



MAESTRIA EN CIENCIAS CON MENCIÓN EN TECNOLOGÍA DE LA INFORMACIÓN

Semestre Académico 2023 – II

CURSO:

MACHINE LEARNING

Trabajo final- CASO-CHICAGO

Docente : Ing. Dr. Juan Orlando Riascos Armas

**Maestranes : David Melvin Requejo Santa Cruz
Jhon Erikson Melendrez Huaman**

Octubre 2023 – Tarapoto



1. Descripción y Objetivo:	3
a. Caso de Crímenes en Chicago	3
2. Objetivo	3
3. Evaluación los algoritmos (turbo pre - Automodel)	3
a. Pasos Para Realizar el Autodel	3
4. Entender el negocio	5
a. Contexto y Origen de los Datos:	5
b. Privacidad y Anonimato:	6
c. Naturaleza de los Datos:	6
5. Entender los datos	6
a. Descripción de las Variables:	6
b. Definición de las variables a utilizar	7
c. Definición del tipo de algoritmo	7
6. Preparación de los datos	7
7. Construcción del modelo	8
a. Carga de la data al RapidMiner	8
b. Procedimiento del para el entrenamiento	8
8. Evaluación	10
9. Despliegue	10

1. Descripción:

a. Caso de Crímenes en Chicago

Este conjunto de datos refleja los incidentes de delitos reportados (con excepción de los asesinatos donde existen datos para cada víctima) que ocurrieron en la ciudad de Chicago desde 2001 hasta el presente, menos los siete días más recientes. Los datos se extraen del sistema CLEAR (Análisis e informes de aplicación de la ley ciudadana) del Departamento de Policía de Chicago. Para proteger la privacidad de las víctimas de delitos, las direcciones se muestran únicamente a nivel de bloque y no se identifican ubicaciones específicas. Estos datos incluyen informes no verificados proporcionados al Departamento de Policía. Las clasificaciones preliminares de delitos pueden cambiarse en una fecha posterior basándose en una investigación adicional y siempre existe la posibilidad de que se produzca un error mecánico o humano. Por lo tanto, el Departamento de Policía de Chicago no garantiza (ya sea expresa o implícita) la exactitud, integridad, puntualidad o secuenciación correcta de la información y la información no debe usarse con fines de comparación a lo largo del tiempo.

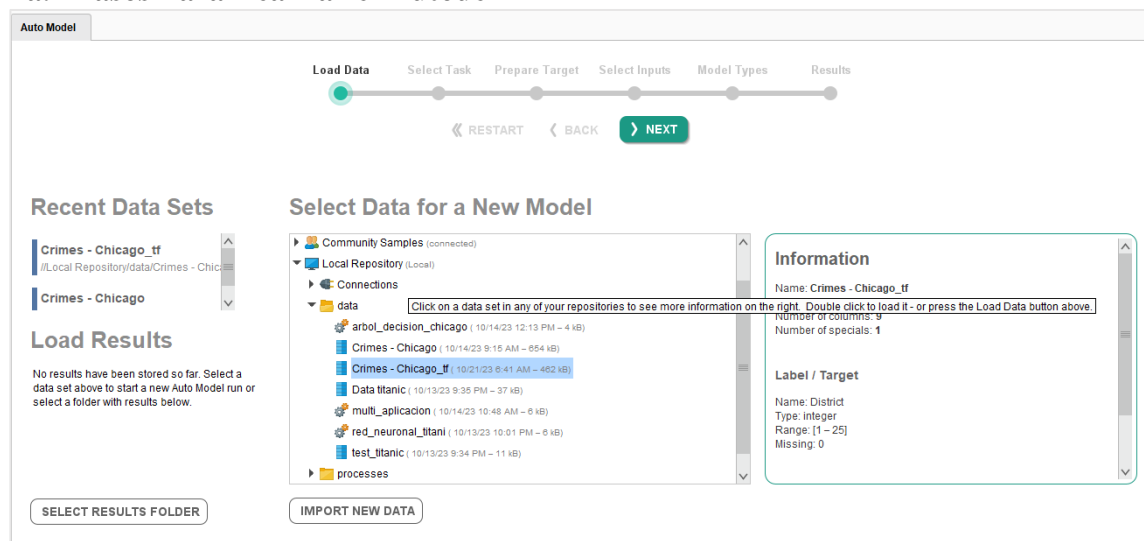
Data : <https://drive.google.com/file/d/14FI2uQKuR0Vt7e34-ac-nEoqnhPMM1km/view>

2. Objetivo

Desarrollar un modelo de predicción de delitos en la ciudad de Chicago utilizando datos históricos de delitos.

3. Evaluación los algoritmos (turbo pre - Automodel)

a. Pasos Para Realizar el Autodel



Paso 1: Seleccionamos la data que deseamos, Esta puede ser desde nuestro ordenador o desde un servidor

Load Data Select Task Prepare Target Select Inputs Model Types Results

« RESTART < BACK > NEXT

Predict
Want to predict the values of a column?

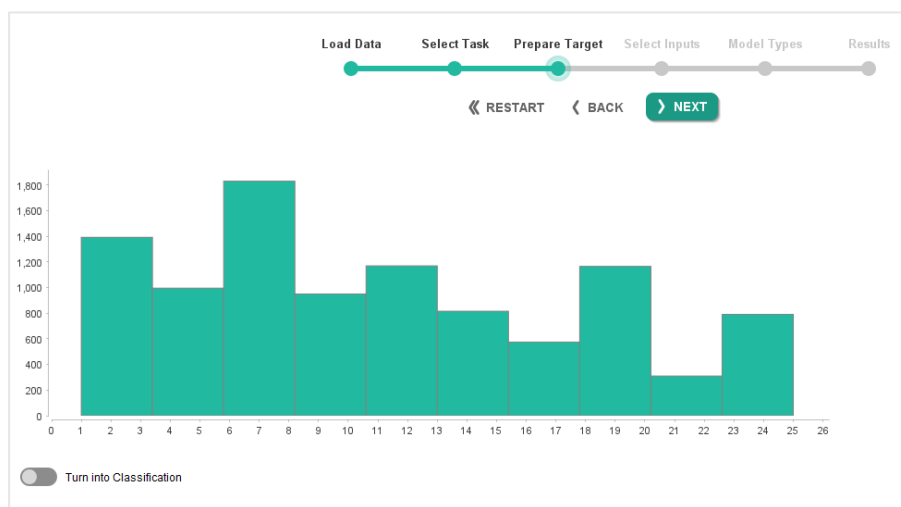
Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

Date / Time	IUCR Number	Primary Type Category	Location Description	Arrest Category	Domestic Category	Ward Number	Community Area	FBI Code Category	District Number
?	266	CRIMINAL SEXUA...	RESIDENCE	true	true	8	45	2	4
Jan 10, 2020	1753	OFFENSE INVOLV...	RESIDENCE	false	true	14	63	2	9
Apr 9, 2020	1754	OFFENSE INVOLV...	RESIDENCE	false	true	9	49	2	5
Jun 8, 2020	495	BATTERY	RESIDENCE	false	true	18	70	04B	8
?	486	BATTERY	PARK PROPERTY	true	true	40	2	08B	24
?	1477	WEAPONS VIOLA...	RESIDENCE	false	false	38	15	15	16
Jan 5, 2020	1710	OFFENSE INVOLV...	RESIDENCE	false	true	9	49	20	5
Nov 6, 2020	1153	DECEPTIVE PRA...	RESIDENCE	false	false	10	54	11	5

9,988 rows, 10 columns (5 nominal, 3 numerical, 1 date)

Paso 2: Visualización de los datos y se selecciona la columna sobre la que se quiere que trabaje.



Paso 3: Reportes previos (Visualización por distritos)

Auto Model

Load Data Select Task Prepare Target Select Inputs Model Types Results

« RESTART < BACK > RUN

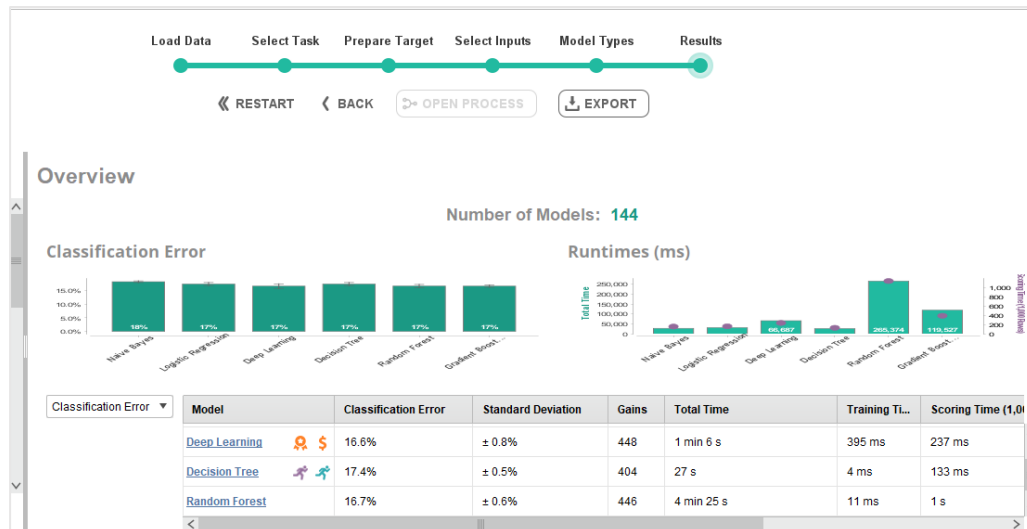
Models

- ☒ Naive Bayes
- ☐ Generalized Linear Model
 - ☒ Use Regularization
 - ☐ Calculate p-Values
- ☒ Logistic Regression
- ☐ Fast Large Margin
 - ☒ Automatically Optimize
- ☒ Deep Learning
- ☒ Decision Tree
 - ☒ Automatically Optimize
 - Maximal Depth: 20

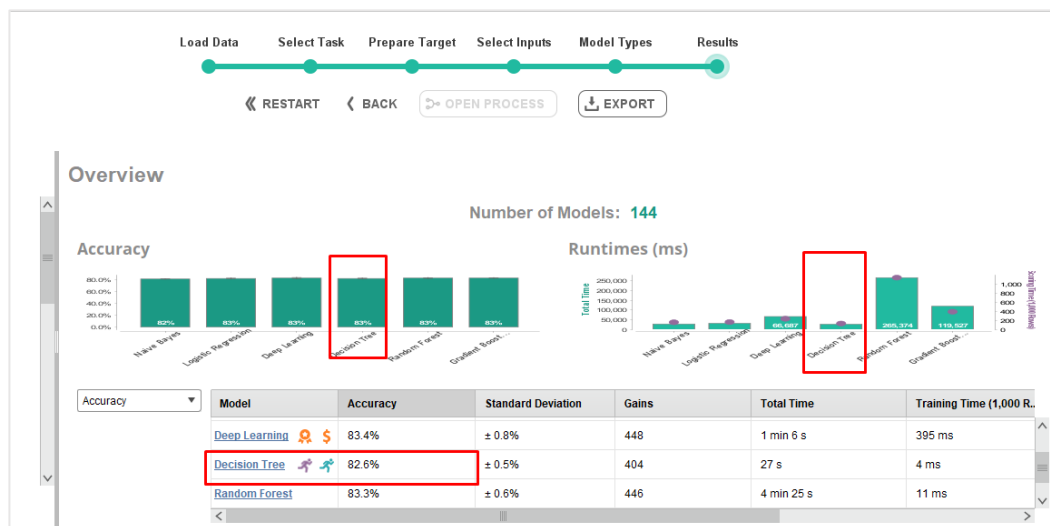
Data Preparation

- ☒ Remove Columns with Too Many Values
 - Maximum Number of Values: 50
- ☒ Extract Date Information
- ☐ Extract Text Information
 - Select Text Columns (0) ...
- Number of Extracted Features: 1,000
- ☐ Automatic Feature Selection
 - Additional Minutes (Maximum): 60
 - Final Feature Set should be: Accurate
- ☐ Automatic Feature Generation

Paso 4: Seleccionamos los modelos que deseamos que el auto modelo analice



Paso 5: En esta imagen se visualiza el error en los diferentes modelos que realizo el entrenamiento.



Paso 6: Para el siguiente Trabajo vamos a tomar el modelo que tiene como Accuracy 82.6%

4. Entender el negocio.

a. Contexto y Origen de los Datos:

El proyecto se enfoca en el análisis de datos que reflejan incidentes de delitos reportados en la ciudad de Chicago desde el año 2001 hasta la fecha actual, excluyendo los siete días más recientes. Estos datos provienen del sistema CLEAR (Análisis e informes de aplicación de la ley ciudadana) del Departamento de Policía de Chicago. El objetivo principal de esta iniciativa es comprender y utilizar esta información para mejorar la seguridad pública y la toma de decisiones relacionadas con la aplicación de la ley.

b. Privacidad y Anonimato:

Para salvaguardar la privacidad de las víctimas de delitos, los datos se han procesado de manera que las direcciones se muestran únicamente a nivel de bloque, sin identificar ubicaciones específicas.

c. Naturaleza de los Datos:

Es importante destacar que los datos pueden contener informes no verificados, lo que significa que algunas de las entradas pueden no haber pasado por un proceso de verificación o validación exhaustiva. Además, las clasificaciones preliminares de delitos están sujetas a cambios posteriores, basados en investigaciones adicionales. Por lo tanto, se debe tener en cuenta que la precisión de los datos no está garantizada.

5. Entender los datos

a. Descripción de las Variables.

Case Number	Número de caso
Date	Fecha en que ocurrió el incidente
Block	Descripción del bloque
IUCR	Código de Referencia Uniforme de Incidentes
Primary Type	Tipo principal de delito
Description	Descripción detallada del incidente
Location Description	Descripción de la ubicación
Arrest	Indica si hubo arresto (verdadero o falso)
Domestic	Indica si el incidente involucra una relación doméstica (verdadero o falso)
Beat	Número o código de unidad policial que patrulla la zona
District	Distrito policial en el que ocurrió el incidente
Ward	Número o código del pabellón o distrito electoral
Community Area	Área comunitaria en la que ocurrió el incidente
FBI Code	Código del FBI que clasifica el tipo de incidente
X Coordinate	Coordenada X
Y Coordinate	Coordenada Y
Year	Año en que ocurrió el incidente
Updated On	Fecha de actualización de los datos
Latitude	Latitud de la ubicación del incidente
Longitude	Longitud de la ubicación del incidente
Location	Coordenadas geospaciales de la ubicación

b. Definición de las variables a utilizar

Nombre	Descripción	Tipo
Location Description	Descripción de la ubicación	Polynomial
Arrest	Indica si hubo arresto (verdadero o falso)	Binomial
Domestic	Indica si el incidente involucra una relación doméstica (verdadero o falso)	Binomial
District	Distrito policial en el que ocurrió el incidente	Integer
Community Area	Área comunitaria en la que ocurrió el incidente	Integer
FBI Code	Código del FBI que clasifica el tipo de incidente	Polynomial

c. Definición del tipo de algoritmo

El algoritmo que se va utilizar para el entrenamiento de los datos es modelo de Árboles de decisión ya que este se adapta a la naturaleza de los datos de incidentes de delitos en Chicago además en el auto-model tiene un Accuracy 82.6%

6. Preparación de los datos.

La preparación y limpieza de datos en **Excel** es una etapa crucial, ya que nos permite seleccionar y depurar los datos relevantes, lo que simplifica el proceso en RapidMiner al llevar únicamente las columnas necesarias.

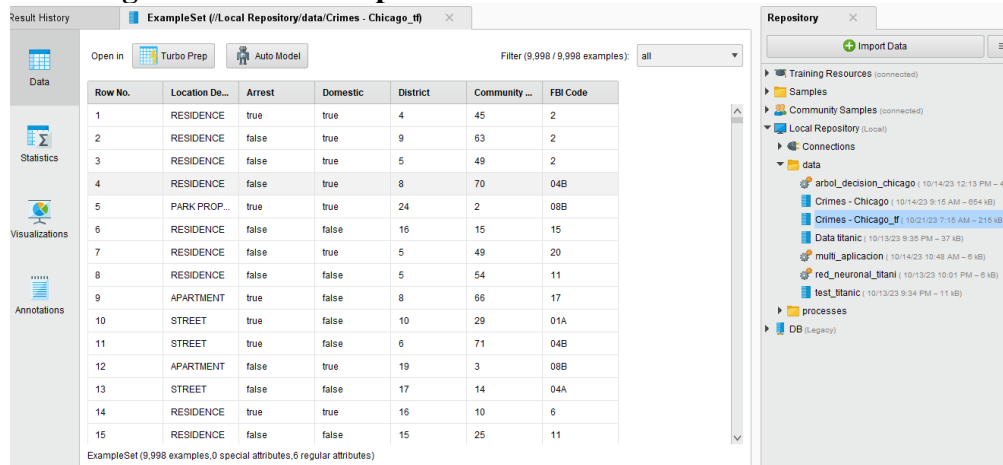
En RapidMiner, el enfoque se centra en asignar el tipo de variable adecuado a cada columna, lo que facilita la interpretación y el procesamiento de los datos. Esto asegura que trabajemos con información consistente y coherente, lo que es esencial para la construcción exitosa de nuestro modelo de Árboles de Decisión y la obtención de información significativa de los datos de incidentes de delitos en Chicago.

Format your columns.						
<input type="checkbox"/> Replace errors with missing values ⓘ						
	Location D...	Arrest	Domestic	District	Community...	FBI Code
	polynomial	binomial	binomial	integer	integer	polynomial
3	RESIDENCE	false	true	5	49	2
4	RESIDENCE	false	true	8	70	04B
5	PARK PROPERTY	true	true	24	2	08B

Asignación del tipo de variable a cada columna

7. Construcción del modelo.

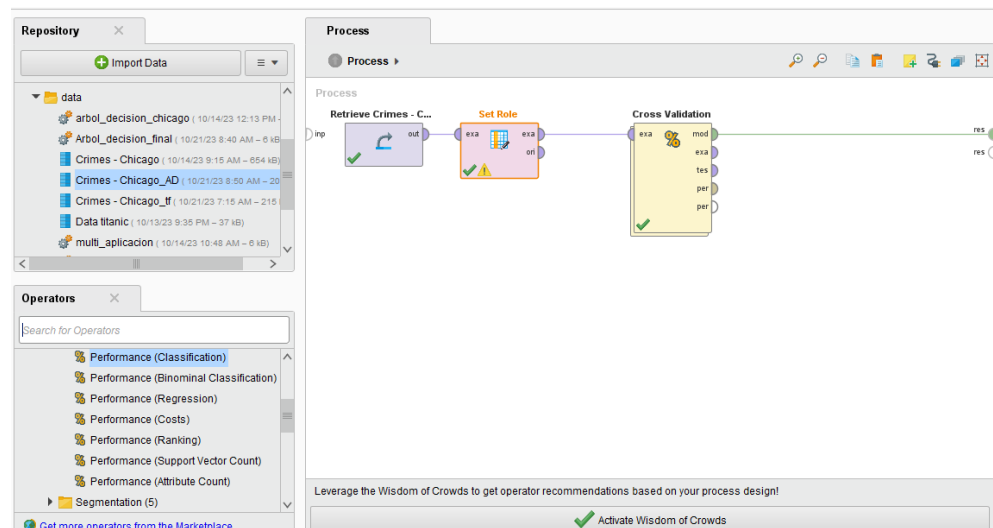
a. Carga de la data al RapidMiner.



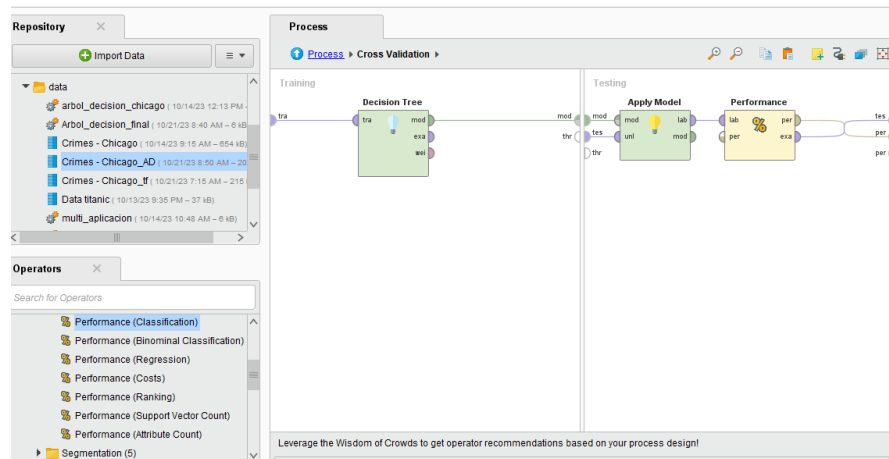
Row No.	Location De...	Arrest	Domestic	District	Community ...	FBI Code
1	RESIDENCE	true	true	4	45	2
2	RESIDENCE	false	true	9	63	2
3	RESIDENCE	false	true	5	49	2
4	RESIDENCE	false	true	8	70	04B
5	PARK PROP...	true	true	24	2	08B
6	RESIDENCE	false	false	16	15	15
7	RESIDENCE	false	true	5	49	20
8	RESIDENCE	false	false	5	54	11
9	APARTMENT	true	false	8	66	17
10	STREET	true	false	10	29	01A
11	STREET	true	false	6	71	04B
12	APARTMENT	false	true	19	3	08B
13	STREET	false	false	17	14	04A
14	RESIDENCE	true	true	16	10	6
15	RESIDENCE	false	false	15	25	11

Se realiza la carga de la data limpia, además de poder ver la data RapidMiner también nos muestra opciones como estadísticas y Dashboards con la intención de mejorar el entendimiento de la data.

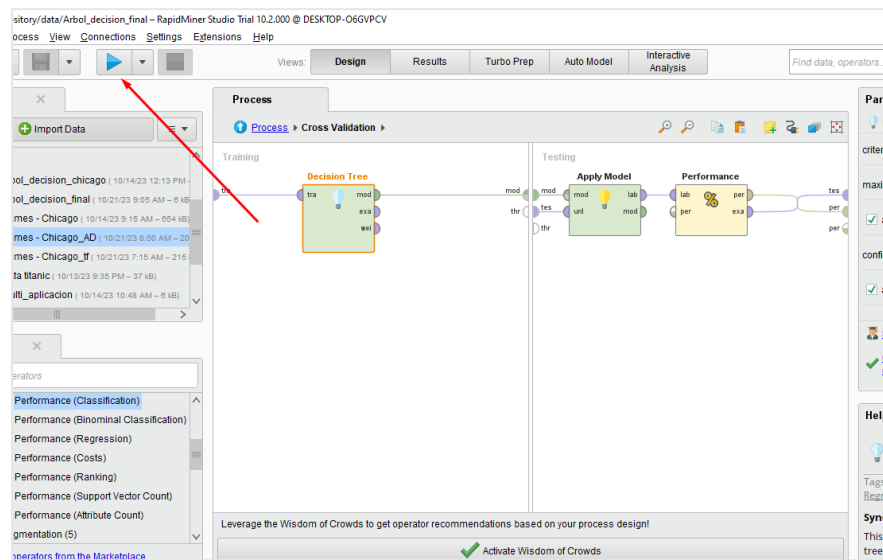
b. Procedimiento del para el entrenamiento.



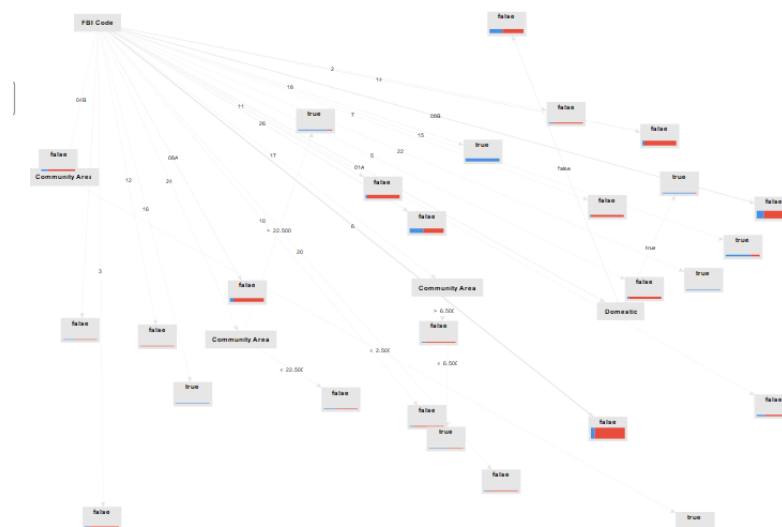
- Del repositorio arrastramos la data a la ventana en blanco.
- De Operaciones buscamos el set Rule, este nos permitirá poner cambiar la variable a tipo label.
- De Operaciones buscamos el Cross Validation, este albergara la el modelo.



Estando dentro del Cross Validation buscamos el modelo que deseamos en Operaciones en este caso Decisión Tree, asimismo el apply model. Y por ultimo el Performance, este nos permitirá ver la matriz de confusión



Finalmente Guardamos y corremos el modelo, obteniendo como resultado el siguiente resultado.





8. Evaluación.

accuracy: 82.44% +/- 3.66% (micro average: 82.44%)			
	true true	true false	class precision
pred. true	722	79	90.14%
pred. false	1677	7520	81.77%
class recall	30.10%	98.96%	

9. Despliegue.