

# DataSandBox - MinTIC



El futuro  
es de todos

Unidad para la atención  
y reparación integral  
a las víctimas

## Identificación de posibles casos de fraude en el RUV



	<b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas		
			Versión	1.0

## CONTENIDO

1	OBJETIVO .....	3
2	PROBLEMÁTICA .....	3
3	IMPLEMENTACIÓN MACHINE LEARNING CASOS TIPIFICADOS COMO POSIBLES DENUNCIAS ...	4
3.1	CONTEXTO DE LA SITUACIÓN .....	4
3.2	CONJUNTO DE DATOS .....	4
3.3	ALGORITMO SELECCIONADO .....	5
3.3.1	EXPERIMENTOS (IMPLEMENTACIÓN DEL MODELO).....	5
3.3.2	RENDIMIENTO DEL MODELO .....	7
3.3.3	PRUEBA DEL MODELO .....	7

	<b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas		
			Versión	1.0

## 1 OBJETIVO

Ejercicio de clúster a partir de los fraudes que ya fueron identificados en el Registro Único de Víctimas (RUV) y que ya fueron denunciados ante la fiscalía por la Oficina Asesora Jurídica (OAJ) de la Unidad. Comportamientos que son identificados como fraude. Buscar en la base de datos (identificación de patrones en el RUV) Aprendizaje de máquina para que identifique dentro del RUV posibles casos de fraude a partir de las características previamente identificadas. Se parte de 3000 registros para que el modelo “Aprenda” y luego correrlo sobre los 9.000.000 de registros de víctimas.

## 2 PROBLEMÁTICA

Esta iniciativa nace en base a la necesidad de poder identificar posibles casos de fraudes en el RUV. La entidad pretende realizar un proyecto de Big Data y analítica que permita cumplir con el objetivo propuesto, la realización de esta iniciativa demanda una cantidad de recursos tecnológicos que permitan la ejecución.

 <b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas			
			Versión	1.0

### 3 IMPLEMENTACIÓN MACHINE LEARNING CASOS TIPIFICADOS COMO POSIBLES DENUNCIAS

Para la implementación del Machine Learning se tomó como insumo la base de datos de la Oficina Asesora Jurídica de la unidad para las Víctimas (OAJ), se hizo el planteamiento de un escenario en particular, el cual se describirá a continuación.

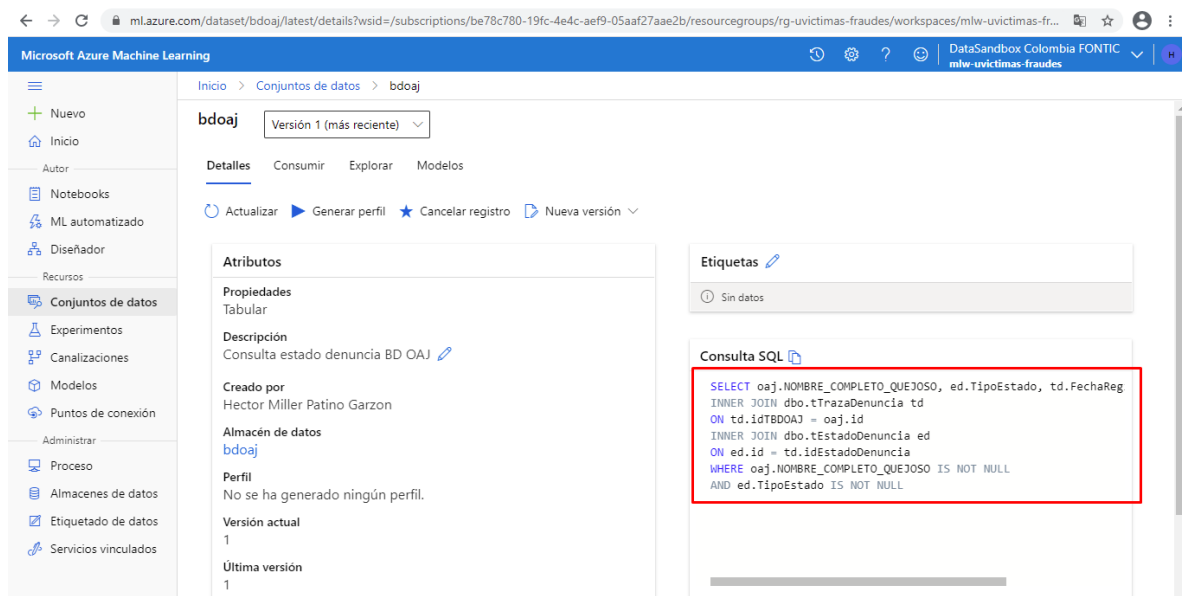
#### 3.1 CONTEXTO DE LA SITUACIÓN

La oficina Asesora Jurídica (OAJ) de la Unidad de Víctimas, entre otras funciones tiene la recibir los casos que son tipificados como posibles fraudes, dentro de dicha gestión, esta recibir la queja, y a su vez hacer la gestión correspondiente ante la entidad competente, para este caso particular es la Fiscalía General de la Nación.

Bajo este contexto, se hizo la implementación de un Machine Learning con el objeto de evaluar el comportamiento de las denuncias, interpuestas ante la Fiscalía dependiendo la etapa del proceso. Los resultados de esta implementación se describen a continuación:

#### 3.2 CONJUNTO DE DATOS

Para crear el conjunto de datos y hacer el entrenamiento del modelo, se hizo la construcción de una consulta a la base de datos implementada en el Azure DATASANDBOX de la OAJ, dicha consulta toma una muestra del nombre completo de la persona que interpone la queja, el estado de la denuncia, y la fecha de creación del registro:



The screenshot shows the Microsoft Azure Machine Learning interface for a dataset named 'bdoaj'. The 'Consultas SQL' section contains the following query:

```
SELECT oaj.NOMBRE_COMPLETO_QUEJOSO, ed.TipoEstado, td.FechaReg
INNER JOIN dbo.tTrazaDenuncia td
ON td.idTBDOAJ = oaj.id
INNER JOIN dbo.tEstadoDenuncia ed
ON ed.id = td.idEstadoDenuncia
WHERE oaj.NOMBRE_COMPLETO_QUEJOSO IS NOT NULL
AND ed.TipoEstado IS NOT NULL
```

 <b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas			
			Versión	1.0

### 3.3 ALGORITMO SELECCIONADO

De acuerdo con el conjunto de datos creados, el mejor algoritmo para hacer la implementación del Machine Learning es el algoritmo VotingEnsemble, el cual consiste en que cada uno de los clasificadores utilizados, realiza una predicción independiente y al final se selecciona la que ha sido clasificada por la mayoría.

#### 3.3.1 EXPERIMENTOS (IMPLEMENTACIÓN DEL MODELO)

A continuación, se describe el resumen del modelo y datos adicionales de su implementación:

Inicio > Experimentos > estadoDenuncia > Ejecución 1

**Ejecución 1** Completado  
Actualizar Cancelar

Detalles Límites de protección de datos Modelos Resultados y registros Ejecuciones secundarias Instantánea

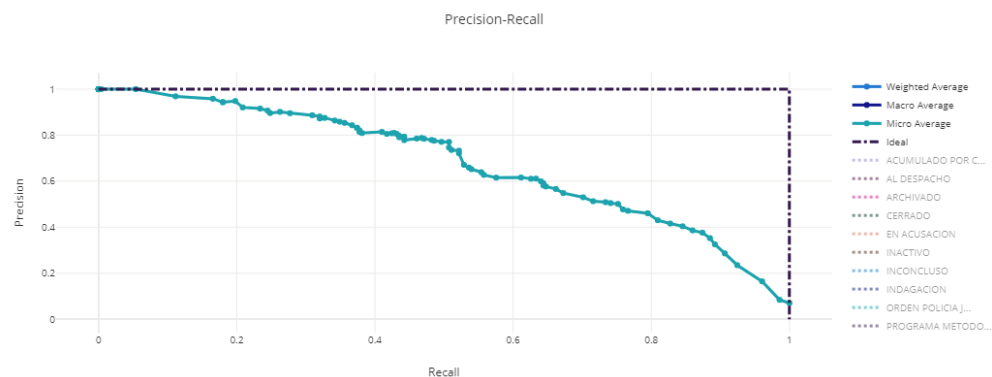
**Propiedades**  
Estado: Completado  
Creada: Mar 10, 2021 3:42 PM  
Iniciado: Mar 10, 2021 3:43 PM  
Duración: 32 m 1.188 s  
Destino de proceso: [cpu-cluster](#)  
Id. de ejecución: AutoML\_010e7904-eb45-4c27-a9f8-9fd1621940d0  
Nombre del script: --  
Creado por: Hector Miller Patino Garzon  
Conjuntos de datos de entrada: Nombre de entrada: training\_data; id: 403c52c1-e65a-45de-93be-ed5d6c152c21  
Conjuntos de datos de salida: Ninguno  
Argumentos: Ninguno

**Mejor resumen del modelo**  
Nombre del algoritmo: [VotingEnsemble](#)  
Precisión: 0.61941 [Ver todas las demás métricas](#)  
Muestreo: 100.00 % [O](#)  
Modelos registrados: [AutoML010e7904e46:1](#)  
Estado de la implementación: [denunciafraude](#) Correcto

**Resumen de ejecución**  
Tipo de tarea: Clasificación [Ver toda la configuración de la ejecución](#)  
Métrica primaria: Precisión  
Nombre del experimento: estadoDenuncia

**Descripción**

Las graficas que se muestran a continuación, indican la interpretación estadística para la precisión del modelo:





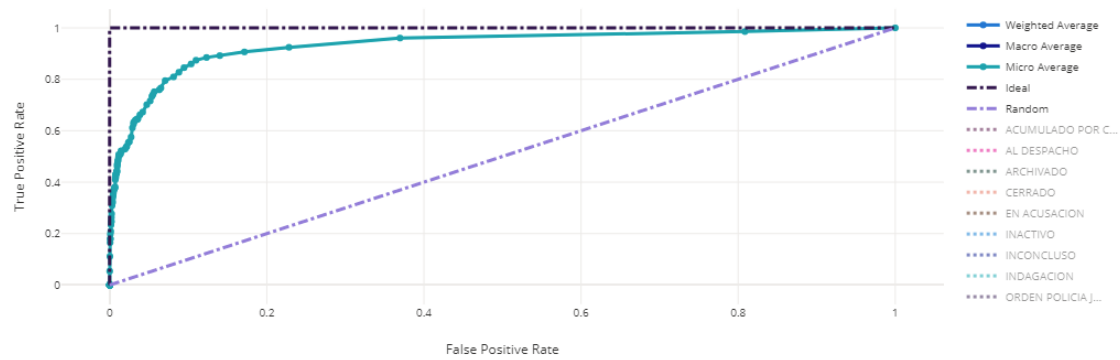
El futuro  
es de todos

Unidad para la atención  
y reparación integral  
a las víctimas

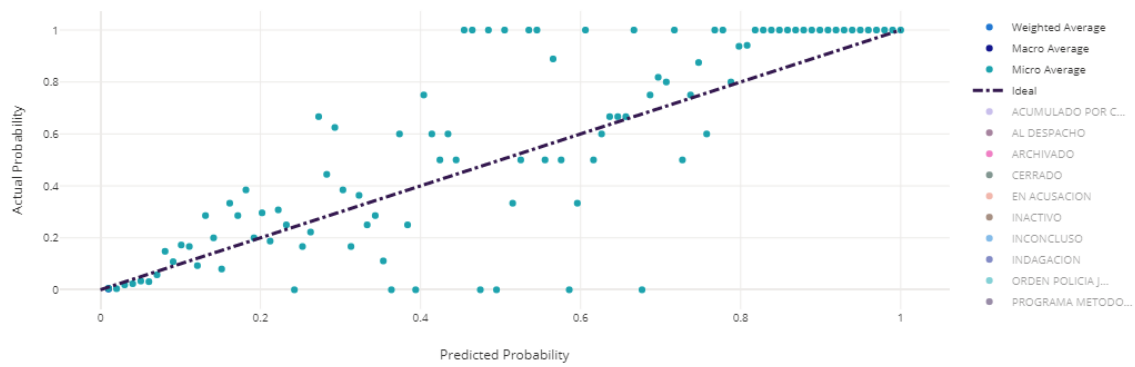
Versión

1.0

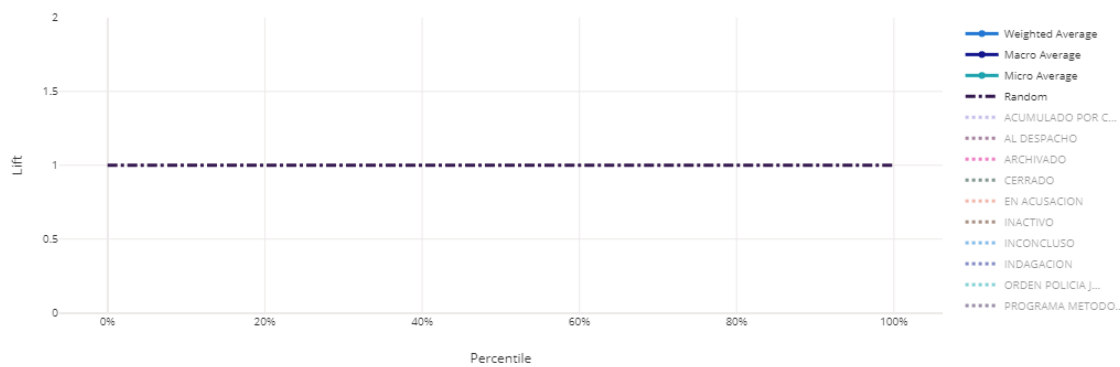
ROC

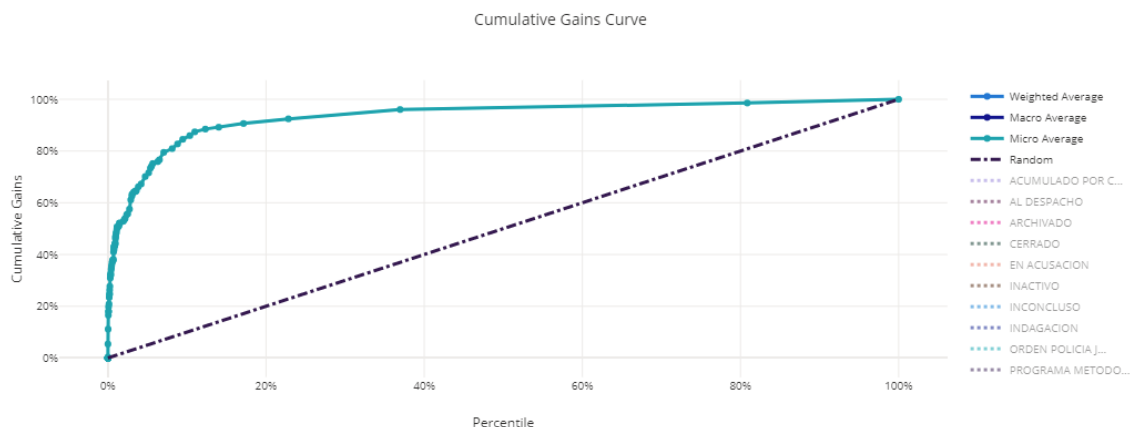


Calibration Curve



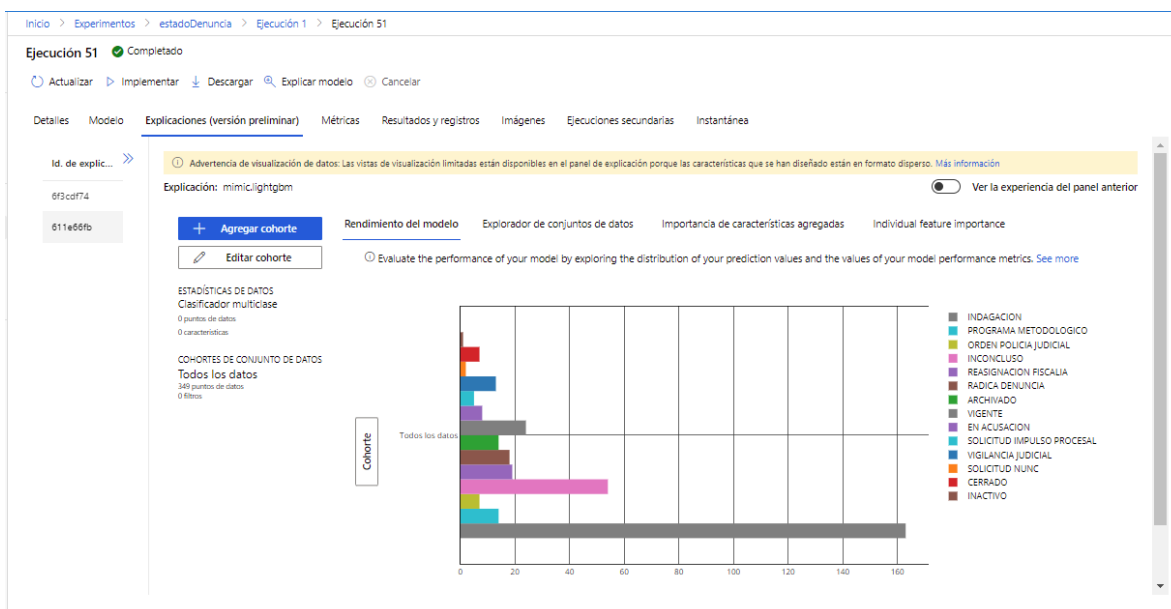
Lift Curve





### 3.3.2 RENDIMIENTO DEL MODELO

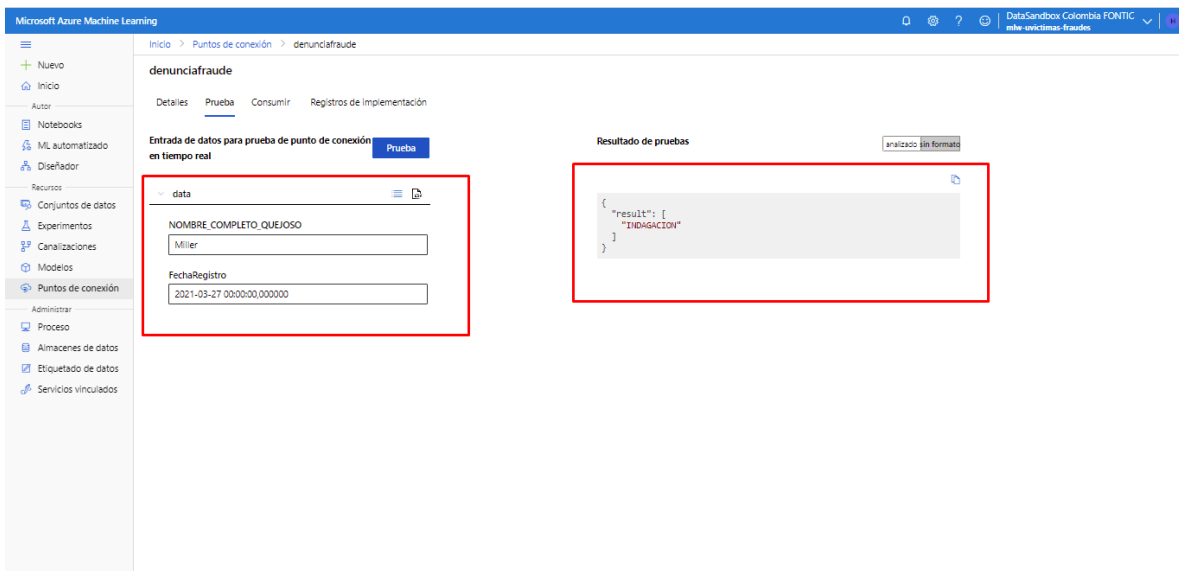
En la imagen que se muestra a continuación, se puede visualizar los datos con los cuales se realizó el entrenamiento del modelo:



### 3.3.3 PRUEBA DEL MODELO

Con el entrenamiento del modelo, pasamos a ingresar datos de prueba. Con este ejercicio el algoritmo nos indica cual es la predicción de un caso reportado como fraude:

 <b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas			
			Versión	1.0



Microsoft Azure Machine Learning

Inicio > Puntos de conexión > denunciafraude

denunciafraude

Detalles **Prueba** Consumir Registros de implementación

Entrada de datos para prueba de punto de conexión en tiempo real **Prueba**

data

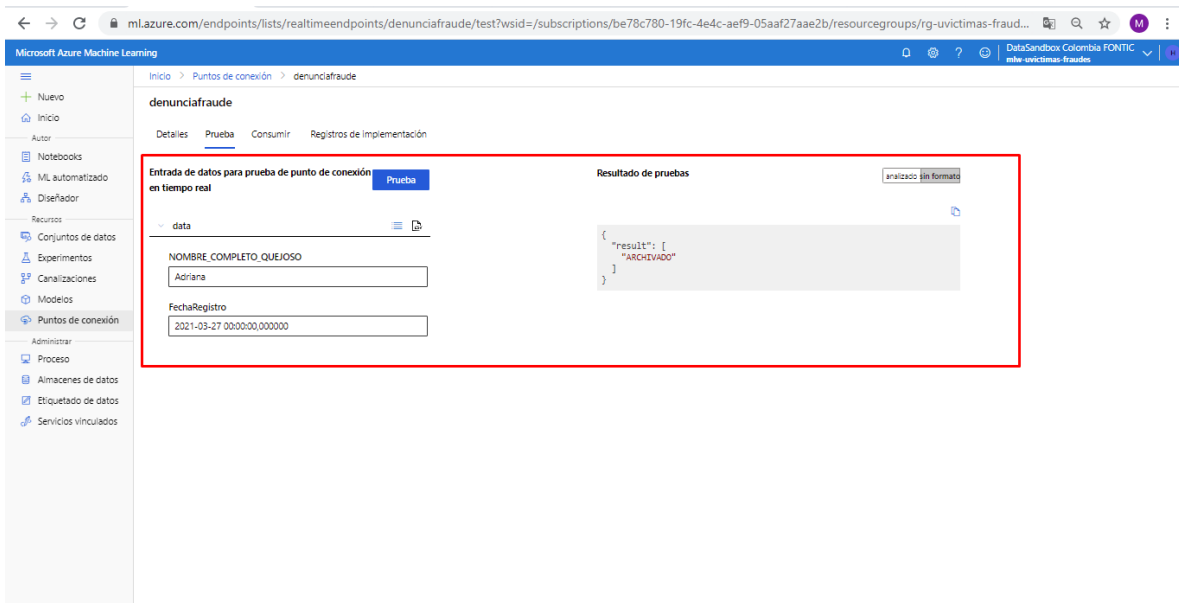
NOMBRE\_COMPLETO\_QUEJOSO  
Miller

FechaRegistro  
2021-03-27 00:00:00.000000

Resultado de pruebas

```
{
  "result": [
    "INDAGACION"
  ]
}
```

En resumen, el algoritmo basado en cierta cantidad de casos ingresados puede determinar cuál será la etapa en la que se encuentre durante el proceso de investigación, de acuerdo con la fecha de la radicación de la denuncia. La imagen que se muestra a continuación arroja un resultado después de haber ingresado cierta cantidad de datos:



Microsoft Azure Machine Learning

Inicio > Puntos de conexión > denunciafraude

denunciafraude

Detalles **Prueba** Consumir Registros de implementación

Entrada de datos para prueba de punto de conexión en tiempo real **Prueba**

data

NOMBRE\_COMPLETO\_QUEJOSO  
Adriana


FechaRegistro  
2021-03-27 00:00:00.000000

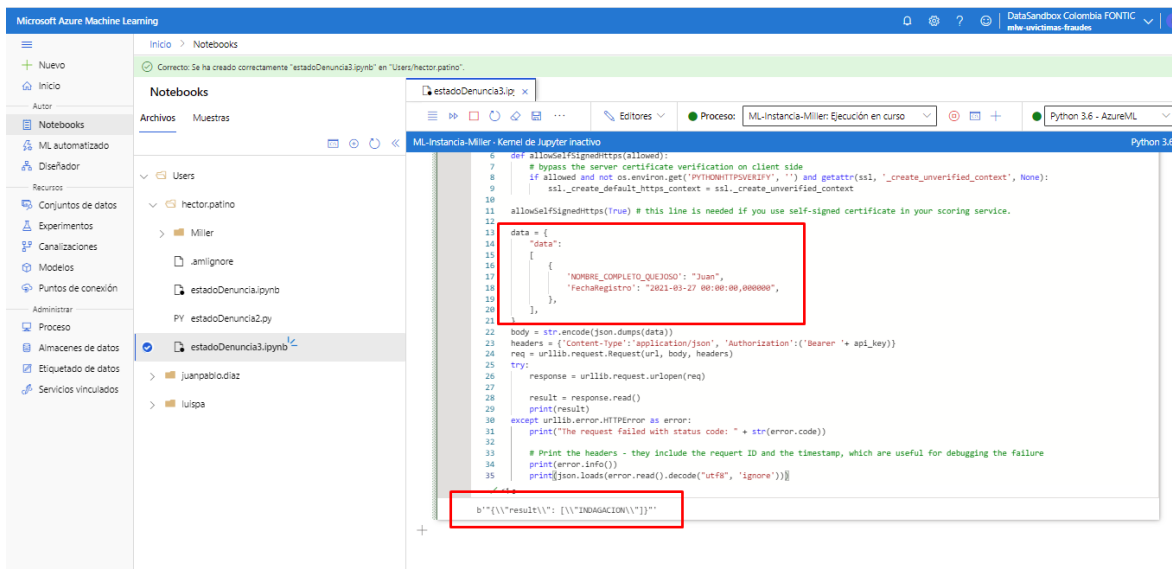
Resultado de pruebas

```
{
  "result": [
    "ARCHIVADO"
  ]
}
```

Para el consumo del modelo, se muestra como hacerlo por medio Notebooks haciendo uso de las librerías y código fuente de Python:



 <b>El futuro es de todos</b>	Unidad para la atención y reparación integral a las víctimas			
			Versión	1.0



```

6 # allow self-signed https (allowed):
7 # bypass the server certificate verification on client side
8 if allowed and not os.environ.get('PYTHONHTTPSVERIFY', '') and getattr(ssl, '_create_unverified_context', None):
9     ssl._create_default_https_context = ssl._create_unverified_context
10
11 allowSelfSignedHttps(True) # this line is needed if you use self-signed certificate in your scoring service.
12
13 data = {
14     "data":
15     {
16         "nombre_completo_quejoso": "Juan",
17         "fecha_registro": "2021-03-27 00:00:00.000000",
18     },
19 },
20
21 body = str.encode(json.dumps(data))
22 headers = {'Content-Type': 'application/json', 'Authorization': ('Bearer ' + api_key)}
23 req = urllib.request.Request(url, body, headers)
24 try:
25     response = urllib.request.urlopen(req)
26     result = response.read()
27     print(result)
28 except urllib.error.HTTPError as error:
29     print("The request failed with status code: " + str(error.code))
30     # Print the headers - they include the request ID and the timestamp, which are useful for debugging the failure
31     print(error.info())
32     print(json.loads(error.read().decode("utf-8", 'ignore'))))
33
34 b'{"result": [{"INVESTIGACION"}]}'

```

## 4 CONCLUSIONES

Como proyecto piloto de inmersión en la plataforma DataSandBox de MinTIC, se ha logrado alcanzar un resultado de aplicación de tecnologías de Machine Learning sobre un conjunto de datos base aportados por la Oficina Asesora Jurídica de la Unidad para la construcción de un modelo que le permitiera con estos, aprender y poder generar respuesta ante planteamientos de valoración de casos catalogados como posibles fraudes y cual sería su posible estado resultado ante la gestión de la Fiscalía.

Esta experiencia a con llevado no sólo al conocimiento en Machine Learning, sino otras actividades relacionadas que se deben procesar previamente como la gestión en DataFactory para la creación de ETLs y otros artefactos para el cargue de la información en la base de datos SQL Azure de forma que garantice el correcto flujo y uso a través del modelo de ML implementado.

Es de recalcar que el espacio y oportunidad dispuesto por MinTIC a través del DataSandBox ha sido de vital importancia para promover como más relevancia e importancia la transformación digital y la implementación de nuevas tecnologías basadas en la 4ta Revolución Industrial como son Machine Learning, Inteligencia Artificial y otras más, vitales para inmersión de la Unidad en estos aspectos claves y estratégicos en pro de un mejoramiento continuo en la atención y reparación integral para las víctimas.