Cas Kaggle

Luis Fernando Paz Galeano

GITHUB KAGGLE

Introducción

Las base de datos consiste en el análisis de anomalías, de caídas en personas, para ello se han utilizado las etiquetas ,010-000-024-033, 010-000-030-096, 020-000-032-221 y 020-000-033 que son representaciones codificadas en caliente de cada actividad del sensor.

Se utilizaron cuatro sensores durante los experimentos, que se han fijado al pecho, los tobillos y el cinturón de la persona.

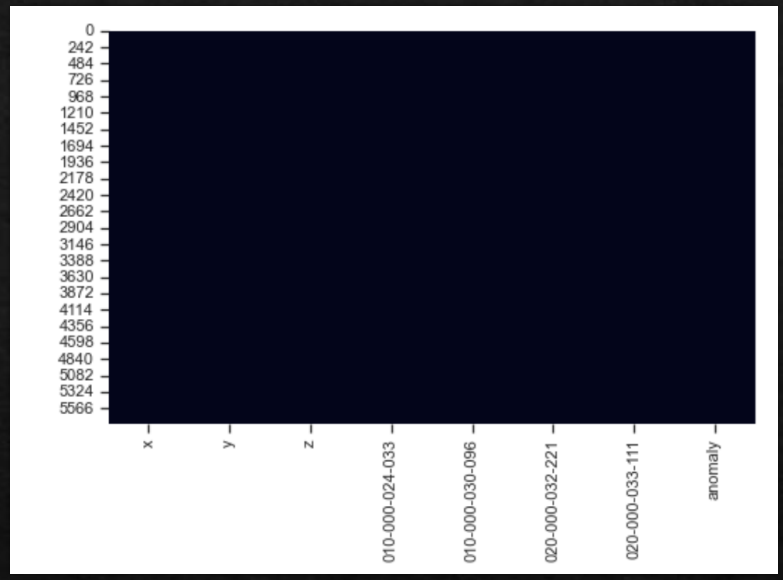
Las etiquetas representan el evento de caída / vida normal, donde 0 es normal y 1 es el evento anómalo de caída.

Cada archivo *.csv* es una persona distinta que se ha observado durante los experimentos. Se dan 5 personas de prueba y 20 personas de formación.

Muestra de Dataset

	x	y	z	010-000-024-033	010-000-030-096	020-000-032-221	020-000-033-111	anomaly
0	18.496	13.767	14.363	0.000	0.000	0.000	1.000	0.000
1	18.501	13.827	14.270	0.000	0.000	1.000	0.000	0.000
2	18.406	13.869	14.095	1.000	0.000	0.000	0.000	0.000
3	18.445	13.911	14.116	0.000	1.000	0.000	0.000	0.000
4	18.418	13.934	14.321	0.000	0.000	0.000	1.000	0.000
...	...	...	...	...	...	...	...	...
134224	9.539	13.049	12.345	0.000	0.000	0.000	1.000	0.000
134225	9.546	13.058	12.364	1.000	0.000	0.000	0.000	0.000
134226	9.575	13.080	12.180	0.000	0.000	1.000	0.000	0.000
134227	9.496	12.996	12.143	0.000	1.000	0.000	0.000	0.000
134228	9.534	13.063	12.104	0.000	0.000	0.000	1.000	0.000

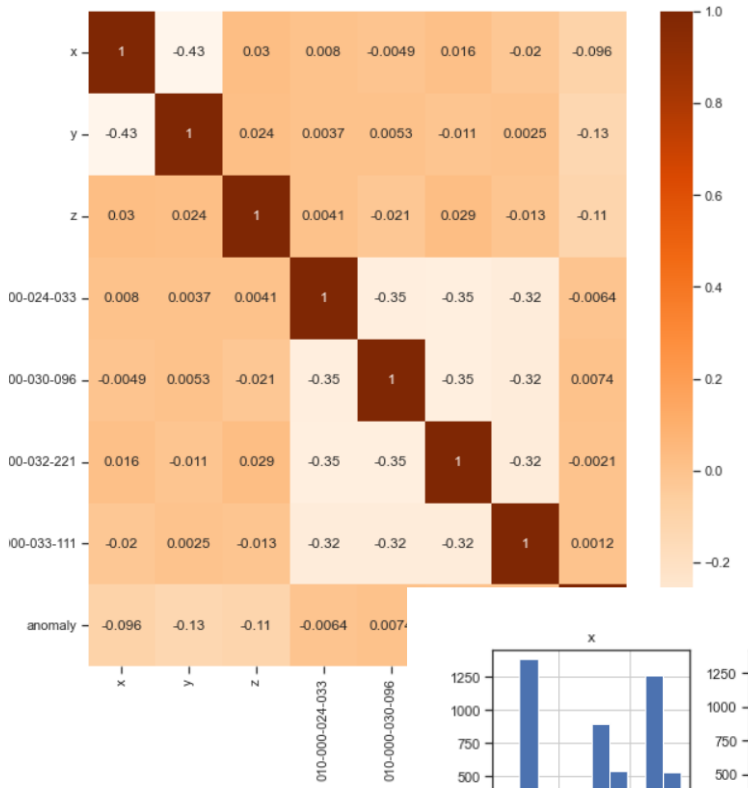
134229 rows × 8 columns



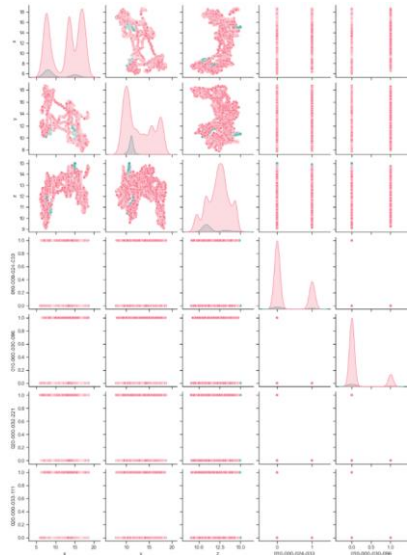
	x	y	z	010-000-024-033	010-000-030-096	020-000-032-221	020-000-033-111	anomaly
count	5805.000	5805.000	5805.000	5805.000	5805.000	5805.000	5805.000	5805.000
mean	12.660	12.698	12.263	0.286	0.149	0.283	0.282	0.056
std	3.948	3.037	1.286	0.452	0.356	0.450	0.450	0.231
min	5.989	7.767	9.053	0.000	0.000	0.000	0.000	0.000
25%	7.992	10.189	11.260	0.000	0.000	0.000	0.000	0.000
50%	13.516	11.710	12.377	0.000	0.000	0.000	0.000	0.000
75%	16.423	15.538	13.103	1.000	0.000	1.000	1.000	0.000
max	18.658	18.700	15.011	1.000	1.000	1.000	1.000	1.000

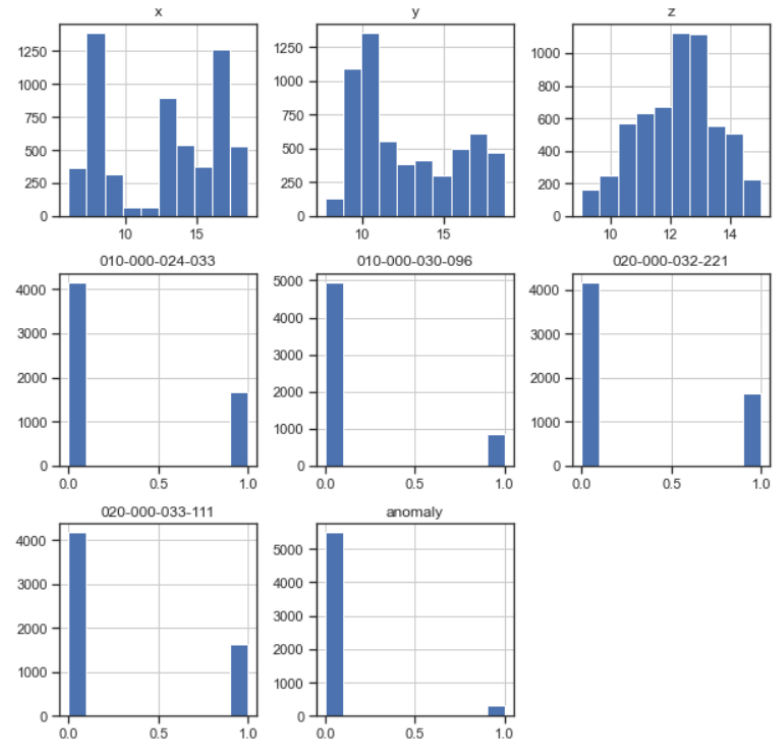
134229 rows × 8 columns

Matriz de correlación, distribución y muestras.

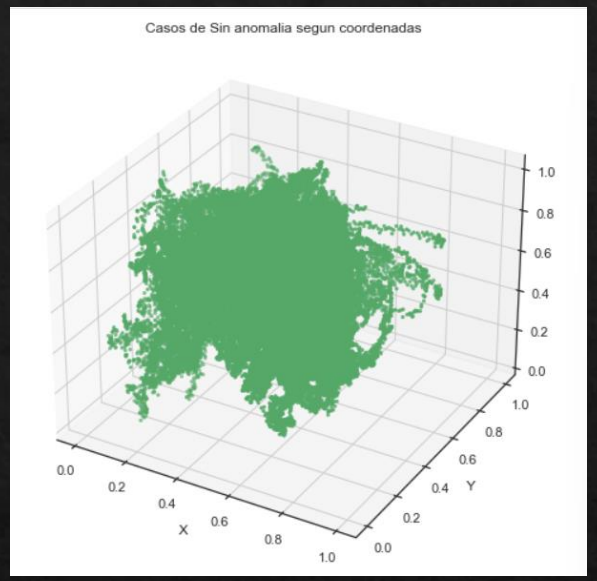


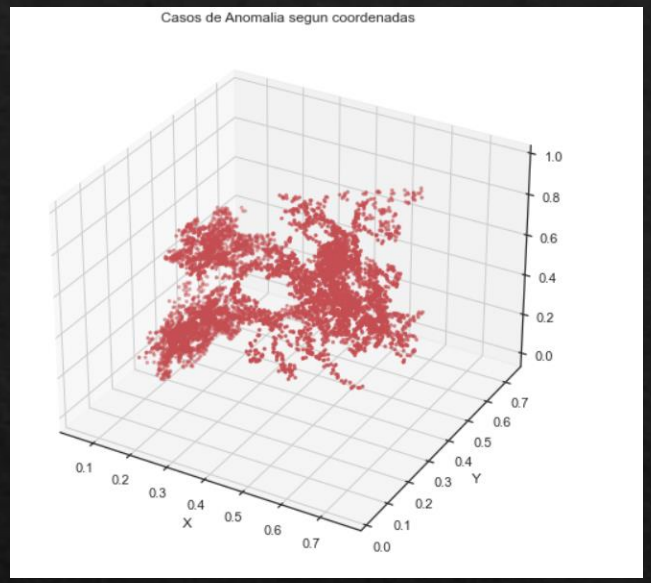
	anomaly
anomaly	1.000000
-000-030-096	0.007371
-000-033-111	0.001192
-000-032-221	-0.002069
-000-024-033	-0.006422
x	-0.096403
z	-0.113559
y	-0.130814

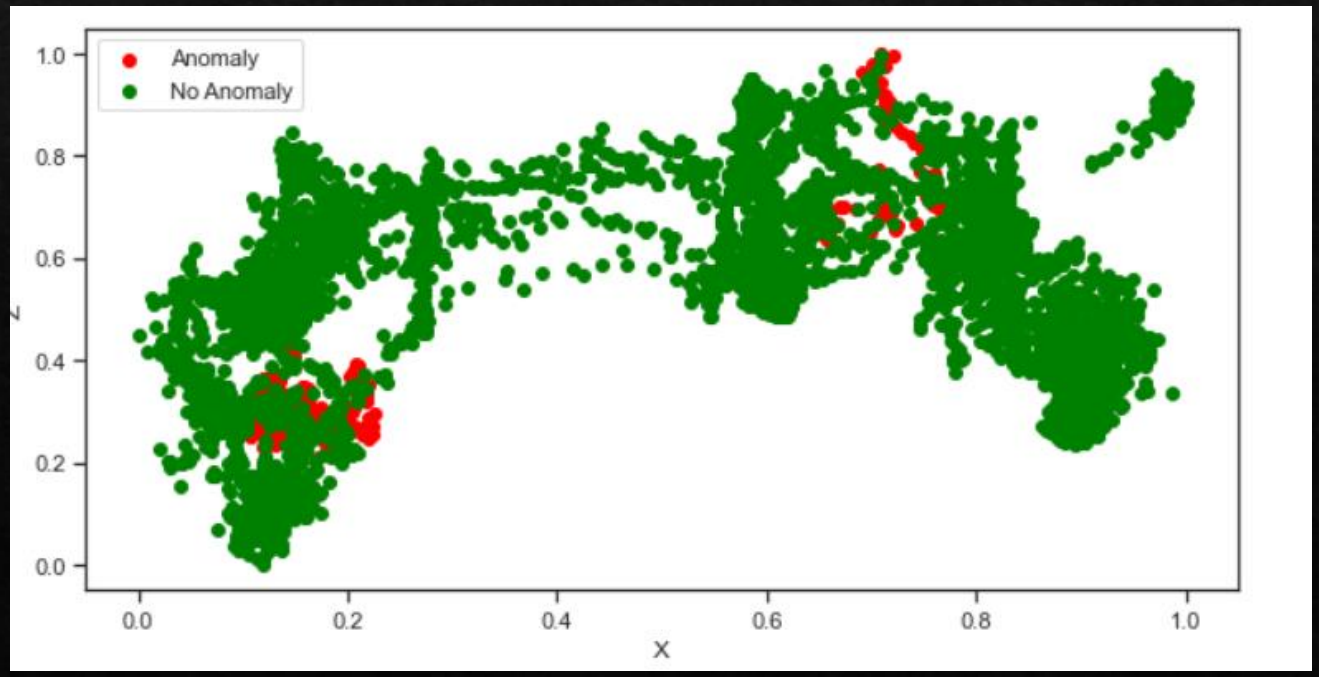




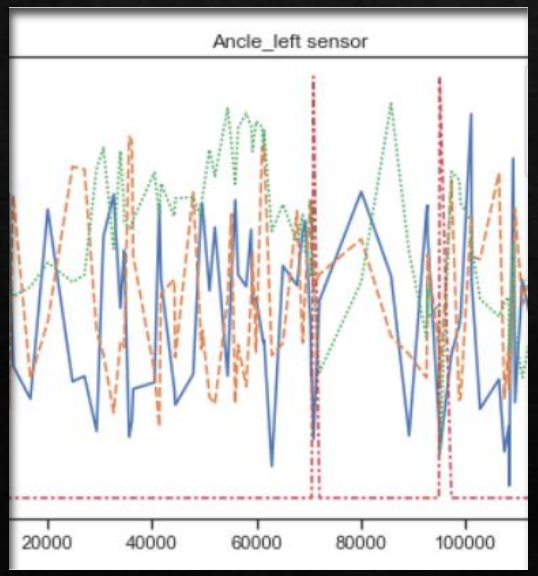
Posición de los sensores

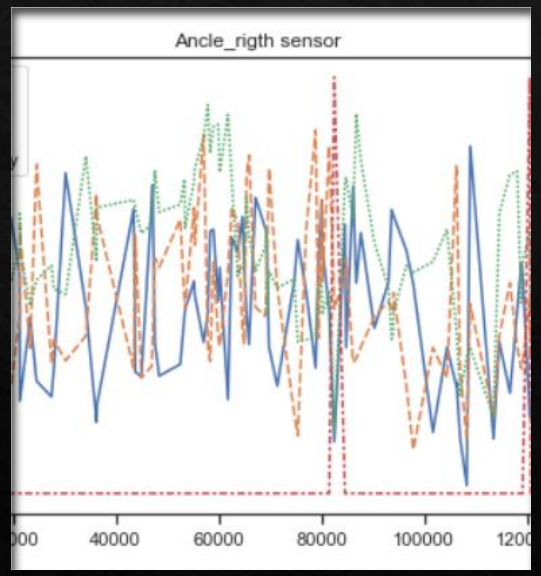


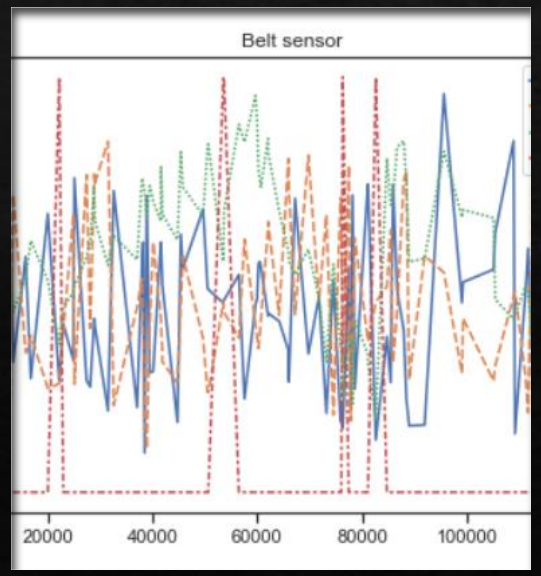


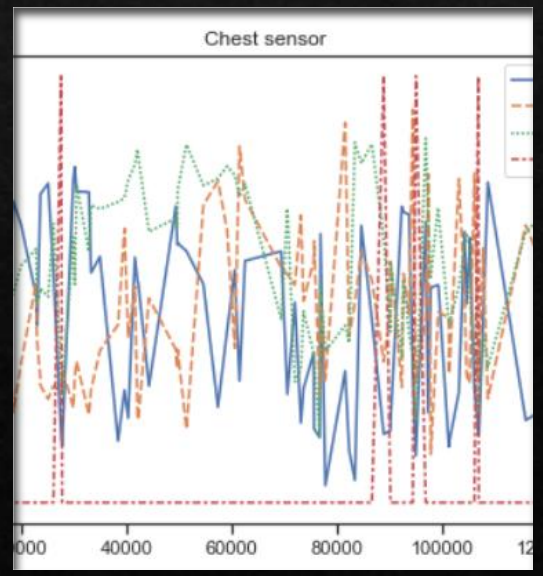


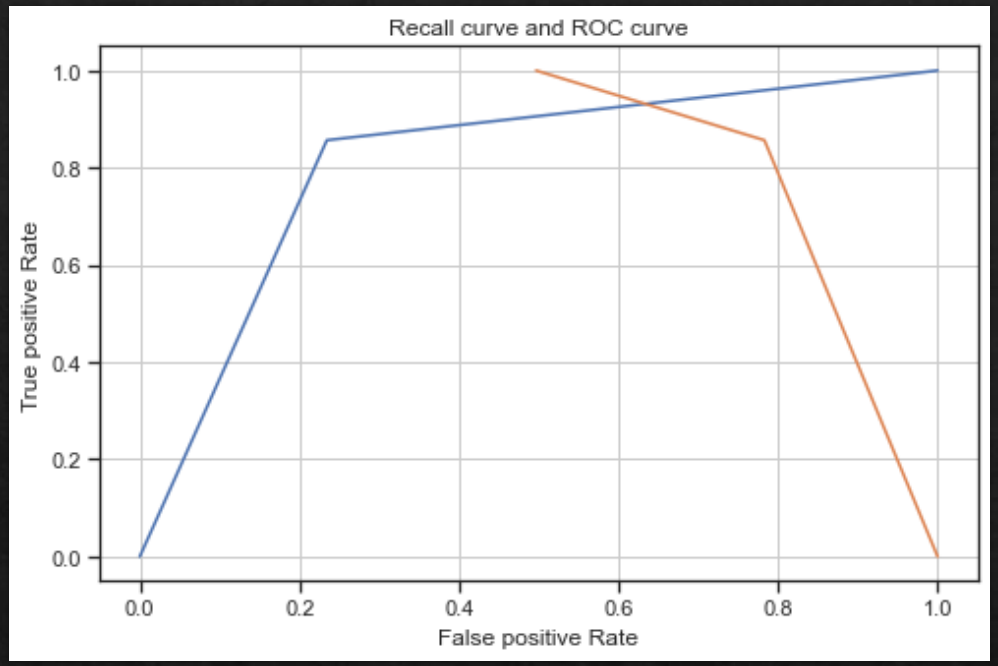
Posición de los sensores, comportamiento.

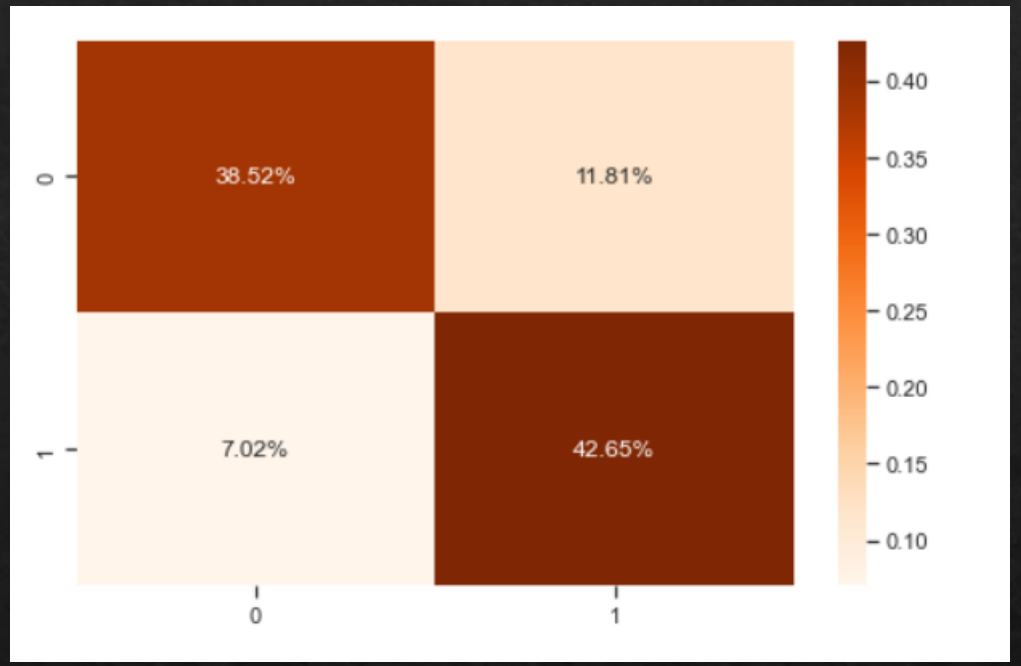










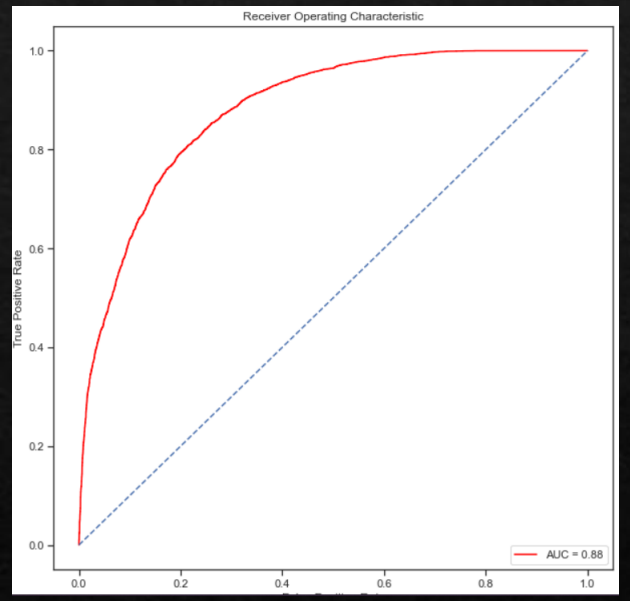


Clasificación RF

	Modelo	Precision	Tiempo
0	Random Forest	0.811	4.207
1	SVG_rbf	0.728	4.883
2	Knn	0.764	0.534
3	Regresion Logistica	0.668	0.070

Hyperparameter Search

```
RandomizedSearchCV(cv=10, estimator=RandomForestClassifier(), n_iter=20,
                   n_jobs=-1,
                   param_distributions={'max_features': ['auto', 'sqrt',
                                                         'log2'],
                                        'min_samples_leaf': range(10, 100, 10),
                                        'n_estimators': range(10, 100, 10)},
                   scoring='accuracy')
```



Conclusiones

- Durante el desarrollo de la implementación del caso Kaggle, pude constatar la importancia de poder interpretar los datos, ya que es imprescindible para un correcto estudio u análisis.
- En mi caso por la distribución de los datos, pude constatar que se producía Overfitting, esto debido a que contaba con un dataset, que si bien solo constaba de dos clases de clasificación, en este caso, aquellos casos que presentaban una anomalía y los que no, existían un mayor número de clases del segundo caso, por lo cual el modelo de entrenamiento, no era capaz de hacer una buena predicción, si los datos de evolución que proporcionaban eran buenos, en la práctica no estaba clasificando correctamente, por lo cual fue necesario implementar técnicas para balancear correctamente los datos.