Laboratorio 4 – BI

María Paula González Escallón - Pruebas
Jessica A. Robles Moreno – Modificación del Pipeline
Juan Esteban Vergara - API

Pruebas endpoint 1


```
{
  "list_of_inputs": [
      {
         "serial_no": 1,
         "gre_score": 301.0,
         "toefl_score": 92.0,
         "University Rating": 1.0,
         "sop": 1.85,
         "lor": 1.5,
         "CGPA": 7.71,
         "Research": 0.0
      },
      {
         "serial_no": 20,
         "gre_score": 320.0,
         "toefl_score": 110.0,
         "University Rating": 5.0,
         "sop": 5.0,
         "lor": 4.5,
         "CGPA": 9.22,
         "Research": 1.0
      },
      {
         "serial_no": 30,
         "gre_score": 299.0,
         "toefl_score": 96.0,
         "University Rating": 2.0,
         "sop": 1.5,
         "lor": 3.39,
         "CGPA": 7.86,
         "Research": 0.0
      },
      {
         "serial_no": 14,
         "gre_score": 328.0,
         "toefl_score": 110.0,
         "University Rating": 2.0,
         "sop": 3.95,
         "lor": 0.0,
         "CGPA": 9.15,
         "Research": 1.0
      },
      {
         "serial_no": 41,
```

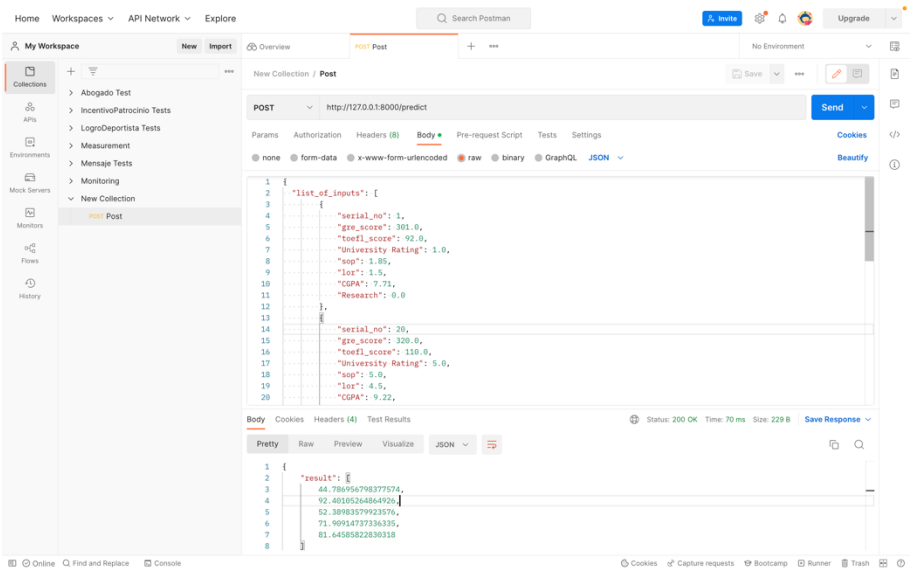
```
            "gre_score": 319.0,
            "toefl_score": 111.0,
            "University Rating": 3.0,
            "sop": 4.0,
            "lor": 0.0,
            "CGPA": 9.65,
            "Research": 1.0
        }
    ]
}
```


Resultados (en orden)


```
{
    "result": [
        44.786956798377574,
        92.40105264864926,
        52.38983579923576,
        71.90914737336335,
        81.64585822830318
    ]
}
```


Resultados esperados (en orden)

45.08
92.00
54.00
126.00
65.00



Al realizar las pruebas del primer endpoint, decidimos mandar 5 instancias (Se pueden observar en el apartado “Pruebas endpoint 1”), de las cuales pudimos observar que 3 nos daban resultados coherentes y 2 nos arrojaban predicciones incoherentes.

La primera instancia ({"serial_no": 1, "gre_score": 301.0, "toefl_score": 92.0, "University Rating": 1.0, "sop": 1.85, "lor": 1.5, "CGPA": 7.71, "Research": 0.0}), la cual en los datos de entrenamiento debería dar como resultado 45.08, nuestro modelo arroja una predicción aproximada de 44.79. Estos valores al ser muy parecidos, podemos interpretar que nuestro modelo dio un resultado coherente.

Para la segunda instancia ({"serial_no": 20, "gre_score": 320.0, "toefl_score": 110.0, "University Rating": 5.0, sop": 5.0, lor": 4.5, "CGPA": 9.22, "Research": 1.0}) el modelo nos arroja una predicción de 92.40, este dato comparado con la predicción esperada (92.00) podemos concluir que la predicción fue coherente ya que la diferencia entre los datos es mínima (0.40).

Del mismo modo, para la instancia 3 ({"serial_no": 30, gre_score": 299.0, "toefl_score": 96.0, "University Rating": 2.0, "sop": 1.5, "lor": 3.39, "CGPA": 7.86, "Research": 0.0}), se esperaba una predicción de 54.00 y nuestro modelo arrojo una predicción de 52.39. Este resultado nos da una diferencia de 1.61, sin embargo, consideramos que este numero no es significativo para clasificar la predicción como incoherente, por lo que podemos interpretar que la predicción fue coherente.

Por el otro lado, para la instancia 4 ({"serial_no": 14, "gre_score": 328.0, "toefl_score": 110.0, "University Rating": 2.0, "sop": 3.95, "lor": 0.0, "CGPA": 9.15, "Research": 1.0}), el modelo nos arroja una predicción de 71.91. Esta predicción la podemos clasificar como incoherente ya que el resultado esperado para esta predicción era de 126.00, dándonos una diferencia de más de 54.

Para la última instancia ({"serial_no": 41, "gre_score": 319.0, "toefl_score": 111.0, "University Rating": 3.0, "sop": 4.0, "lor": 0.0, "CGPA": 9.65, "Research": 1.0}), podemos concluir que el modelo arrojo una predicción incoherente ya que la diferencia entre el dato esperado (65.00) y el resultado obtenido (81.65) fue de aproximadamente 16.

Como el pipeline usado en la fastAPI es resultado de las transformaciones en el laboratorio 3, se asume que la raíz cuadrática de error medio sigue siendo de 10.524. Al ser esto la media, se entiende que las incoherencias mayores a 10 unidades son casos excepcionales y erróneos, así por cada valor excepcional incoherente (como la instancia 4 que tiene una diferencia de 54) habrá varios escenarios donde los valores que tengan errores de menos de 10 unidades como las primeras 3 instancias.

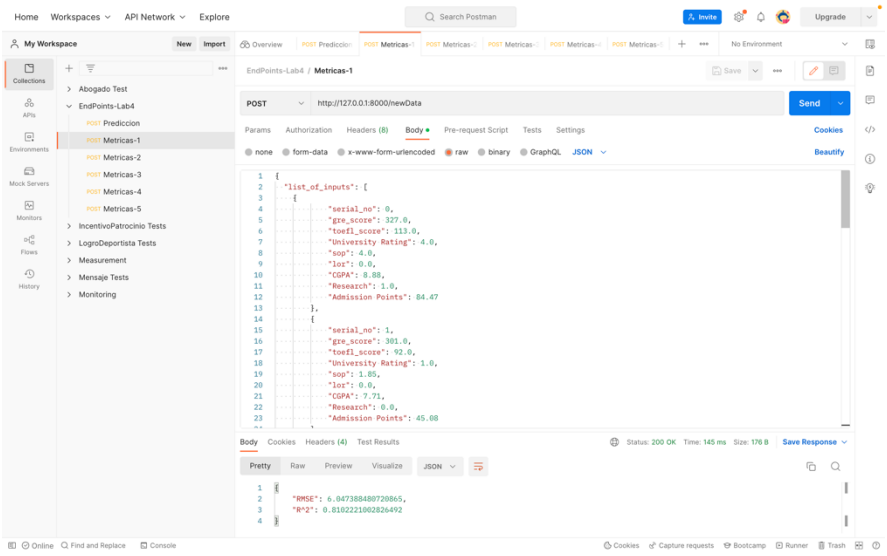
Prueba endpoint 2

1.


```
{
  "list_of_inputs": [
    {
        "serial_no": 0,
        "gre_score": 327.0,
        "toefl_score": 113.0,
        "University Rating": 4.0,
        "sop": 4.0,
        "lor": 0.0,
        "CGPA": 8.88,
        "Research": 1.0,
        "Admission Points": 84.47
```

```
},
{
    "serial_no": 1,
    "gre_score": 301.0,
    "toefl_score": 92.0,
    "University Rating": 1.0,
    "sop": 1.85,
    "lor": 0.0,
    "CGPA": 7.71,
    "Research": 0.0,
    "Admission Points": 45.08
},
{
    "serial_no": 2,
    "gre_score": 297.0,
    "toefl_score": 100.0,
    "University Rating": 1.0,
    "sop": 2.41,
    "lor": 0.0,
    "CGPA": 7.89,
    "Research": 0.0,
    "Admission Points": 47.42
},
{
    "serial_no": 3,
    "gre_score": 303.0,
    "toefl_score": 98.0,
    "University Rating": 3.0,
    "sop": 3.5,
    "lor": 0.0,
    "CGPA": 8.5,
    "Research": 0.0,
    "Admission Points": 62.0
},
{
    "serial_no": 4,
    "gre_score": 320.0,
    "toefl_score": 94.0,
    "University Rating": 2.0,
    "sop": 1.38,
    "lor": 0.0,
    "CGPA": 8.78,
    "Research": 1.0,
    "Admission Points": 73.0
},
{
    "serial_no": 5,
    "gre_score": 321.0,
    "toefl_score": 84.0,
    "University Rating": 4.0,
    "sop": 4.13,
    "lor": 0.0,
```

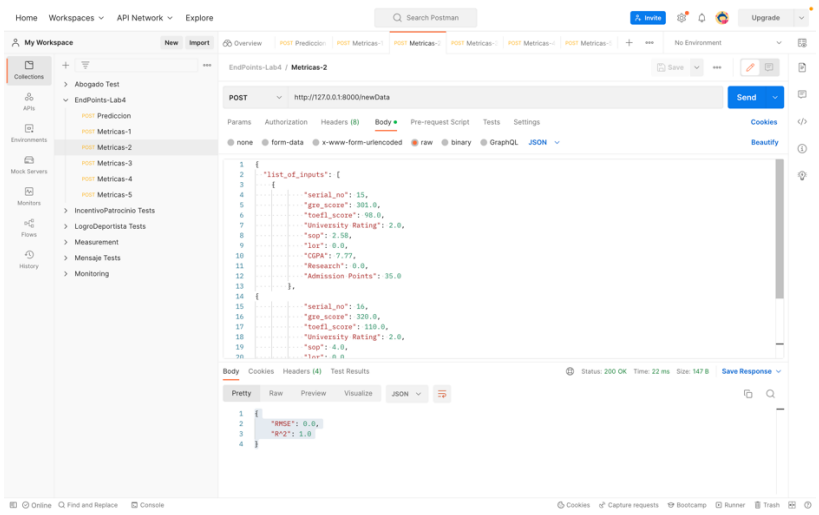
```
      "CGPA": 8.68,
      "Research": 1.0,
      "Admission Points": 69.0
    },
    {
      "serial_no": 6,
      "gre_score": 306.0,
      "toefl_score": 110.0,
      "University Rating": 3.0,
      "sop": 3.0,
      "lor": 0.0,
      "CGPA": 8.0,
      "Research": 0.0,
      "Admission Points": 70.0
    },
    {
      "serial_no": 7,
      "gre_score": 299.0,
      "toefl_score": 112.0,
      "University Rating": 1.0,
      "sop": 0.97,
      "lor": 0.0,
      "CGPA": 6.82,
      "Research": 1.0,
      "Admission Points": 40.03
    },
    {
      "serial_no": 8,
      "gre_score": 311.0,
      "toefl_score": 74.0,
      "University Rating": 4.0,
      "sop": 1.5,
      "lor": 0.0,
      "CGPA": 8.36,
      "Research": 0.0,
      "Admission Points": 57.0
    }
  ]
}
```



En esta prueba los resultados del RMSE y del coeficiente de determinación dan valores óptimos. Nos podemos dar cuenta de esto ya que el RMSE es un valor muy pequeño comparado al rango de nuestras instancias ingresadas y R^2 es un valor muy cercano al 1, ambos nos indican que se está aprendiendo bien.

2.


```
{
  "list_of_inputs": [
    {
        "serial_no": 15,
        "gre_score": 301.0,
        "toefl_score": 98.0,
        "University Rating": 2.0,
        "sop": 2.58,
        "lor": 0.0,
        "CGPA": 7.77,
        "Research": 0.0,
        "Admission Points": 35.0
    },
{
        "serial_no": 16,
        "gre_score": 320.0,
        "toefl_score": 110.0,
        "University Rating": 2.0,
        "sop": 4.0,
        "lor": 0.0,
        "CGPA": 8.56,
        "Research": 0.0,
        "Admission Points": 72.0
    }
  ]
}
```

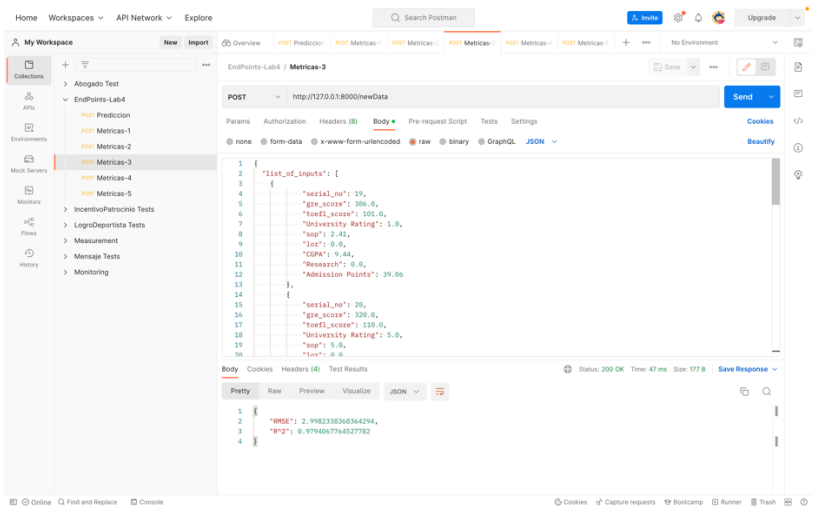


En la prueba los resultados del RMSE y del coeficiente de determinación dan valores óptimos. El RMSE nos da un valor muy pequeño comparado al rango de nuestras instancias ingresadas y el coeficiente de determinación es 1, el valor máximo que se puede alcanzar. Lo que esto nos está indicando es que el valor predicho y el valor real no cambia mucho y que el modelo se ajusta bien a la variable objetivo.

3.


```
{
 "list_of_inputs": [
  {
        "serial_no": 19,
        "gre_score": 306.0,
        "toefl_score": 101.0,
        "University Rating": 1.0,
        "sop": 2.41,
        "lor": 0.0,
        "CGPA": 9.44,
        "Research": 0.0,
        "Admission Points": 39.06
    },
    {
        "serial_no": 20,
        "gre_score": 320.0,
        "toefl_score": 110.0,
        "University Rating": 5.0,
        "sop": 5.0,
        "lor": 0.0,
        "CGPA": 9.22,
        "Research": 1.0,
        "Admission Points": 92.0
    },
    {
        "serial_no": 21,
        "gre_score": 293.0,
        "toefl_score": 102.0,
        "University Rating": 2.0,
        "sop": 2.42,
```

```
      "lor": 0.0,
      "CGPA": 7.17,
      "Research": 0.0,
      "Admission Points": 49.02
    },
    {
      "serial_no": 22,
      "gre_score": 324.0,
      "toefl_score": 74.0,
      "University Rating": 5.0,
      "sop": 4.19,
      "lor": 0.0,
      "CGPA": 8.88,
      "Research": 1.0,
      "Admission Points": 87.02
    },
    {
      "serial_no": 23,
      "gre_score": 310.0,
      "toefl_score": 100.0,
      "University Rating": 3.0,
      "sop": 3.41,
      "lor": 0.0,
      "CGPA": 7.76,
      "Research": 0.0,
      "Admission Points": 45.93
    },
    {
      "serial_no": 24,
      "gre_score": 298.0,
      "toefl_score": 109.0,
      "University Rating": 2.0,
      "sop": 2.98,
      "lor": 0.0,
      "CGPA": 7.21,
      "Research": 0.0,
      "Admission Points": 45.0
    },
    {
      "serial_no": 25,
      "gre_score": 325.0,
      "toefl_score": 107.0,
      "University Rating": 4.0,
      "sop": 4.5,
      "lor": 0.0,
      "CGPA": 9.06,
      "Research": 1.0,
      "Admission Points": 79.0
    }
  ]
}
```

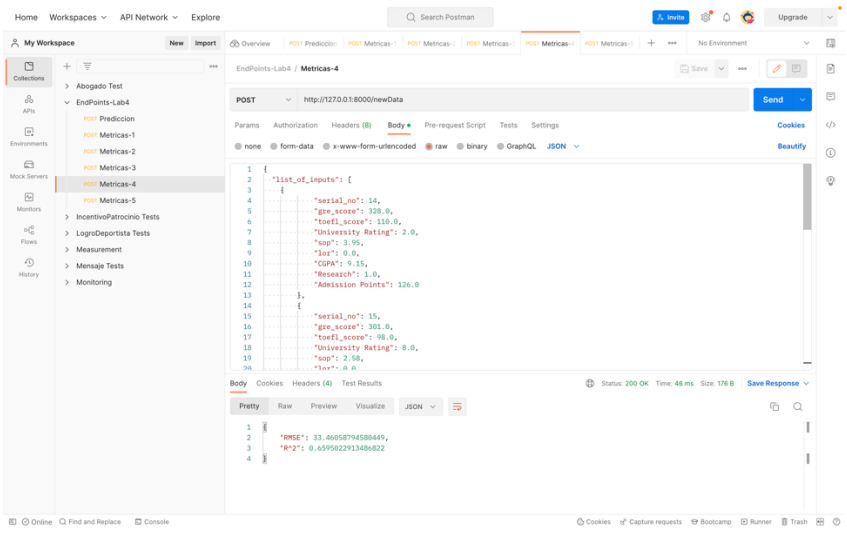


Con las pruebas obtenidas nos damos cuenta de que el valor predicho y el valor real no cambia mucho y que el modelo se ajusta bien a la variable objetivo. Podemos deducir esto ya que el RMSE es pequeño y el R^2 es cercano al 1. Creemos que esto sucede ya que los datos con puntos de admisión similares tienen características similares, y esto hace que para el modelo sea más fácil aprender a predecir.

4.


```
{
  "list_of_inputs": [
    {
        "serial_no": 14,
        "gre_score": 328.0,
        "toefl_score": 110.0,
        "University Rating": 2.0,
        "sop": 3.95,
        "lor": 0.0,
        "CGPA": 9.15,
        "Research": 1.0,
        "Admission Points": 126.0
    },
    {
        "serial_no": 15,
        "gre_score": 301.0,
        "toefl_score": 98.0,
        "University Rating": 8.0,
        "sop": 2.58,
        "lor": 0.0,
        "CGPA": 10.77,
        "Research": 0.0,
        "Admission Points": 0.0
    },
    {
        "serial_no": 16,
        "gre_score": 320.0,
        "toefl_score": 110.0,
        "University Rating": 2.0,
        "sop": 4.0,
```

```
        "lor": 0.0,
        "CGPA": 2.56,
        "Research": 0.0,
        "Admission Points": 0.0
    },
    {
        "serial_no": 67,
        "gre_score": 275.0,
        "toefl_score": 94.0,
        "University Rating": 0.0,
        "sop": 4.0,
        "lor": 0.0,
        "CGPA": 0,
        "Research": 0.0,
        "Admission Points": 127.5
    },
{
        "serial_no": 68,
        "gre_score": 321.0,
        "toefl_score": 114.0,
        "University Rating": 1.0,
        "sop": 4.0,
        "lor": 0.0,
        "CGPA": 9.12,
        "Research": 0.0,
        "Admission Points": 85.0
    }
  ]
}
```

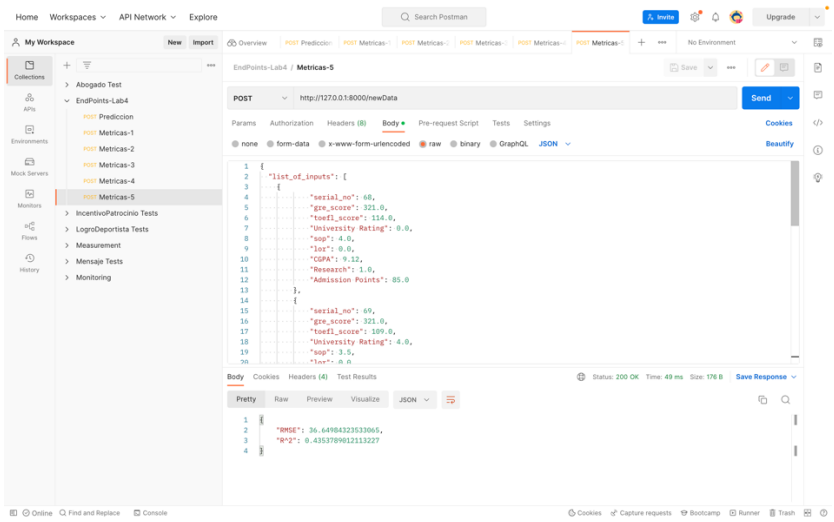
En esta prueba los datos no fueron los más adecuados ya que el RMSE es un valor muy grande comparado a el rango de los valores ingresados y el coeficiente de determinación no es tan alto como en los otros casos, aquí nos da 0.65. Lo que nos indica que no está prediciendo de la mejor manera.

5.


```
{
```

```
"list_of_inputs": [
  {
        "serial_no": 68,
        "gre_score": 321.0,
        "toefl_score": 114.0,
        "University Rating": 0.0,
        "sop": 4.0,
        "lor": 0.0,
        "CGPA": 9.12,
        "Research": 1.0,
        "Admission Points": 85.0
    },
    {
        "serial_no": 69,
        "gre_score": 321.0,
        "toefl_score": 109.0,
        "University Rating": 4.0,
        "sop": 3.5,
        "lor": 0.0,
        "CGPA": 8.35,
        "Research": 0.0,
        "Admission Points": 15.0
    },
    {
        "serial_no": 69,
        "gre_score": 321.0,
        "toefl_score": 109.0,
        "University Rating": 10.0,
        "sop": 3.5,
        "lor": 0.0,
        "CGPA": 8.35,
        "Research": 1.0,
        "Admission Points": 140.0
    },
    {
        "serial_no": 69,
        "gre_score": 321.0,
        "toefl_score": 109.0,
        "University Rating": 1.0,
        "sop": 3.5,
        "lor": 0.0,
        "CGPA": 8.35,
        "Research": 0.0,
        "Admission Points": 136.0
    },
    {
        "serial_no": 69,
        "gre_score": 321.0,
        "toefl_score": 109.0,
        "University Rating": 0.0,
        "sop": 3.5,
        "lor": 0.0,
```

```
            "CGPA": 8.35,
            "Research": 1.0,
            "Admission Points": 140.0
        }
    ]
}
```

En esta prueba los datos no fueron los más adecuados ya que el RMSE es un valor muy grande comparado a el rango de los valores ingresados y el coeficiente de determinación es más pequeño que el 50%. Lo que nos indica que no está prediciendo de la mejor manera. Esto creemos que sucede ya que los datos no siguen un patrón de características, valores con admission points similares, no tienen una puntuación de investigación y universidad similar.

Mitigación de incoherencias y errores de ejecución

Una posible estrategia para mitigar incoherencias constaría de comparar los resultados de todos los datos con los datos que se esperan y compararlos. Se colocaría un margen de aceptación según la raíz de error cuadrático (en este caso alrededor de 90%) y se haría un cálculo sencillo de diferencia entre ambos datos a forma de porcentaje de la siguiente forma:

$$\%\ de\ registro = \frac{|dato\ esperado - resultado|}{dato\ esperado} * 100$$

Este cálculo se les haría a todos los datos del csv original del laboratorio. Con el margen establecido se aislarían los registros por debajo del margen en otro DataFrame.

Al nuevo DataFrame se le haría una revisión de datos para identificar sus patrones, también se podría guardar en un nuevo archivo .csv para hacer un tablero en las herramientas Power BI o Tableau para identificar como están distribuidos los datos, por aparte hacer el mismo proceso con los datos que si pasaron el margen y ver cuáles son las diferencias más fuertes. De esta forma se podría identificar en que parte se encuentra la posible razón por la que los datos pueden presentar incoherencias y tratarla con limpieza de datos o comunicando al negocio las irregularidades identificadas.

Los errores de ejecución pueden tratarse con manejo de excepciones en la entrada de información en las peticiones POST, los errores en el formato JSON no son identificados por POSTMAN ni por la API así que habría que hacer un manejo amplio de excepciones o dejar instrucciones extremadamente claras de como son los formatos de ingreso.