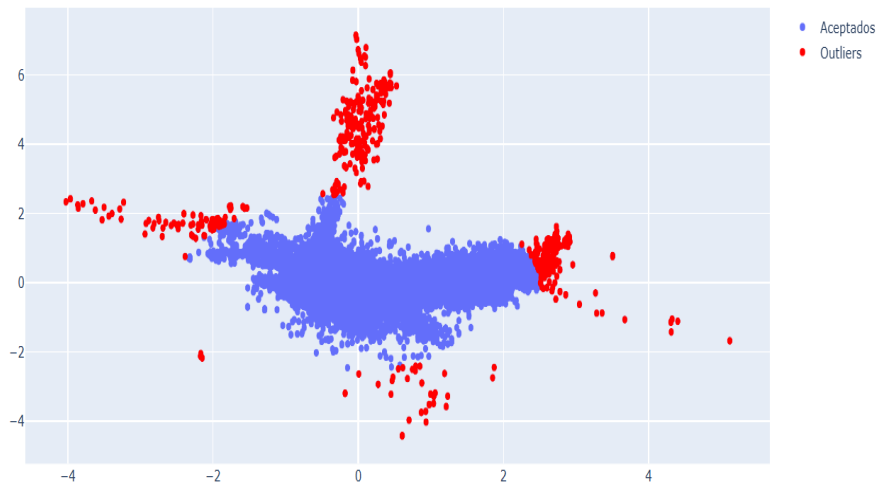
Desarrollo del modelo

Etapa 3 - Ingeniería de características

Remove outliers de feats categóricas

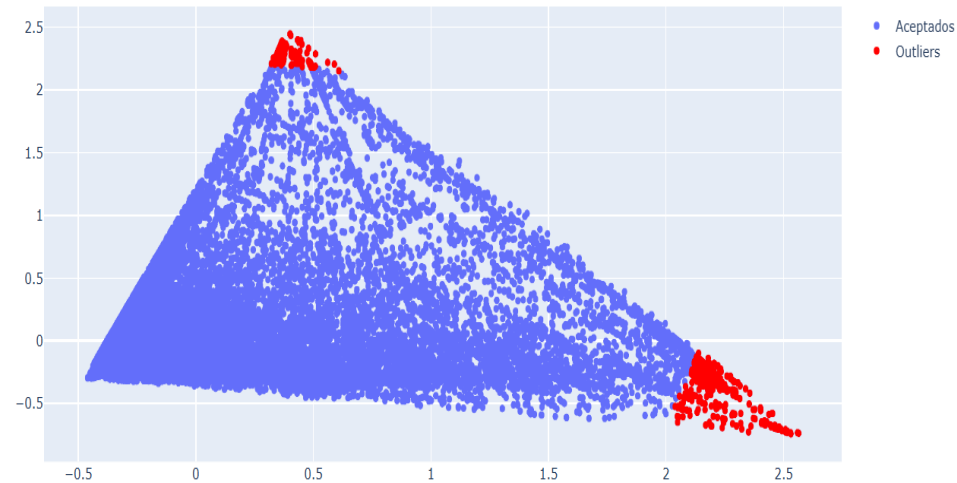
Multiple Correspondence Analysis (MCA)

PC0 vs PC1



MCA con One-Hot encoding

PC0 vs PC1



MCA con factorización

Desarrollo del modelo

Etapa 3 - Ingeniería de características

Remove outliers calculados

Outliers de variables numéricas - Mahalanobis: 4450

Outliers de variables numéricas - DBSCAN: 406

Outliers de variables categóricas - One Hot: 2092

Outliers de variables categóricas - Factorización: 2561

Registros del dataset inicial: 76258

Registros del dataset sin outliers: 71228


Porcentaje de datos removidos como outliers: 6.6%

estancia_total	
count	71228.000000
mean	9.174426
std	9.826989
min	0.000000
1%	0.000000
10%	2.000000
20%	3.000000
25%	3.000000
40%	5.000000
50%	6.000000
75%	11.000000
90%	21.000000
95%	30.000000
99%	54.000000
max	54.000000

También se acota la variable de respuesta por encima percentil 99 (Winsorizing)

Desarrollo del modelo

Etapa 4 - Preparación de los datos

- Particionar el dataset 
- Reemplazar moda en categorías con poca frecuencia
- Estandarizar datos de entrenamiento
- Validar multicolinealidad en data de entrenamiento

Porcentaje de datos en partición train: 90.0% - registros: 64105
Porcentaje de datos en partición test: 10.0% - registros: 7123

	Variable	VIF
0	ir_cdm	1.181699
1	ir_grd_base	2.939995
2	nivel_de_complejidad	2.304809
3	procedimiento_principal	1.592875
4	diagnostico_principal	1.322133
5	estancia_en_uci	2.112041
6	edad	1.038850
7	costo_operativo_estimado	2.187669
8	peso_ir_estimado	1.956964

Desarrollo del modelo

Etapa 5 - Entrenamiento del modelo

Torneo de modelos

Torneo 3: Removiendo outliers en todo el dataset - Mejor torneo

- Feats numéricas: DBSCAN
- Feats categoricas: MCA con factorización y One-Hot
- Estandarización: StandardScaler
- Regularización target encoding: 1
- Acotación de feat target: Si

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	3.2918	24.3708	4.9362	0.7489	0.5008	0.6081	5.8130
knn	K Neighbors Regressor	3.5773	29.2414	5.4071	0.6987	0.5381	0.6440	0.3620
lr	Linear Regression	3.6818	30.6978	5.5400	0.6838	0.5475	0.6930	0.3280
ridge	Ridge Regression	3.6818	30.6978	5.5400	0.6838	0.5475	0.6930	0.0160
lasso	Lasso Regression	3.8024	32.5426	5.7040	0.6648	0.5641	0.7645	0.0160
dt	Decision Tree Regressor	4.5544	48.4978	6.9629	0.5006	0.6818	0.7782	0.1380

Desarrollo del modelo

Regresión lineal

OLS Regression Results

```
=====
Dep. Variable:      estancia_total    R-squared:                0.683
Model:              OLS              Adj. R-squared:           0.683
Method:             Least Squares     F-statistic:             1.537e+04
Date:               Sun, 01 Dec 2024  Prob (F-statistic):         0.00
Time:               17:17:54          Log-Likelihood:          -2.0060e+05
No. Observations:   64105            AIC:                     4.012e+05
Df Residuals:       64095            BIC:                     4.013e+05
Df Model:           9
Covariance Type:    nonrobust
=====
```

Mean Squared Error: 5.528730935339455
R^2 Score: 0.6812341378235647
Mean Squared Error Traim: 5.506468327398332
R^2 Score: 0.6831368000142211

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
const         9.1748      0.022     419.998      0.000         9.132         9.218
x1             0.2947      0.032       9.283      0.000         0.232         0.357
x2            -0.1826      0.022      -8.202      0.000        -0.226        -0.139
x3             6.6736      0.032     206.548      0.000         6.610         6.737
x4            -0.2395      0.031      -7.836      0.000        -0.299        -0.180
x5            -0.1745      0.024      -7.347      0.000        -0.221        -0.128
x6             0.4636      0.037     12.377      0.000         0.390         0.537
x7             0.0004      0.033       0.012      0.991        -0.065         0.065
x8             1.0777      0.028     39.088      0.000         1.024         1.132
x9             1.3475      0.025     53.647      0.000         1.298         1.397
=====
```

```
=====
Omnibus:                13791.735    Durbin-Watson:                2.000
Prob(Omnibus):           0.000      Jarque-Bera (JB):             130949.994
Skew:                    0.766      Prob(JB):                     0.00
Kurtosis:                9.832      Cond. No.                     4.11
=====
```

Desarrollo del modelo

Regresión polinómica

Cross-Validation RMSE para polinomio de grado: 3: [5.08447335 5.12517997 4.99600152 5.05107016 5.05102584]

RMSE promedio: 5.0616

Desviación estandar de RMSE: 0.0426

Test MAE: 3.4588

Test RMSE: 5.1448

Test R-squared: 0.7255

OLS Regression Results

```

=====
Dep. Variable:      estancia_total    R-squared:      0.740
Model:              OLS              Adj. R-squared: 0.739
Method:             Least Squares    F-statistic:    834.6
Date:               Sun, 01 Dec 2024  Prob (F-statistic): 0.00
Time:               23:43:30          Log-Likelihood: -1.9432e+05
No. Observations:   64105            AIC:            3.891e+05
Df Residuals:       63886            BIC:            3.911e+05
Df Model:           218
Covariance Type:    nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	2.1688	0.017	126.109	0.000	2.135	2.203
x1	2.1688	0.017	126.109	0.000	2.135	2.203
x2	-0.0374	0.024	-1.537	0.124	-0.085	0.010
x3	-0.2792	0.033	-8.407	0.000	-0.344	-0.214
x4	0.0066	0.027	0.244	0.807	-0.046	0.060
x5	0.3785	0.023	16.475	0.000	0.333	0.423
x6	0.5996	0.021	28.461	0.000	0.558	0.641
x7	-0.6599	0.043	-15.426	0.000	-0.744	-0.576
x8	-0.0153	0.021	-0.712	0.477	-0.057	0.027
x9	2.7756	0.029	95.459	0.000	2.719	2.833

Variables iniciales del modelo

x255	-0.0125	0.017	-0.746	0.456	-0.045	0.020
x256	-0.0407	0.010	-4.012	0.000	-0.061	-0.021
x257	-0.0723	0.028	-2.560	0.010	-0.128	-0.017
x258	0.0843	0.020	4.279	0.000	0.046	0.123
x259	0.0414	0.025	1.656	0.098	-0.008	0.090
x260	-0.1143	0.029	-4.005	0.000	-0.170	-0.058
x261	0.0786	0.028	2.767	0.006	0.023	0.134
x262	-0.0136	0.034	-0.402	0.687	-0.080	0.053
x263	-0.1351	0.011	-12.727	0.000	-0.156	-0.114
x264	-0.0042	0.028	-0.153	0.878	-0.058	0.050
x265	-0.0094	0.023	-0.403	0.687	-0.055	0.036
x266	-0.0509	0.005	-10.444	0.000	-0.060	-0.041
x267	0.0199	0.017	1.200	0.230	-0.013	0.052
x268	-0.0519	0.012	-4.414	0.000	-0.075	-0.029
x269	-0.0049	0.012	-0.419	0.675	-0.028	0.018
x270	0.1334	0.037	3.623	0.000	0.061	0.206
x271	0.0464	0.029	1.617	0.106	-0.010	0.103
x272	0.0189	0.034	0.550	0.582	-0.049	0.086
x273	0.1076	0.012	9.039	0.000	0.084	0.131
x274	-0.0267	0.025	-1.086	0.278	-0.075	0.022
x275	0.0658	0.017	3.823	0.000	0.032	0.100
x276	-0.0773	0.023	-3.414	0.001	-0.122	-0.033
x277	0.0044	0.037	0.119	0.905	-0.068	0.077
x278	-0.0265	0.033	-0.806	0.420	-0.091	0.038
x279	-0.0491	0.016	-2.990	0.003	-0.081	-0.017
x280	0.0542	0.038	1.423	0.155	-0.020	0.129
x281	-0.0422	0.024	-1.761	0.078	-0.089	0.005
x282	-0.0002	0.006	-0.044	0.965	-0.011	0.011
x283	0.0349	0.016	2.122	0.034	0.003	0.067
x284	-0.0283	0.023	-1.218	0.223	-0.074	0.017
x285	-0.0420	0.010	-4.312	0.000	-0.061	-0.023

```

=====
Omnibus:              14004.260    Durbin-Watson:      1.998
Prob(Omnibus):        0.000        Jarque-Bera (JB):    106406.344
Skew:                 0.849        Prob(JB):            0.00
Kurtosis:             9.079        Cond. No.            2.73e+16
=====

```

Variables asociadas a términos cuadráticos, cúbicos e interacciones.



Desarrollo del modelo

Regresión KNN

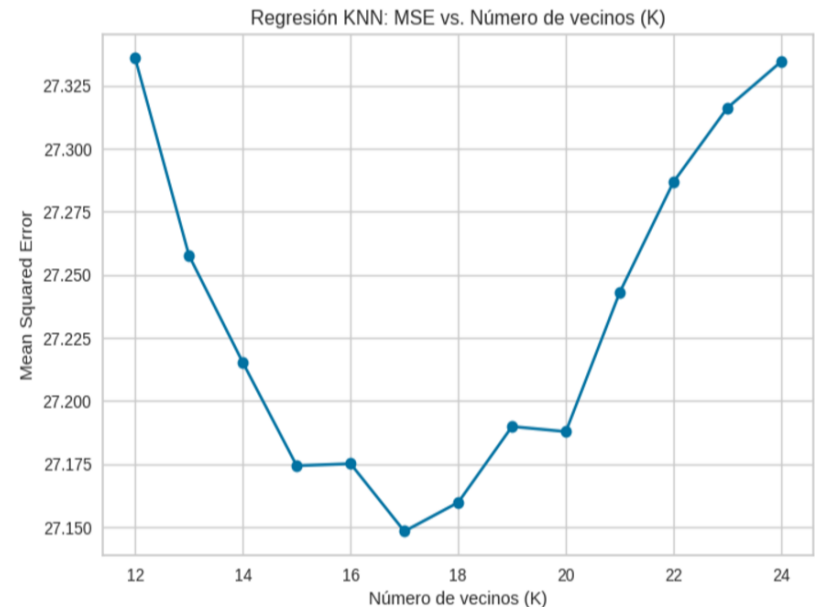
Se utilizó un enfoque de validación cruzada con cinco pliegues para evaluar diferentes combinaciones de hiperparámetros clave:

- **weights:** 'distance' (ponderación basada en la distancia de los vecinos).
- **n_neighbors:** 18 (cantidad óptima de vecinos).
- **metric:** 'euclidean' (distancia euclidiana como métrica principal).
- **algorithm:** 'auto' (selección automática del algoritmo más eficiente según el tamaño y estructura de los datos).

MAE: 3.4886

RMSE: 5.2528

R-squared: 0.7139

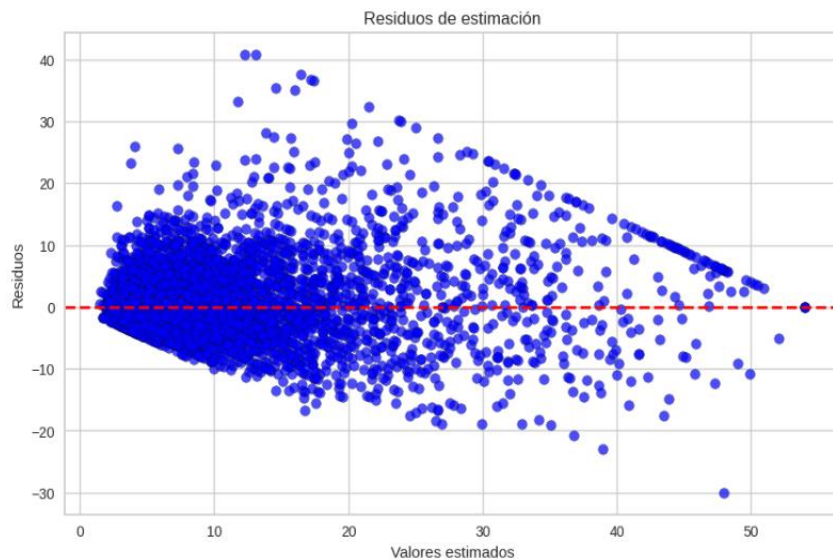




Desarrollo del modelo

Regresión KNN

Análisis de residuos



Conclusiones

- El modelo proporciona predicciones confiables, mejorando la planificación de recursos como camas y personal médico. Al incluir variables clave como diagnósticos y costos operativos, refuerza la eficiencia hospitalaria y la calidad de atención al paciente.
- El modelo K-Nearest Neighbors (KNN) se destacó por su simplicidad, interpretabilidad y facilidad de implementación, lo que lo hace adecuado para entornos hospitalarios. Aunque no es el modelo más avanzado, equilibra rendimiento técnico y usabilidad.
- El tratamiento meticuloso de los datos (eliminación de valores atípicos, normalización y transformación de variables) garantizó un modelo representativo y generalizable, resaltando la importancia de la ingeniería de datos para el éxito del proyecto.
- El uso de tecnologías como AWS Glue y PostgreSQL permite procesar grandes volúmenes de datos de manera escalable y eficiente. Herramientas de visualización y notificaciones automatizadas facilitan la integración de los resultados en los flujos operativos del hospital.
- El modelo no solo contribuye a la sostenibilidad financiera del hospital, sino también a mejorar la experiencia del paciente. A futuro, se recomienda combinar la simplicidad de KNN con modelos avanzados como Random Forest, realizar pruebas en tiempo real y considerar nuevas fuentes de datos para incrementar su precisión y alcance.

Interacción con el usuario



INFORME DE ESTIMACIÓN DE ESTANCIA HOSPITALARIA



Diagnostico principal

Procedimiento principal

Ir cdm

Ir grd base

Nivel de complejidad

Edad

9
Predicción (días)

\$17,1 mill.
Costo operativo estimado

Diagnostico principal	Procedimiento principal	Ir cdm	Ir grd base	Nivel de complejidad	Costo operativo estimado	Peso ir estimado
aborto espontaneo incompleto, complicado con infeccion genital y pelviana	dilatacion y legrado despues de parto o aborto	enfermedades y trastornos del aparato reproductor femenino	ph procedimientos sobre utero y cuello, dilatacion y legrado	baja complejidad	\$3.857.328,73	2,24
aborto espontaneo incompleto, complicado con infeccion genital y pelviana	dilatacion y legrado despues de parto o aborto	enfermedades y trastornos del aparato reproductor femenino	ph procedimientos sobre utero y cuello, dilatacion y legrado	baja complejidad	\$12.470.047,03	0,18
aborto espontaneo incompleto, complicado con infeccion genital y pelviana	legrado por aspiracion despues de parto o aborto	enfermedades y trastornos del aparato reproductor femenino	ph procedimientos sobre utero y cuello, dilatacion y legrado	baja complejidad	\$8.328.209,3	0,34
aborto espontaneo	hincia de medula nec	enfermedades y trastornos	ph procedimientos sobre	mediana complejidad	\$61.300.112	2,52

Prediccion estancia y Estancia total

Interacción con el usuario



EVALUACIÓN DE MODELO DE REGRESIÓN KNN



Grupo edad

- ☐ 0-20
- ☐ 21-40
- ☐ 41-60
- ☐ 61+

Nivel de complejidad

- ☐ alta complejidad
- ☐ baja complejidad
- ☐ mediana complejidad

3,23

MAE

0,75

R2

4,87

RMSE

23,69

MSE

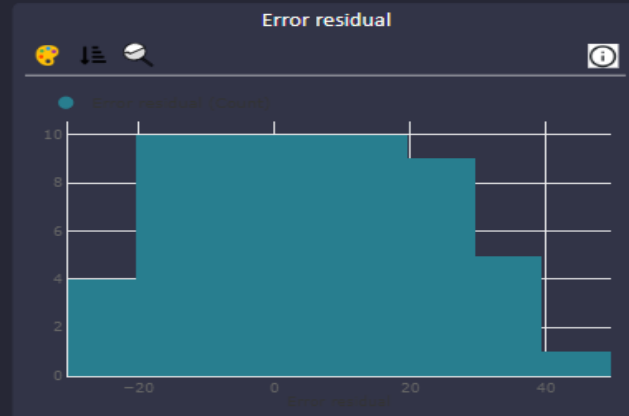
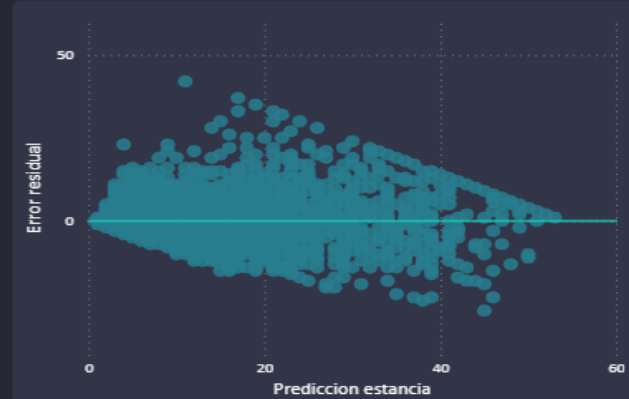
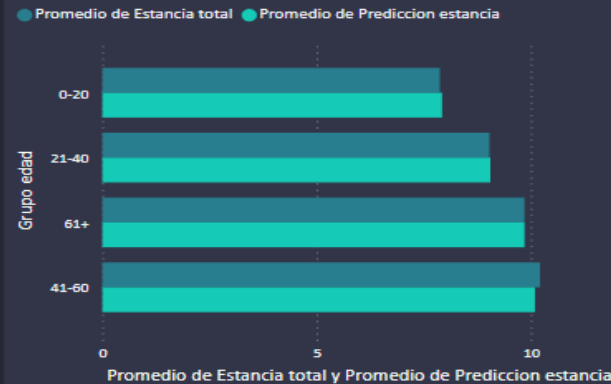
Ir grd base

Search



Procedimiento principal

Search



GRACIAS