



INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO



DATA MINING

Proyecto Final

Equipo #4:

- García Cruz Octavio Arturo
- Sampayo Hernández Mauro
- Flores Ponce Alan Marcelo

Grupo:

3CV15

Profesora:

Fabiola Ocampo Botello

Fecha:

20 de junio de 2023

Índice

CONJUNTOS DE DATOS	1
i. FUENTES DE DATOS	1
ii. DESCRIPCIÓN DE DATOS	2
iii. TRATAMIENTO DE DATOS.	5
I. CLASIFICACIÓN DE ARBOLES:	22
I.1 Descripción del ejercicio	22
I.2 Diccionarios de Datos.	22
I.2.1 Diccionario de Datos CART	22
I.2.2 Diccionario de datos C4.5	23
I.3 Resultados	30
I.4 Análisis de los resultados	33
II. Multi Clasificación. Bagging y Boosting	36
II.1 Descripción del ejercicio	36
II.2 Diccionario de datos	36
II.2.1 Diccionario de datos Random Forest.	36
II.2.3 Diccionario de datos Gradient Boost.	37
II.3. Resultados	38
II.4. Análisis de los resultados.	42
III. Agrupamiento (Jerárquico y No Jerárquico)	45
III.1 Descripción del ejercicio	45
III.2. Diccionario de Datos	45
III.2.1 Diccionario de Datos Jerárquico	45
III.2.2 Diccionario No Jerárquico	46
III.3. Resultados.	47
III.4 Análisis de Resultados	52
IV. Reglas de Asociación	54
IV.1 Descripción del ejercicio	54
IV.2 Diccionario de Datos	54
IV.3 Resultados	55
IV.4 Análisis de Resultados	56
V. Regresión Lineal	58
V.1 Descripción del ejercicio	58
V.2 Diccionario de Datos.	58
V.3 Resultados	58
V.4 Análisis de Resultados	59

INSTITUTO POLITÉCNICO NACIONAL ESCUELA SUPERIOR DE CÓMPUTO

Unidad de Aprendizaje: Minería de datos

Ciclo escolar: 2023–2

Proyecto Final

CONJUNTOS DE DATOS

i. FUENTES DE DATOS

1. **Conjunto de datos 1.** *Afluencia en estaciones del STC de la CDMX.*

El conjunto de datos que se usara es proporcionado por el Gobierno de la Ciudad de México el cual proporciona información sobre la afluencia diaria en diferentes estaciones del sistema de transporte colectivo (Metro).

A su vez usaremos un segundo conjunto de datos que nos proporciona los ingresos percibidos por cada día en cada estación del Metro de la ciudad de México.

Link del conjunto de datos *afluencia*:

<https://datos.cdmx.gob.mx/ne/dataset/afluencia-diaria-del-metro-cdmx>

Autor: SEMOVI (Secretaría de Movilidad)

Link del conjunto de datos *ingresos*:

<https://datos.cdmx.gob.mx/ne/dataset/ingresos-del-sistema-de-transporte-colectivo-metro>

Autor: SEMOVI (Secretaría de Movilidad)

2. **Conjunto de datos 2.** *FIFA World Cup Attendance 1930-2022*

El conjunto presentado en el siguiente enlace presenta la asistencia de aficionados a la Copa Mundial de la FIFA desde 1930.

Link del conjunto:

<https://www.kaggle.com/datasets/rajkumarpandey02/fifa-world-cup-attendance-19302022>

Autor: Raj Kumar Pandey

3. **Conjunto de datos 3.** *Credit Cards Approvals*

En este conjunto de datos encontraremos las características de personas que hicieron una petición para tramitar una tarjeta de crédito, y si la misma les fue aprobada o no.

Link del conjunto:

<https://www.kaggle.com/datasets/samueltcortinhas/credit-card-approval-clean-data>

Autor: Samuel Cortinhas

ii. DESCRIPCIÓN DE DATOS

Conjunto de Afluencia

Este conjunto está formado por un total de 955,305 registros; A continuación, se describe el significado de cada columna en el conjunto de datos:

Nombre	Significado	Tipo	Dominio
Fecha	La fecha en la que se registró la afluencia.	Date	01/01/2010 31/05/2023
Año	La fecha en la que se registró la afluencia.	Numérico	2010 - 2023
Mes	El mes correspondiente a la fecha.	Categorico	Enero - Diciembre
Línea	La línea del sistema de transporte público	Categorico	Línea 1 – Línea 12 Línea A Línea B
Estación	El nombre de la estación donde se registró la afluencia	Categorico	Nombre de la estación
Afluencia	El número de personas que utilizaron esa estación en particular en el día especificado	Numérico	Numero entero Positivo

Conjunto de Ingresos

Este conjunto está formado por un total de 151,260 registros; A continuación, se describe el significado de cada columna en el conjunto de datos:

Nombre	Significado	Tipo	Dominio
Fecha	La fecha en la que se registró la afluencia.	Date	01/01/2010 - 31/05/2023
Tipo_ingreso	El tipo de ingreso registrado	Categorico	Forma de pago
Ingreso	El valor numérico que representa la cantidad de ingresos registrados para esa línea en particular en el día especificado.	Numérico	Numero entero positivo
Línea	La línea del sistema de transporte público.	Categorico	Línea 1 – Línea 12 Linea A Linea B

Conjunto de FIFA World Cup Attendance

Nombre	Significado	Tipo	Dominio
Game(s)	Última fase a la que llegó el equipo anfitrión, resultado y equipo con el que se enfrentó	Texto	Texto
Venue	Estadio de inauguración y clausura	Categorico	Nombre del estadio sede
Number	Número de asistencia en el torneo	Numérico	Número no especificado
Host	Anfitrión de la Copa del Mundo	Categorico	Sede elegida por la FIFA
Total_Attendance	Asistencia total de personas	Numérico	Numero entero Positivo
Year	El año que se disputó el mundial	Numérico	1930 – 2022
Average_Attendance	Promedio de personas con asistencia en vivo	Numérico	Numero entero Positivo
Matches	Número total de partidos Jugados	Numerico	Numero entero Positivo

Conjunto de Credit Cards Approvals

Nombre	Significado	Tipo	Dominio
Genders	Genero de la persona	Categórico	Masculino Femenino
Age	Edad del solicitante	Numérico	0-99
Debt	Deuda pendiente (la característica se ha escalado)	Numérico	0-10
Married	Estado civil	Numérico	0- Soltero, divorciado, etc. 1- Casado
BankCustomer	Cliente del banco	Numérico	0- No tiene cuenta 1- Tiene cuenta
Industry	Sector de la industria en el que trabaja	Categórico	Nombre del sector
Ethnicity	Etnia del solicitante	Categórico	Asiático, Latino, Blanco, Negro, etc.
YearsEmployed	Años trabajando	Numérico	0-30
PriorDefault	Valor predeterminado con anterioridad.	Numérico	0- Sin valor 1- Con valor
Employed	Estado laboral	Numérico	0- Con empleo 1- Desempleado
CreditScore	Puntaje de crédito (esta función se ha escalado)	Numérico	0 – 99
DriversLicense	Licencia de conducir	Numérico	0- No tiene licencia 1- Tiene licencia
Citizen	Ciudadanía del solicitante	Categórico	Por nacimiento Por otras vías
ZipCode	Código postal	Numérico	00000 – 00XXX
Income	Ingresos del solicitante (previamente escalados)	Numérico	0 – 999,999
Approved	Aprobación de expedición de tarjeta	Numérico	0- No aprobado 1- Aprobado

iii. TRATAMIENTO DE DATOS.

Tratamiento de datos Ingresos

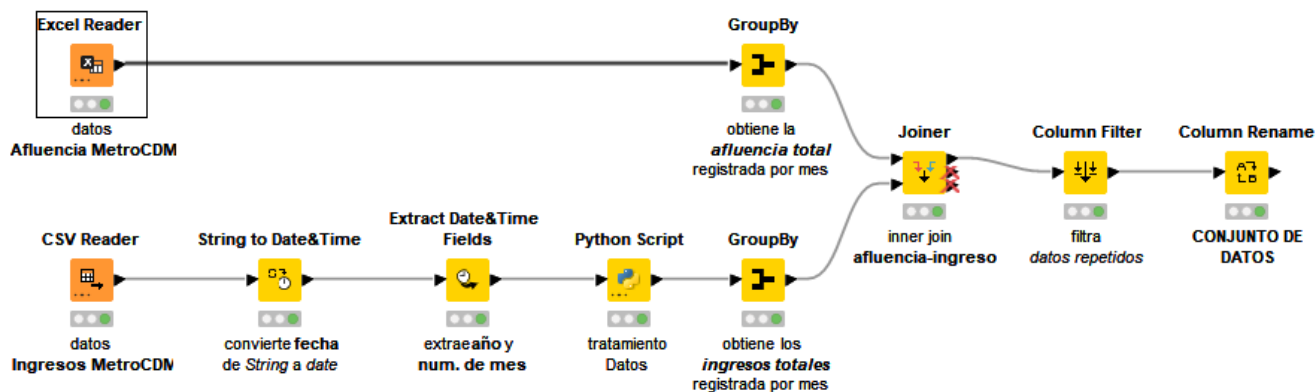
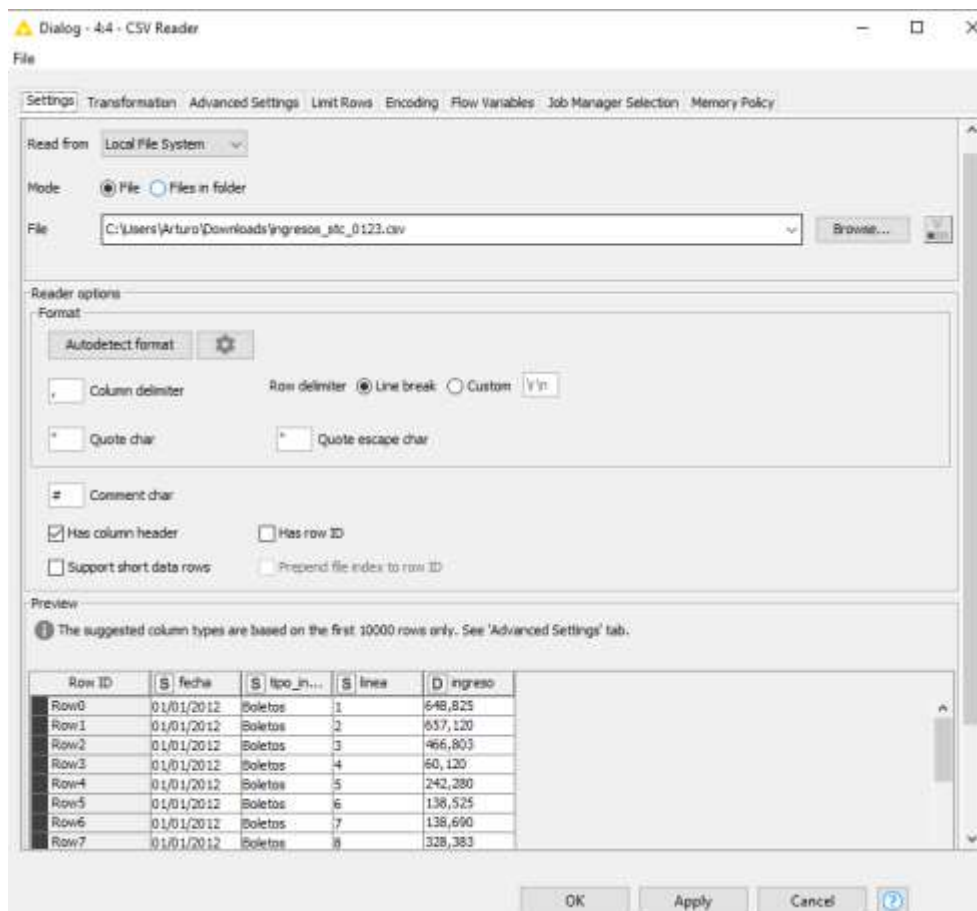


Diagrama de Tratamiento de datos

- 1) Cargaremos nuestro conjunto de datos (ingresos_stc_0123).



- 2) Nuestro conjunto de datos pasara por el nodo String Date & Time para cambiar el formato de la fecha.

Table "default" - Rows: 151260					Spec - Columns: 4	Properties	Flow Variables
Row ID	fecha	tipo_in...	linea	ingreso			
Row0	2012-01-01	Boletos	1	648,825			
Row1	2012-01-01	Boletos	2	657,120			
Row2	2012-01-01	Boletos	3	466,803			
Row3	2012-01-01	Boletos	4	60,120			
Row4	2012-01-01	Boletos	5	242,280			
Row5	2012-01-01	Boletos	6	138,525			
Row6	2012-01-01	Boletos	7	138,690			
Row7	2012-01-01	Boletos	8	328,383			
Row8	2012-01-01	Boletos	9	176			
Row9	2012-01-01	Boletos	A	275,370			
Row10	2012-01-01	Boletos	B	472,665			
Row11	2012-01-01	Boletos	12	NaN			
Row12	2012-01-02	Boletos	1	1,532,940			
Row13	2012-01-02	Boletos	2	1,491,780			
Row14	2012-01-02	Boletos	3	1,168,590			
Row15	2012-01-02	Boletos	4	165,690			
Row16	2012-01-02	Boletos	5	437,025			
Row17	2012-01-02	Boletos	6	286,620			
Row18	2012-01-02	Boletos	7	433,323			
Row19	2012-01-02	Boletos	8	777,951			
Row20	2012-01-02	Boletos	9	504			
Row21	2012-01-02	Boletos	A	514,455			
Row22	2012-01-02	Boletos	B	965,385			
Row23	2012-01-02	Boletos	12	NaN			
Row24	2012-01-03	Boletos	1	1,398,639			

Dialog - Job - String to Date&Time (convierte fecha)

File

Options Flow Variables Job Manager Selection Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

tipo_ingreso
linea

☐ Enforce exclusion

Include

Filter

fecha

☒ Enforce inclusion

Replace/Append Selection

☐ Append selected columns Suffix of appended columns: (Date&Time)

☒ Replace selected columns

Type and Format Selection

New type: Date Date format: MM/YYYY

Locale: en-US Content of the first cell: 01/01/2012

Guess data type and format

Abort Execution

☒ Fail on error

OK Apply Cancel

- 3) Aplicaremos el no extract date & time el cual se encargará de extraer el año y el mes y los pondrá en columnas separadas.

Table "default" - Rows: 151260 Spec - Columns: 6 Properties Flow Variables						
Row ID	fecha	tipo_in...	linea	ingreso	Year	Month (number)
Row0	2012-01-01	Boletos	1	648,825	2012	1
Row1	2012-01-01	Boletos	2	657,120	2012	1
Row2	2012-01-01	Boletos	3	466,803	2012	1
Row3	2012-01-01	Boletos	4	60,120	2012	1
Row4	2012-01-01	Boletos	5	242,280	2012	1
Row5	2012-01-01	Boletos	6	138,525	2012	1
Row6	2012-01-01	Boletos	7	138,690	2012	1
Row7	2012-01-01	Boletos	8	328,383	2012	1
Row8	2012-01-01	Boletos	9	176	2012	1
Row9	2012-01-01	Boletos	A	275,370	2012	1
Row10	2012-01-01	Boletos	B	472,665	2012	1
Row11	2012-01-01	Boletos	12	NaN	2012	1
Row12	2012-01-02	Boletos	1	1,532,940	2012	1
Row13	2012-01-02	Boletos	2	1,491,780	2012	1
Row14	2012-01-02	Boletos	3	1,168,590	2012	1
Row15	2012-01-02	Boletos	4	165,690	2012	1
Row16	2012-01-02	Boletos	5	437,025	2012	1
Row17	2012-01-02	Boletos	6	286,620	2012	1
Row18	2012-01-02	Boletos	7	433,323	2012	1
Row19	2012-01-02	Boletos	8	777,951	2012	1
Row20	2012-01-02	Boletos	9	504	2012	1
Row21	2012-01-02	Boletos	A	514,455	2012	1
Row22	2012-01-02	Boletos	B	965,385	2012	1
Row23	2012-01-02	Boletos	12	NaN	2012	1
Row24	2012-01-03	Boletos	1	1,398,639	2012	1

Dialog - 5:7 - Extract Date&Time Fields (extrae año y)

File

Options Flow Variables Job Manager Selection Memory Policy

Column Selection

Date&Time column: fecha

Date Fields

☒ Year

☐ Year (week-based)

☐ Quarter

☒ Month (number)

☐ Month (name)

☐ Week

☐ Day of year

☐ Day of month

☐ Day of week (number)

☐ Day of week (name)

Time Fields

☐ Hour

☐ Minute

☐ Second

☐ Subsecond in: milliseconds

Time Zone Fields

☐ Time zone name

☐ Time zone offset

Localization (month and day names, etc.)

Locale: en_US

OK Apply Cancel

4) Nodo de python:

Este código se encarga de darle un tratamiento mas profundo a los datos contenidos en algunos de las columnas del conjunto de datos, con el objetivo de hacer su significado más claro.

```
import knime.scripting.io as knio

# Lista Auxiliar de Meses
meses = {"Enero", "Febrero", "Marzo", "Abril", "Mayo", "Junio", "Julio",
"Agosto",
        "Septiembre", "Octubre", "Noviembre", "Diciembre"}

# Conversion de KNIME a DataFrame
ingreso = knio.input_tables[0].to_pandas()

# ----- TRATAMIENTO DE DATOS -----

# Rellena todos los valores NaN de INGRESO con 0s
ingreso['ingreso'] = ingreso['ingreso'].fillna(0)

# Reemplaza el NUMERO DE MES por el NOMBRE DEL MES
ingreso['Month (number)'] = ingreso['Month
(number)'].replace(list(range(1, 13)), meses)

# Agrega la Leyenda "Linea " a los numeros de linea
lineas_toReplace = ingreso['linea'].unique() # obtener el dominio de
LINEA
new_lineas = [] # lista auxiliar

# guardar "Linea #" en 'new_lineas', usando los valores contenidos en
'lineas_toReplace'
for l in lineas_toReplace:
    new_lineas.append('Linea ' + l)

# Reemplazar los valores de 'lineas_toReplace', por los de 'new_linea'
ingreso['linea'] = ingreso['linea'].replace(lineas_toReplace, new_lineas)

# Reemplazamiento de un solo valor de RIPO_INGRESO, para mejorar su
claridad
ingreso['tipo_ingreso'] = ingreso['tipo_ingreso'].replace('Tarjetas',
'Compra de Tarjetas')

# ----- TRATAMIENTO DE DATOS -----

# Tabla de Salida
knio.output_tables[0] = knio.Table.from_pandas(ingreso)
```

Este nodo nos permitira saber la suma los ingresos totales registrados por cada mes del año.

File Edit Hilite Navigation View

Dialog - 3B - GroupBy (obtiene los)

OK Apply Cancel

Tratamiento de datos Afluencia de datos

1) Carga del conjunto de datos *afluenciastc_simple_05_2023*

Settings Transformation Advanced Settings Encryption Flow Variables Job Manager Selection Memory Policy

Read from: Local File System

Mode: ☒ File ☐ Files in folder

File: C:\Users\Arturo\Downloads\afuenciastc_simple_05_2023.xlsx Browse...

Sheet selection

☒ Select first sheet with data (afuenciastc_simple_05_2023)

☐ Select sheet with name: afuenciastc_simple_05_2023

☐ Select sheet at index: 0 (Sheet indexes start with 0.)

Column header

☒ Use Excel column name e.g. A, B, C ☐ Use column index e.g. Col0, Col1, Col2

☒ Table contains column names in row number 1 (Row numbers start with 1. See 'File Content' tab to identify row numbers.)

Empty column name prefix: empty_

Row ID

☒ Generate row IDs ☐ Table contains row IDs in column: A

Sheet area

☒ Read entire data of the sheet ☐ Read only data in columns from A to and rows from 1 to (See 'File Content' tab to identify columns and rows.)

Preview File Content

Preview with current settings

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	fecha	anio	mes	linea	estacion	afluencia
Row0	2020-01-01	2020	Enero	Linea 1	Zaragoza	20227

2) Group by

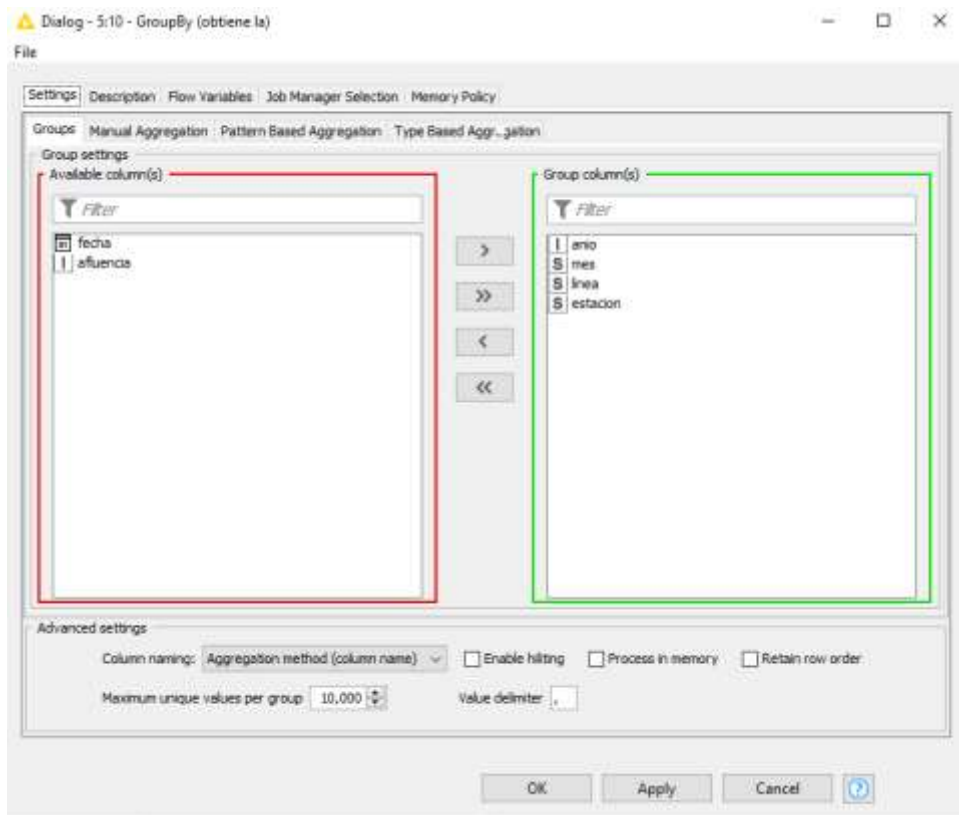
Obtiene la Afluencia total que se registra por cada mes.

Group table - 5:10 - GroupBy (obtiene la)

File Edit Help Navigation View

Table 'default' - Rows: 31394 Spec: Columns: 5 - Properties Flow Variables

Row ID	anio	mes	linea	estacion	Sum(afluencia)
Row0	2020	Abril	Linea 1	Bobuena	428911
Row1	2020	Abril	Linea 1	Itabasca	593076
Row2	2020	Abril	Linea 1	Boulevard Puerto Aéreo	930660
Row3	2020	Abril	Linea 1	Candelaria	659634
Row4	2020	Abril	Linea 1	Chapultepec	1911301
Row5	2020	Abril	Linea 1	Cuauhtémoc	620730
Row6	2020	Abril	Linea 1	Gómez Farias	1291789
Row7	2020	Abril	Linea 1	Insurgentes	1683470
Row8	2020	Abril	Linea 1	Isabel la Católica	627375
Row9	2020	Abril	Linea 1	Juanacatlán	204215
Row10	2020	Abril	Linea 1	Merced	1574063
Row11	2020	Abril	Linea 1	Moctezuma	808383
Row12	2020	Abril	Linea 1	Observatorio	1789760
Row13	2020	Abril	Linea 1	Pantitlán	1531208
Row14	2020	Abril	Linea 1	Pino Suárez	828230
Row15	2020	Abril	Linea 1	Salto del Agua	742609
Row16	2020	Abril	Linea 1	San Lázaro	901348
Row17	2020	Abril	Linea 1	Sevilla	787892
Row18	2020	Abril	Linea 1	Tacubaya	1040385
Row19	2020	Abril	Linea 1	Zaragoza	1580493
Row20	2020	Abril	Linea 12	Atila	0
Row21	2020	Abril	Linea 12	Calle 11	0
Row22	2020	Abril	Linea 12	Culhuacán	0
Row23	2020	Abril	Linea 12	Eje Central	0
Row24	2020	Abril	Linea 12	Ermita	0



Tratamiento de datos Ingresos y Afluencia de datos

1) Inner join

Se usa para juntar ambos conjuntos de datos con sus respectivos tratamientos hechos.

Join result - 5:11 - Joiner (inner join)

File Edit Hints Navigation View

Table "default" - Rows: 80727 - Spec - Columns: 10 - Properties - Flow Variables

Row ID	I año	S mes	S línea	S estación	I Sum(af...	S tpo_in...	S línea (i...	I Year	S Month (...)	D Sum(jn...
Row4680_Row0	2012	Abril	Linea 1	Balbuena	761584	Boletos	Linea 1	2012	Abril	38,551,920
Row4680_Row...	2012	Abril	Linea 1	Balbuena	761584	Compra de ...	Linea 1	2012	Abril	10
Row4680_Row...	2012	Abril	Linea 1	Balbuena	761584	Recargas	Linea 1	2012	Abril	17,237,236.6
Row4681_Row0	2012	Abril	Linea 1	Balderas	509031	Boletos	Linea 1	2012	Abril	38,551,920
Row4681_Row...	2012	Abril	Linea 1	Balderas	509031	Compra de ...	Linea 1	2012	Abril	10
Row4681_Row...	2012	Abril	Linea 1	Balderas	509031	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4682_Row0	2012	Abril	Linea 1	Boulevard P...	1447999	Boletos	Linea 1	2012	Abril	38,551,920
Row4682_Row...	2012	Abril	Linea 1	Boulevard P...	1447999	Compra de ...	Linea 1	2012	Abril	10
Row4682_Row...	2012	Abril	Linea 1	Boulevard P...	1447999	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4683_Row0	2012	Abril	Linea 1	Candelaria	704132	Boletos	Linea 1	2012	Abril	38,551,920
Row4683_Row...	2012	Abril	Linea 1	Candelaria	704132	Compra de ...	Linea 1	2012	Abril	10
Row4683_Row...	2012	Abril	Linea 1	Candelaria	704132	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4684_Row0	2012	Abril	Linea 1	Chapultepec	1323732	Boletos	Linea 1	2012	Abril	38,551,920
Row4684_Row...	2012	Abril	Linea 1	Chapultepec	1323732	Compra de ...	Linea 1	2012	Abril	10
Row4684_Row...	2012	Abril	Linea 1	Chapultepec	1323732	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4685_Row0	2012	Abril	Linea 1	Cuauhtémoc	591260	Boletos	Linea 1	2012	Abril	38,551,920
Row4685_Row...	2012	Abril	Linea 1	Cuauhtémoc	591260	Compra de ...	Linea 1	2012	Abril	10
Row4685_Row...	2012	Abril	Linea 1	Cuauhtémoc	591260	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4686_Row0	2012	Abril	Linea 1	Gómez Farias	1009699	Boletos	Linea 1	2012	Abril	38,551,920
Row4686_Row...	2012	Abril	Linea 1	Gómez Farias	1009699	Compra de ...	Linea 1	2012	Abril	10
Row4686_Row...	2012	Abril	Linea 1	Gómez Farias	1009699	Recargas	Linea 1	2012	Abril	17,257,236.6
Row4687_Row0	2012	Abril	Linea 1	Insurgentes	1120377	Boletos	Linea 1	2012	Abril	38,551,920
Row4687_Row...	2012	Abril	Linea 1	Insurgentes	1120377	Compra de ...	Linea 1	2012	Abril	10
Row4687_Row...	2012	Abril	Linea 1	Insurgentes	1120377	Recargas	Linea 1	2012	Abril	17,237,236.6
Row4688_Row0	2012	Abril	Linea 1	Isabel la Cat...	693093	Boletos	Linea 1	2012	Abril	38,551,920

Dialog - 5:11 - Joiner (inner join)

File

Joiner Settings Column Selection Performance Flow Variables Job Manager Selection Memory Policy

Top Input (left table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude No columns in this list

☒ Enforce exclusion

Include

 ☐ Enforce inclusion

Bottom Input (right table)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude No columns in this list

☒ Enforce exclusion

Include

 ☐ Enforce inclusion

Duplicate column names:
☐ Do not execute
☒ Append custom suffix (right)

OK Apply Cancel

2) Column filter

Se usa para eliminar datos que se encuentren repetidos.

Filtered table - 5:12 - Column Filter (filtra)

File Edit Filter Navigation View

Table "default" - Rows: 80727 Spec - Columns: 7 Properties Flow Variables

Row ID	S linea	S estacion	I Sum(afl...	S tipo_in...	I Year	S Month (...)	D Sum(In...
Row4680_Row0	Linea 1	Babuada	761584	Boletos	2012	Abril	38,551,920
Row4680_Row...	Linea 1	Babuada	761584	Compra de ...	2012	Abril	10
Row4680_Row...	Linea 1	Babuada	761584	Recargas	2012	Abril	17,257,236.6
Row4681_Row0	Linea 1	Balderas	509031	Boletos	2012	Abril	38,551,920
Row4681_Row...	Linea 1	Balderas	509031	Compra de ...	2012	Abril	10
Row4681_Row...	Linea 1	Balderas	509031	Recargas	2012	Abril	17,257,236.6
Row4682_Row0	Linea 1	Boulevard P...	1447599	Boletos	2012	Abril	38,551,920
Row4682_Row...	Linea 1	Boulevard P...	1447599	Compra de ...	2012	Abril	10
Row4682_Row...	Linea 1	Boulevard P...	1447599	Recargas	2012	Abril	17,257,236.6
Row4683_Row0	Linea 1	Candelaria	704132	Boletos	2012	Abril	38,551,920
Row4683_Row...	Linea 1	Candelaria	704132	Compra de ...	2012	Abril	10
Row4683_Row...	Linea 1	Candelaria	704132	Recargas	2012	Abril	17,257,236.6
Row4684_Row0	Linea 1	Chapultepec	1323732	Boletos	2012	Abril	38,551,920
Row4684_Row...	Linea 1	Chapultepec	1323732	Compra de ...	2012	Abril	10
Row4684_Row...	Linea 1	Chapultepec	1323732	Recargas	2012	Abril	17,257,236.6
Row4685_Row0	Linea 1	Cuauhtémoc	591260	Boletos	2012	Abril	38,551,920
Row4685_Row...	Linea 1	Cuauhtémoc	591260	Compra de ...	2012	Abril	10
Row4685_Row...	Linea 1	Cuauhtémoc	591260	Recargas	2012	Abril	17,257,236.6
Row4686_Row0	Linea 1	Gómez Farías	1009699	Boletos	2012	Abril	38,551,920
Row4686_Row...	Linea 1	Gómez Farías	1009699	Compra de ...	2012	Abril	10
Row4686_Row...	Linea 1	Gómez Farías	1009699	Recargas	2012	Abril	17,257,236.6
Row4687_Row0	Linea 1	Insurgentes	1120377	Boletos	2012	Abril	38,551,920
Row4687_Row...	Linea 1	Insurgentes	1120377	Compra de ...	2012	Abril	10
Row4687_Row...	Linea 1	Insurgentes	1120377	Recargas	2012	Abril	17,257,236.6
Row4688_Row0	Linea 1	Isabel la Cat...	693053	Boletos	2012	Abril	38,551,920

Dialog - 5:12 - Column Filter (filtra)

File

Column Filter Flow Variables Job Manager Selection Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

- I año
- S mes
- S linea (right)

☒ Enforce exclusion

Include

Filter

- S linea
- S estacion
- I Sum(afluencia)
- S tipo_ingreso
- I Year
- S Month (number)
- D Sum(Ingreso)

☐ Enforce inclusion

OK Apply Cancel ?

3) Column Rename

Se usa para renombrar el nombre de los atributos.

Renamed/Retyped table - 5:13 - Column Rename (CONJUNTO DE)

File Edit Hilit Navigation View

Table "default" - Rows: 80727 Spec - Columns: 7 Properties Flow Variables

Row ID	S Línea	S Estacion	I Afluencia	S TipoIng...	I Año	S Mes	D Ingreso
Row4680_Row0	Línea 1	Balbuena	761584	Boletos	2012	Abril	38,551,920
Row4680_Row...	Línea 1	Balbuena	761584	Compra de ...	2012	Abril	10
Row4680_Row...	Línea 1	Balbuena	761584	Recargas	2012	Abril	17,257,236.6
Row4681_Row0	Línea 1	Balderas	509031	Boletos	2012	Abril	38,551,920
Row4681_Row...	Línea 1	Balderas	509031	Compra de ...	2012	Abril	10
Row4681_Row...	Línea 1	Balderas	509031	Recargas	2012	Abril	17,257,236.6
Row4682_Row0	Línea 1	Boulevard P...	1447599	Boletos	2012	Abril	38,551,920
Row4682_Row...	Línea 1	Boulevard P...	1447599	Compra de ...	2012	Abril	10
Row4682_Row...	Línea 1	Boulevard P...	1447599	Recargas	2012	Abril	17,257,236.6
Row4683_Row0	Línea 1	Candelaria	704132	Boletos	2012	Abril	38,551,920
Row4683_Row...	Línea 1	Candelaria	704132	Compra de ...	2012	Abril	10
Row4683_Row...	Línea 1	Candelaria	704132	Recargas	2012	Abril	17,257,236.6
Row4684_Row0	Línea 1	Chapultepec	1323732	Boletos	2012	Abril	38,551,920
Row4684_Row...	Línea 1	Chapultepec	1323732	Compra de ...	2012	Abril	10
Row4684_Row...	Línea 1	Chapultepec	1323732	Recargas	2012	Abril	17,257,236.6
Row4685_Row0	Línea 1	Cuauhtémoc	591260	Boletos	2012	Abril	38,551,920
Row4685_Row...	Línea 1	Cuauhtémoc	591260	Compra de ...	2012	Abril	10
Row4685_Row...	Línea 1	Cuauhtémoc	591260	Recargas	2012	Abril	17,257,236.6
Row4686_Row0	Línea 1	Gómez Farías	1009699	Boletos	2012	Abril	38,551,920
Row4686_Row...	Línea 1	Gómez Farías	1009699	Compra de ...	2012	Abril	10
Row4686_Row...	Línea 1	Gómez Farías	1009699	Recargas	2012	Abril	17,257,236.6
Row4687_Row0	Línea 1	Insurgentes	1120377	Boletos	2012	Abril	38,551,920
Row4687_Row...	Línea 1	Insurgentes	1120377	Compra de ...	2012	Abril	10
Row4687_Row...	Línea 1	Insurgentes	1120377	Recargas	2012	Abril	17,257,236.6
Row4688_Row0	Línea 1	Isabel la Cat...	693093	Boletos	2012	Abril	38,551,920

Dialog - 5:13 - Column Rename (CONJUNTO DE)

File

Change columns Flow Variables Job Manager Selection Memory Policy

Column Search

Filter Options
None

S | línea
S | estacion
I | Sum(afluencia)
S | tipo_ingreso
I | Year
S | Month (number)
D | Sum(ingreso)

Sum(afluencia) Remove
[X] Change: Afluencia [I] IntValue

Year Remove
[X] Change: Año [I] IntValue

Month (number) Remove
[X] Change: Mes [S] StringValue

Sum(ingreso) Remove
[X] Change: Ingreso [D] DoubleValue

estacion Remove
[X] Change: Estacion [S] StringValue

linea Remove
[X] Change: Linea [S] StringValue

tipo_ingreso Remove
[X] Change: TipoIngreso [S] StringValue

OK Apply Cancel ?

Conjunto de Datos Final generado con el Tratamiento de Datos realizado

Nombre	Significado	Tipo	Dominio
Línea	Número de línea del sistema del transporte público.	Categórico	Línea 1 – Línea 12 Línea A Línea B
Estacion	Nombre de la estación donde se encuentra la afluencia.	Categórico	Nombre de la estación.
Afluencia	Número de personas que utilizaron esa estación en particular en el día especificado.	Numérico	Numero entero Positivo.
TipoIngreso	El valor numérico que representa la cantidad de ingresos registrados para esa línea en particular en el día especificado.	Categórico	Forma de pago.
Año	La fecha en la que se registró la afluencia.	Numérico	2010 – 2023
Mes	El mes correspondiente a la fecha.	Categórico	Enero – Diciembre
Ingreso	El valor numérico que representa la cantidad de ingresos registrados para esa línea en particular en el día especificado.	Numérico	Numero entero positivo.

Tratamiento al conjunto de FIFA World Cup Attendance

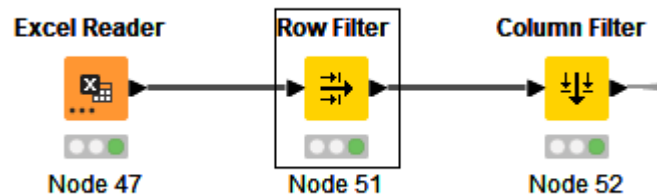


Diagrama de tratamiento

1) Aplicación de Row Filter

Eliminamos el renglón que se encargaba de sumar el número total de asistencia durante cada una de las sedes del mundial.

The screenshot shows the 'Filter Criteria' dialog box with the 'Filter Criteria' tab selected. On the left, there are radio buttons for filtering options: 'Include rows by attribute value', 'Exclude rows by attribute value' (selected), 'Include rows by number', 'Exclude rows by number', 'Include rows by row ID', and 'Exclude rows by row ID'. On the right, the 'Column value matching' section is expanded, showing 'Column to test:' set to 'Year'. Below this, there are checkboxes for 'filter based on collection elements', 'use pattern matching' (selected), 'use range checking', and 'only missing values match'. The 'use pattern matching' section includes a dropdown set to 'overall' and checkboxes for 'case sensitive match', 'contains wild cards', and 'regular expression'.

2) Aplicación de Column Filter

Apartamos las columnas de “Número”, “Venue” y “Game(s)” ya que no las necesitaremos para nuestros modelos de predicción.

The screenshot shows the 'Column Filter' dialog box with the 'Column Filter' tab selected. At the top, there are radio buttons for 'Manual Selection' (selected), 'Wildcard/Regex Selection', and 'Type Selection'. The dialog is divided into two main sections: 'Exclude' on the left and 'Include' on the right. The 'Exclude' section has a red border and contains a list of columns: 'Number', 'Venue', and 'Game(s)'. The 'Include' section has a green border and contains a list of columns: 'Year', 'Hosts', 'Total_Attendance', 'Matches', and 'Average_Attendance'. At the bottom, there are radio buttons for 'Enforce exclusion' (selected) and 'Enforce inclusion'.

Tenemos una tabla ya tratada de la siguiente manera:

Row ID	\$ Year	\$ Hosts	Total_Attendance	Matches	Average_Attendance
0	1930	Uruguay	590549	18	32808
1	1934	Italy	363000	17	21353
2	1938	France	375700	18	20872
3	1950	Brazil	1045246	22	47511
4	1954	Switzerland	768607	26	29562
5	1958	Sweden	819810	35	23423
6	1962	Chile	893172	32	27912
7	1966	England	1563135	32	48848
8	1970	Mexico	1603975	32	50124
9	1974	West Germany	1865753	38	49099
10	1978	Argentina	1545791	38	40679
11	1982	Spain	2109723	52	40572
12	1986	Mexico	2394031	52	46039
13	1990	Italy	2516215	52	48389
14	1994	United States	3587538	52	68991
15	1998	France	2785100	64	43517
16	2002	South Korea Japan	2705197	64	42269
17	2006	Germany	3359439	64	52491
18	2010	South Africa	3178856	64	49670
19	2014	Brazil	3429873	64	53592
20	2018	Russia	3031768	64	47371
21	2022	Qatar	3404252	64	53191

Tabla con el Tratamiento de Datos realizado

Nombre	Significado	Tipo	Dominio
Host	Anfitrión de la Copa del Mundo	Categorico	Sede elegida por la FIFA
Total_Attendance	Asistencia total de personas	Numérico	Numero entero Positivo
Year	El año que se disputo el mundial	Numérico	1930 – 2022
Average_Attendance	Promedio de personas con asistencia en vivo	Numérico	Numero entero Positivo
Matches	Número total de partidos Jugados	Numerico	Numero entero Positivo

Tratamiento al conjunto de Credit Cards Approval

- 1) Se inserta un nodo de Python para renombrar valores en columnas.

```
import knime.scripting.io as knio

# Lista Auxiliar
binary = [0, 1]

# Conversion de KNIME a DataFrame
credit = knio.input_tables[0].to_pandas()

# ----- TRATAMIENTO DE DATOS -----

# 0=Female, 1=Male
credit['Gender'] = credit['Gender'].replace(binary, ['Female', 'Male'])

# 0=Single/Divorced/etc, 1=Married
credit['Married'] = credit['Married'].replace(binary,
['Single/Divorced/etc', 'Married'])

# 0=does not have a bank account, 1=has a bank account
credit['BankCustomer'] = credit['BankCustomer'].replace(binary, ['does
not have a bank account', 'has a bank account'])

# 0=no prior defaults, 1=prior default
credit['PriorDefault'] = credit['PriorDefault'].replace(binary, ['no
prior defaults', 'prior default'])

# 0=not employed, 1=employed
credit['Employed'] = credit['Employed'].replace(binary, ['not
employed', 'employed'])

# 0=no license, 1=has license
credit['DriversLicense'] = credit['DriversLicense'].replace(binary, ['no
license', 'has license'])

# 0=not approved, 1=approved
credit['Approved'] = credit['Approved'].replace(binary, ['NOT
APPROVED', 'APPROVED'])

# ----- TRATAMIENTO DE DATOS -----

# Tabla de Salida
knio.output_tables[0] = knio.Table.from_pandas(credit)
```

S	Gender	D	Age	D	Debt	S	Married	S	BankCustomer	S	Industry	S	Ethnicity	D	YearAt...	S	PriorDefault	S	Employed	I	CreditScore	S	Overseas...	S	Citizen	I	ZipCode	I	Bus...	S	Approved
Male	30.83	0					Married		has a bank account		Industrials		White	1.25		prior default		employed	1			no license		ByBirth	212	0					APPROVED
Female	38.67	4.46					Married		has a bank account		Materials		Black	3.04		prior default		employed	6			no license		ByBirth	43	560					APPROVED
Female	24.5	0.5					Married		has a bank account		Materials		Black	1.5		prior default		not employed	0			no license		ByBirth	380	824					APPROVED
Male	27.83	1.54					Married		has a bank account		Industrials		White	3.75		prior default		employed	5			has license		ByBirth	300	3					APPROVED
Male	26.17	5.625					Married		has a bank account		Industrials		White	1.71		prior default		not employed	0			no license		ByOtherMeans	120	0					APPROVED
Male	22.08	4					Married		has a bank account		Communicat...		White	2.5		prior default		not employed	0			has license		ByBirth	360	0					APPROVED
Male	23.17	1.04					Married		has a bank account		Transport		Black	6.5		prior default		not employed	0			has license		ByBirth	364	31285					APPROVED
Female	22.92	11.585					Married		has a bank account		Information...		White	0.04		prior default		not employed	0			no license		ByBirth	80	1346					APPROVED
Male	34.42	0.5					Single/Divorced...		does not have a ba...		Financials		Black	3.96		prior default		not employed	0			no license		ByBirth	180	314					APPROVED
Male	42.5	4.915					Single/Divorced...		does not have a ba...		Industrials		White	3.165		prior default		not employed	0			has license		ByBirth	12	1442					APPROVED
Male	22.08	0.83					Married		has a bank account		Energy		Black	2.165		no prior defaults		not employed	0			has license		ByBirth	128	0					APPROVED
Male	25.92	1.835					Married		has a bank account		Energy		Black	4.325		prior default		not employed	0			no license		ByBirth	360	200					APPROVED
Female	26.25	6					Married		has a bank account		Financials		White	1		prior default		not employed	0			has license		ByBirth	0	0					APPROVED
Male	46.38	6.04					Married		has a bank account		Financials		White	0.04		no prior defaults		not employed	0			no license		ByBirth	0	2640					APPROVED
Female	45.83	60.5					Married		has a bank account		Materials		White	5		prior default		employed	7			has license		ByBirth	0	0					APPROVED
Male	26.67	4.415					Single/Divorced...		does not have a ba...		Financials		White	0.25		prior default		employed	30			has license		ByBirth	120	0					APPROVED
Male	28.25	0.875					Married		has a bank account		Communicat...		White	0.96		prior default		employed	3			has license		ByBirth	296	0					APPROVED
Female	25.25	5.875					Married		has a bank account		Materials		White	3.17		prior default		employed	30			no license		ByBirth	120	245					APPROVED
Male	21.83	0.25					Married		has a bank account		Real Estate		Black	0.665		prior default		not employed	0			has license		ByBirth	0	0					APPROVED
Female	18.17	8.585					Married		has a bank account		Information...		Black	0.75		prior default		employed	7			no license		ByBirth	96	0					APPROVED
Male	28	11.25					Married		has a bank account		Energy		White	2.5		prior default		employed	17			no license		ByBirth	300	1208					APPROVED
Male	23.25	1					Married		has a bank account		Energy		White	0.635		prior default		not employed	0			no license		ByOtherMeans	300	0					APPROVED
Female	47.75	8					Married		has a bank account		Energy		White	7.675		prior default		employed	6			has license		ByBirth	0	1246					APPROVED
Female	27.42	14.5					Married		has a bank account		Utilities		Black	3.085		prior default		employed	1			no license		ByBirth	120	11					APPROVED
Female	41.17	4.1					Married		has a bank account		Materials		White	0.1		prior default		employed	3			has license		ByBirth	145	0					APPROVED
Female	15.83	0.385					Married		has a bank account		Energy		Black	1.5		prior default		employed	2			no license		ByBirth	300	0					APPROVED
Female	47	12					Married		has a bank account		ConsumerD...		Asian	3.165		prior default		employed	9			has license		ByBirth	0	0					APPROVED
Male	36.58	38.5					Married		has a bank account		Real Estate		Asian	15		prior default		employed	17			has license		ByBirth	0	0					APPROVED
Male	37.42	8.5					Married		has a bank account		Education		Black	7		prior default		employed	3			no license		ByBirth	0	0					APPROVED
Male	42.08	1.04					Married		has a bank account		Industrials		White	5		prior default		employed	6			has license		ByBirth	300	16000					APPROVED
Male	29.25	34.79					Married		has a bank account		ConsumerD...		White	5.04		prior default		employed	5			has license		ByBirth	268	0					APPROVED
Male	42	9.79					Married		has a bank account		Utilities		Black	7.96		prior default		employed	8			no license		ByBirth	0	0					APPROVED
Male	46.5	7.585					Married		has a bank account		ConsumerD...		Asian	7.985		prior default		employed	15			has license		ByBirth	0	5060					APPROVED
Female	36.75	5.125					Married		has a bank account		Education		White	5		prior default		not employed	0			has license		ByBirth	0	4060					APPROVED
Female	22.58	30.75					Married		has a bank account		Materials		White	0.415		prior default		employed	5			has license		ByBirth	0	560					APPROVED
Male	27.83	1.5					Married		has a bank account		Industrials		White	2		prior default		employed	11			has license		ByBirth	434	35					APPROVED
Male	29.25	11.585					Married		has a bank account		Information...		Black	1.835		prior default		employed	12			has license		ByBirth	583	713					APPROVED
Female	23	11.75					Married		has a bank account		Utilities		Black	0.5		prior default		employed	2			has license		ByBirth	300	551					APPROVED
Male	27.75	0.585					Single/Divorced...		does not have a ba...		Information...		White	0.25		prior default		employed	2			no license		ByBirth	260	560					APPROVED
Male	34.58	9.415					Married		has a bank account		Healthcare		Latino	14.415		prior default		employed	11			has license		ByBirth	30	360					APPROVED
Male	34.17	9.17					Married		has a bank account		Energy		White	4.5		prior default		employed	12			has license		ByBirth	0	221					APPROVED
Male	28.92	15					Married		has a bank account		Energy		Black	5.358		prior default		employed	11			no license		ByBirth	0	1283					APPROVED

Tabla con el Tratamiento de Datos Realizado

Nombre	Significado	Tipo	Dominio
Genders	Genero de la persona	Categórico	Masculino Femenino
Age	Edad del solicitante	Numérico	0-99
Debt	Deuda pendiente (la característica se ha escalado)	Numérico	0-10
Married	Estado civil	Categórico	Soltero/ divorciado Casado
BankCustomer	Cliente del banco	Categórico	No tiene cuenta Tiene cuenta
Industry	Sector de la industria en el que trabaja	Categórico	Nombre del sector
Ethnicity	Etnia del solicitante	Categórico	Asiático, Latino, Blanco, Negro, etc.
YearsEmployed	Años trabajando	Numérico	0-30
PriorDefault	Valor predeterminado con anterioridad.	Categórico	Sin valor Con valor
Employed	Estado laboral	Categórico	Con empleo Desempleado
CreditScore	Puntaje de crédito (esta función se ha escalado)	Numérico	0 – 99
DriversLicense	Licencia de conducir	Categórico	No tiene licencia Tiene licencia
Citizen	Ciudadanía del solicitante	Categórico	Por nacimiento Por otras vías
ZipCode	Código postal	Numérico	0 – 999
Income	Ingresos del solicitante (previamente escalados)	Numérico	0 – 999,999
Approved	Aprobación de expedición de tarjeta	Categórico	No aprobado Aprobado

I. Clasificación. *Árboles*

Equipo 4

Data Mining

3CV15

I. CLASIFICACIÓN DE ARBOLES:

I.1 Descripción del ejercicio

CART, es un algoritmo creado por Breiman en 1984. El algoritmo CART construye árboles de clasificación y regresión. El árbol de clasificación es construido por CART mediante la división binaria del atributo. El índice de Gini se utiliza para seleccionar el atributo de división. CART también se utiliza para análisis de regresión con la ayuda de un árbol de regresión. La función de regresión de CART se puede utilizar al pronosticar una variable dependiente dado un conjunto de predictores variable durante un período de tiempo determinado. CART tiene una velocidad de procesamiento promedio y admite tanto continuo como datos de atributos nominales, aunque el atributo objetivo tiene que ser nominal.

Para esta sección usamos un Árbol de Clasificación tipo CART para poder predecir la Afluencia Registrada en cada línea del STCM de la CDMX por cada mes entre los años 2012-2023.

C4.5, es un algoritmo desarrollado por JR Quinlan en 1993, como una extensión (mejora) del algoritmo ID3 que desarrolló en 1986. El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente.

El objetivo para nuestro árbol C4.5 es predecir a que línea de Metro pertenece una estación (principalmente aquellas que son correspondencia), de acuerdo con su afluencia.

I.2 Diccionarios de Datos.

I.2.1 Diccionario de Datos CART

Nombre	Significado	Tipo	Dominio
Línea	Número de línea del sistema del transporte público.	Categórico	Línea 1 – Línea 12 Línea A Línea B
Afluencia	Número de personas que utilizaron esa estación en particular en el día especificado.	Numérico	Numero entero Positivo.
Año	La fecha en la que se registró la afluencia.	Numérico	2010 – 2023
Mes	El mes correspondiente a la fecha.	Categórico	Enero – Diciembre

I.2.2 Diccionario de datos C4.5

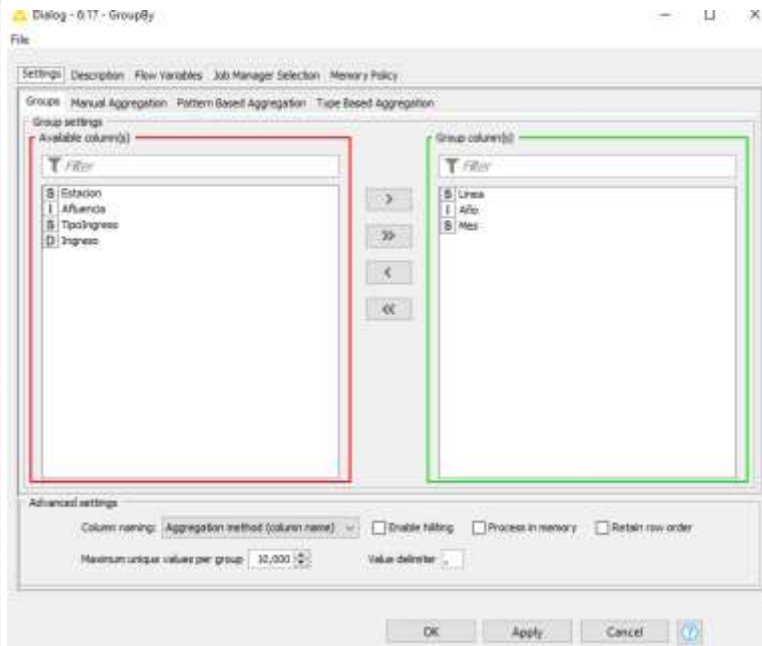
Nombre	Significado	Tipo	Dominio
Línea	Número de línea del sistema del transporte público.	Categórico	Línea 1 – Línea 12 Línea A Línea B
Afluencia	Número de personas que utilizaron esa estación en particular en el día especificado.	Numérico	Numero entero Positivo.
Estacion	Estación de la línea del metro	Categórico	Nombre de estación

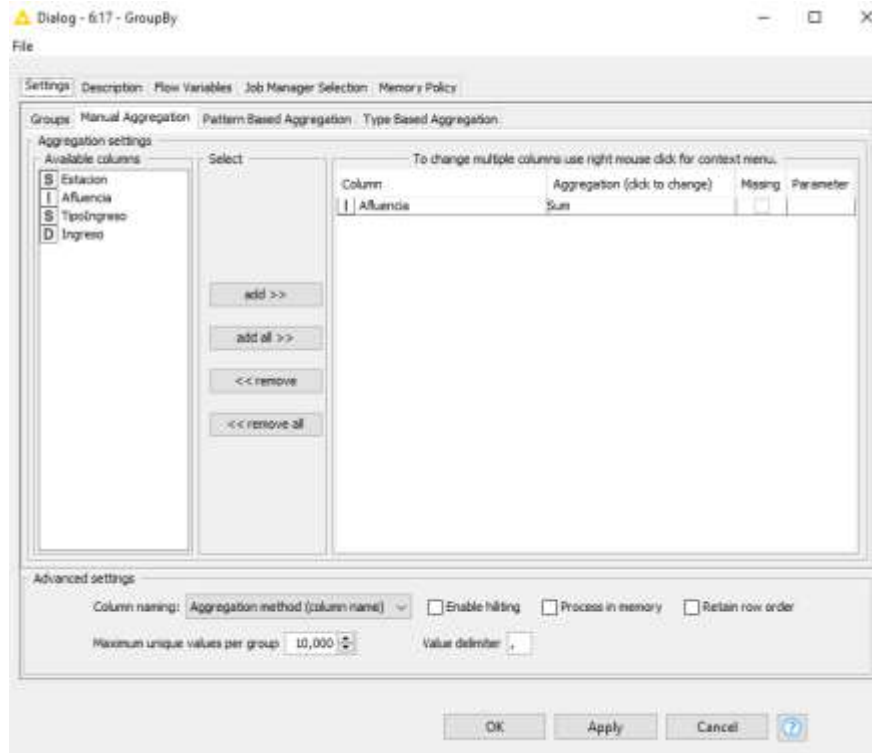
CART

1) GroupBy

Para poder empezar con nuestro árbol de decisión juntamos lo que son las líneas (Metro), Año y Mes; En función de la suma de la afluencia de ese mismo Mes y Año.

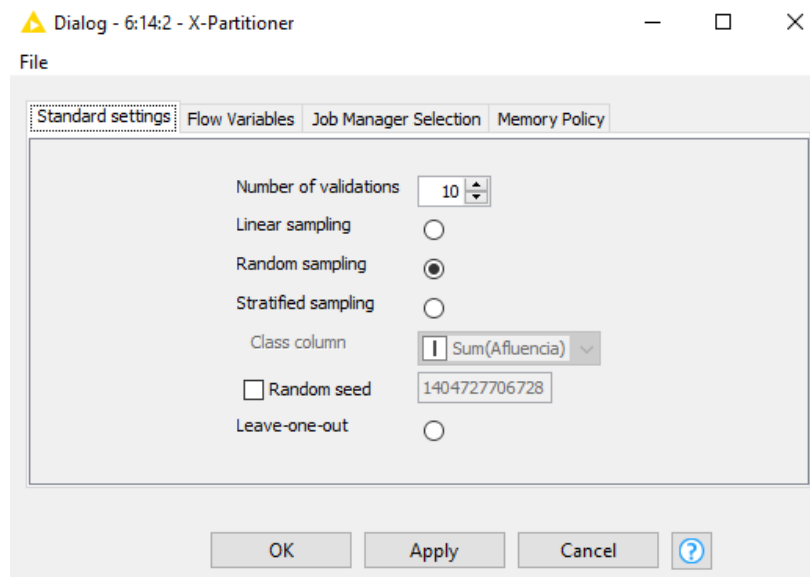
Row ID	S Línea	T Año	S Mes	T Sum(AFLUENCIA)
Row0	Línea 1	2012	Abril	63059673
Row1	Línea 1	2012	Agosto	72022381
Row2	Línea 1	2012	Diciembre	65747124
Row3	Línea 1	2012	Enero	65263842
Row4	Línea 1	2012	Febrero	62623209
Row5	Línea 1	2012	Julio	68325876
Row6	Línea 1	2012	Junio	62878215
Row7	Línea 1	2012	Marzo	67325105
Row8	Línea 1	2012	Mayo	68843796
Row9	Línea 1	2012	Noviembre	65539662
Row10	Línea 1	2012	Octubre	70236618
Row11	Línea 1	2012	Septiembre	66607665
Row12	Línea 1	2013	Abril	85477211
Row13	Línea 1	2013	Agosto	72476754
Row14	Línea 1	2013	Diciembre	67542318
Row15	Línea 1	2013	Enero	64688646
Row16	Línea 1	2013	Febrero	68091677
Row17	Línea 1	2013	Julio	68333136
Row18	Línea 1	2013	Junio	68717651
Row19	Línea 1	2013	Marzo	62105526
Row20	Línea 1	2013	Mayo	68842122
Row21	Línea 1	2013	Noviembre	69673047
Row22	Línea 1	2013	Octubre	71683006
Row23	Línea 1	2013	Septiembre	69079188
Row24	Línea 1	2014	Abril	82072361
Row25	Línea 1	2014	Agosto	70186224
Row26	Línea 1	2014	Diciembre	68321896
Row27	Línea 1	2014	Enero	64588077
Row28	Línea 1	2014	Febrero	58518372
Row29	Línea 1	2014	Julio	68300706
Row30	Línea 1	2014	Junio	63767220
Row31	Línea 1	2014	Marzo	65230468
Row32	Línea 1	2014	Mayo	65241883
Row33	Línea 1	2014	Noviembre	65085453
Row34	Línea 1	2014	Octubre	70541545
Row35	Línea 1	2014	Septiembre	68678973
Row36	Línea 1	2015	Abril	63788700
Row37	Línea 1	2015	Agosto	69988281
Row38	Línea 1	2015	Diciembre	68820876
Row39	Línea 1	2015	Enero	67202838
Row40	Línea 1	2015	Febrero	63481480
Row41	Línea 1	2015	Julio	68672799





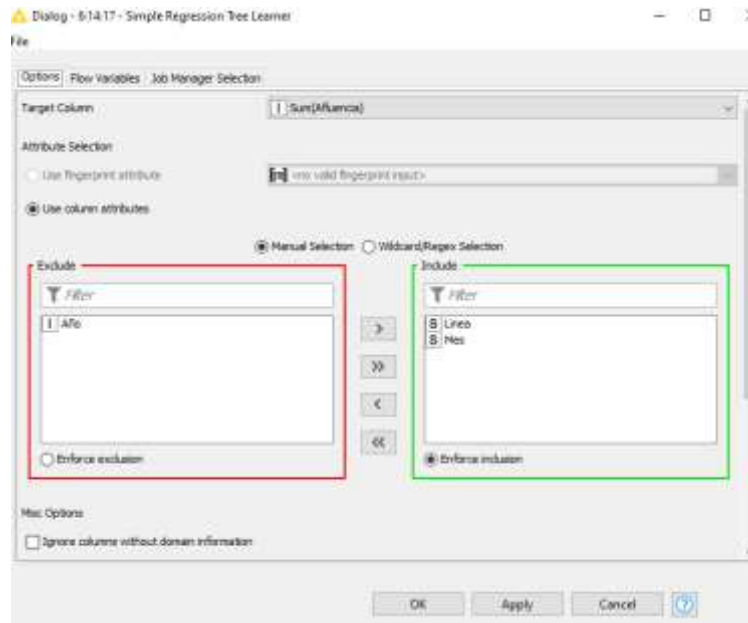
2) X-Partitioner

Este nodo se usará para dividir un conjunto de datos en subconjuntos por medio de un muestreo aleatorio. El número de validaciones se establecerá en 10.



3) Simple Regression Tree Learner

Para este nodo se especificara nuestra Columna Destino la cual será sum(Afluencia) y se incluirán los parámetros de línea y mes.



4) Simple Regression Tree Predictor

El nodo "Simple Regression Tree Predictor" toma el modelo de árbol de decisión entrenado y aplica las reglas de división para asignar una predicción numérica a cada instancia de datos de entrada.

Predicted Output - 6:14:19 - Simple Regression Tree Predictor

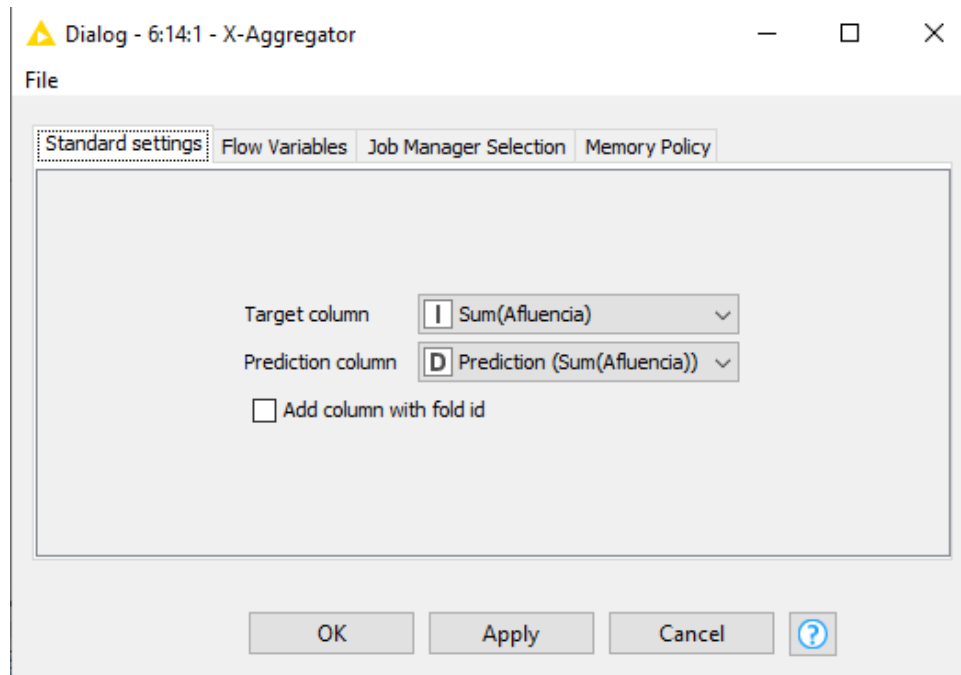
File Edit Hints Navigation View

Table 'default' - Rows: 159 - Spec - Columns: 6 - Properties - Flow Variables

Row ID	S Línea	I Año	S Mes	I Sum(Afluencia)	D Predicción Sum(Afluencia)
Row5	Línea 1	2012	Julio	88325876	55,907,284.4
Row28	Línea 1	2014	Febrero	58518372	53,924,590.9
Row32	Línea 1	2014	Mayo	65741883	55,517,802
Row34	Línea 1	2014	Octubre	70644345	56,710,105.5
Row61	Línea 1	2017	Agosto	67885503	56,631,333.3
Row71	Línea 1	2017	Septiembre	43291527	53,877,636.3
Row80	Línea 1	2018	Mayo	61090797	55,517,802
Row105	Línea 1	2020	Noviembre	34919487	59,130,047.444
Row115	Línea 1	2021	Marzo	30295613	60,282,049.7
Row117	Línea 1	2021	Noviembre	37638774	59,130,047.444
Row135	Línea 12	2012	Diciembre	17415828	19,088,952
Row147	Línea 12	2013	Diciembre	23042718	19,088,952
Row167	Línea 12	2014	Octubre	12967263	18,042,817.333
Row203	Línea 12	2017	Octubre	27335679	18,042,817.333
Row204	Línea 12	2017	Septiembre	20613813	17,818,695
Row229	Línea 12	2020	Abril	11338872	17,655,524.4
Row248	Línea 12	2021	Febrero	13854045	19,851,599.1
Row252	Línea 12	2021	Septiembre	0	17,818,695
Row259	Línea 12	2022	Junio	0	17,874,401.7
Row280	Línea 2	2013	Diciembre	70965564	63,446,640.7
Row287	Línea 2	2013	Noviembre	78170994	63,719,070
Row300	Línea 2	2014	Octubre	76417123	66,933,765.2
Row322	Línea 2	2016	Mayo	70489611	61,063,934.111
Row327	Línea 2	2017	Agosto	73020789	64,668,837
Row334	Línea 2	2017	Mayo	70576842	61,063,934.111

5) X-Aggregator

Recopila el resultado de un nodo predictor, compara la clase predicha y la clase real y genera las predicciones para todas las filas y las estadísticas de iteración.



6) Numeric score

Este análisis nos proporcionará las estadísticas generadas por nuestro árbol de decisión. El coeficiente de determinación (r^2) se utiliza como medida de precisión. En nuestro caso de estudio, se observan algunas inconsistencias en el árbol, sin embargo, al examinar las estadísticas, se puede apreciar un nivel de predicción satisfactorio, ya que el valor de r^2 se acerca a uno.

Statistics - 6:16 - Numeric Scorer

File Edit Hilite Navigation View

Table "Scores" - Rows: 7 Spec - Column: 1 Properties Flow Variables

Row ID	D Predicti...
R^2	0.731
mean absolut...	6,612,701.47
mean square...	104,015,74...
root mean sq...	10,198,810....
mean signed ...	-28,396.166
mean absolut...	NaN
adjusted R^2	0.731

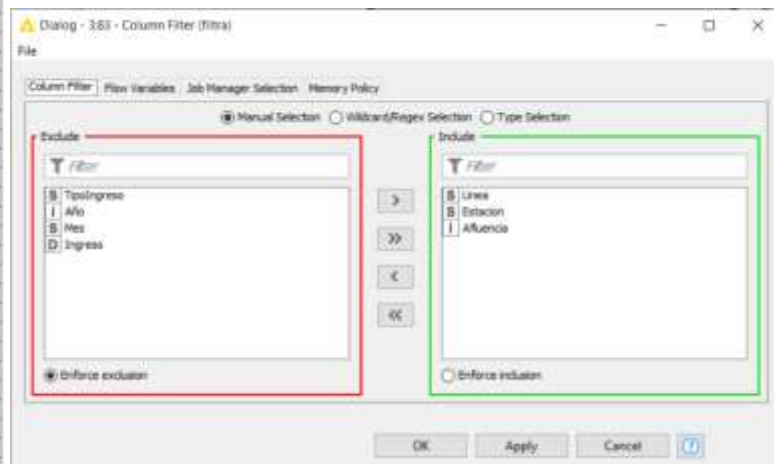
Conforme se señaló previamente, este conjunto de datos fue sometido a una operación de agrupamiento (group by) con el propósito de consolidar las líneas de datos por año y mes, y se realizó una suma de la variable "afluencia". Esto permitió la preparación del conjunto de datos que se empleó posteriormente para entrenar nuestro modelo de árbol de decisiones, utilizando los datos pertenecientes a los conjuntos de ingresos y afluencia.

C4.5

1) ColumnFilter

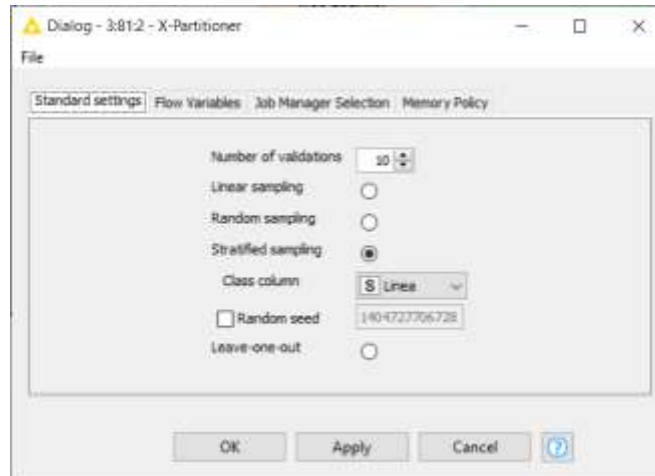
Para poder empezar con nuestro árbol de decisión, eliminamos las columnas de datos que resulten innecesarios.

Row ID	S Línea	S Estación	I Afluencia
Row4580_Row0	Línea 1	Balbuena	761584
Row4580_Row1	Línea 1	Balbuena	761584
Row4580_Row2	Línea 1	Balbuena	761584
Row4581_Row0	Línea 1	Balderas	509031
Row4581_Row1	Línea 1	Balderas	509031
Row4581_Row2	Línea 1	Balderas	509031
Row4582_Row0	Línea 1	Boulevard P...	1447599
Row4582_Row1	Línea 1	Boulevard F...	1447599
Row4582_Row2	Línea 1	Boulevard P...	1447599
Row4583_Row0	Línea 1	Candelaria	704132
Row4583_Row1	Línea 1	Candelaria	704132
Row4583_Row2	Línea 1	Candelaria	704132
Row4584_Row0	Línea 1	Chapultepec	1323732
Row4584_Row1	Línea 1	Chapultepec	1323732
Row4584_Row2	Línea 1	Chapultepec	1323732
Row4585_Row0	Línea 1	Cuauhtémoc	591260
Row4585_Row1	Línea 1	Cuauhtémoc	591260
Row4585_Row2	Línea 1	Cuauhtémoc	591260
Row4586_Row0	Línea 1	Gómez Farías	1009699
Row4586_Row1	Línea 1	Gómez Farías	1009699
Row4586_Row2	Línea 1	Gómez Farías	1009699
Row4587_Row0	Línea 1	Insurgentes	1120377
Row4587_Row1	Línea 1	Insurgentes	1120377
Row4587_Row2	Línea 1	Insurgentes	1120377
Row4588_Row0	Línea 1	Isabel la Cat...	693093
Row4588_Row1	Línea 1	Isabel la Cat...	693093
Row4588_Row2	Línea 1	Isabel la Cat...	693093
Row4589_Row0	Línea 1	Juanacatlán	646632
Row4589_Row1	Línea 1	Juanacatlán	646632
Row4589_Row2	Línea 1	Juanacatlán	646632
Row4590_Row0	Línea 1	Merced	1115579
Row4590_Row1	Línea 1	Merced	1115579
Row4590_Row2	Línea 1	Merced	1115579
Row4591_Row0	Línea 1	Moctezuma	1369178
Row4591_Row1	Línea 1	Moctezuma	1369178
Row4591_Row2	Línea 1	Moctezuma	1369178
Row4592_Row0	Línea 1	Observatorio	1980815
Row4592_Row1	Línea 1	Observatorio	1980815



2) X-Partitioner

Este nodo se usara para dividir un conjunto de datos en subconjuntos basados en una variable objetivo la cual será *Línea*. El número de validaciones se establecerá en 10.



3) Decision Tree Learner

Para este nodo se especificara como variable objetivo a la columna *Línea*.



4) Simple Regression Tree Predictor

El nodo "Decision Tree Predictor" toma el modelo de árbol de decisión entrenado para asignar una predicción categórica a cada instancia de datos de entrada.

[illegible]

5) X-Aggregator

Recopila el resultado de un nodo predictor, compara la clase predicha y la clase real y genera las predicciones para todas las filas y las estadísticas de iteración.

The screenshot shows the 'Standard settings' tab of the 'X-AGGREGATOR' dialog box. The 'Target column' is set to 'Linea' and the 'Prediction column' is set to 'Prediction (Linea)'. The checkbox 'Add column with fold id' is unchecked. The dialog has 'OK', 'Apply', and 'Cancel' buttons at the bottom, along with a help icon.

Target column	Prediction column	Add column with fold id
Linea	Prediction (Linea)	<input type="checkbox"/>

6) Score

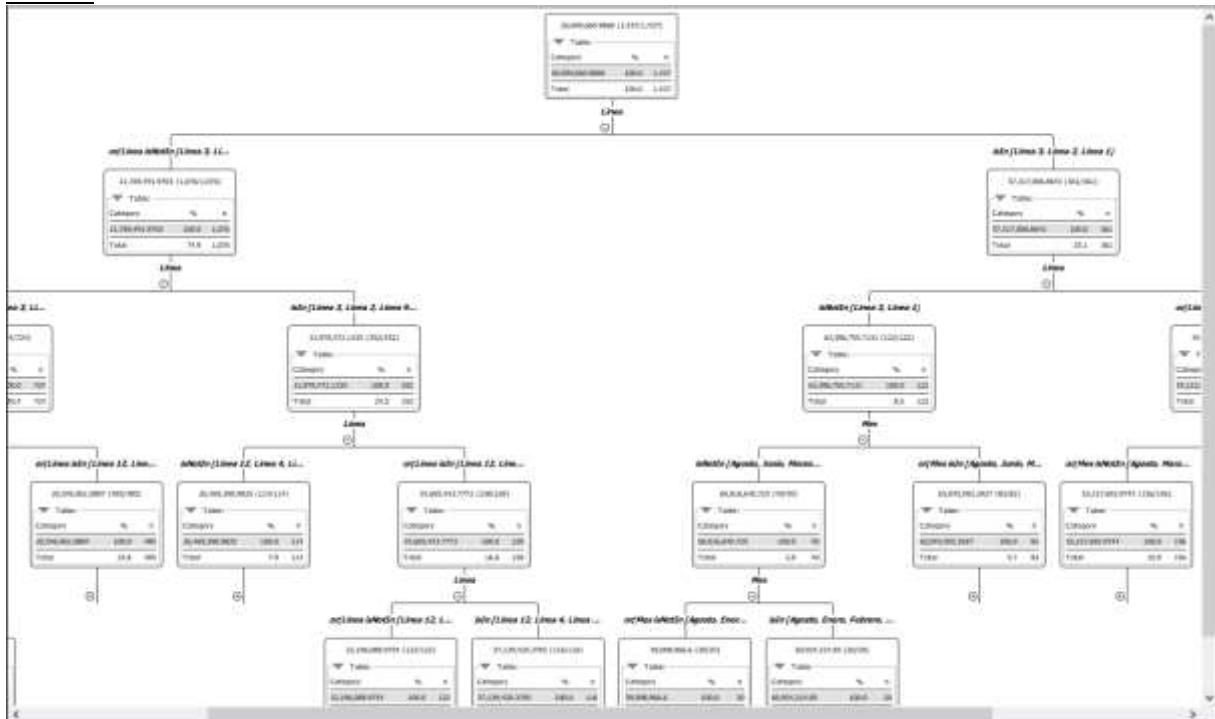
Este análisis nos proporcionará la matriz de confusión generada por nuestro árbol de decisión. En nuestro caso de estudio, se observan algunas inconsistencias en el árbol potencialmente debido a las estaciones que pertenecen a más de una línea de metro, sin embargo, al examinar las estadísticas, se puede apreciar un nivel de predicción satisfactorio, ya que el valor de precisión es del 96%.

Row ID	Línea 1	Línea 6	Línea 9	Línea 8	Línea 5	Línea 7	Línea 3	Línea 4	Línea 2	Línea B	Línea 12	Línea A
Línea 1	7910	0	66	123	10	20	29	20	52	67	0	23
Línea 6	0	4523	0	0	29	42	0	116	0	0	0	0
Línea 9	98	0	4487	20	197	0	0	84	7	0	0	27
Línea 8	40	0	12	7589	0	0	0	35	25	37	52	0
Línea 5	68	79	23	0	5061	0	0	63	0	22	0	14
Línea 7	0	82	0	0	0	5612	0	0	0	0	46	0
Línea 3	52	4	156	0	16	0	8335	0	7	40	40	0
Línea 4	3	43	1	12	90	0	0	3858	0	93	0	0
Línea 2	106	0	12	143	0	4	51	0	9457	0	67	0
Línea B	11	0	0	13	121	0	0	0	0	8462	0	0
Línea 12	0	0	0	30	0	12	0	0	0	0	8158	0
Línea A	33	0	11	0	103	0	0	0	0	0	0	3953

I.3 Resultados

a) Diagrama generado

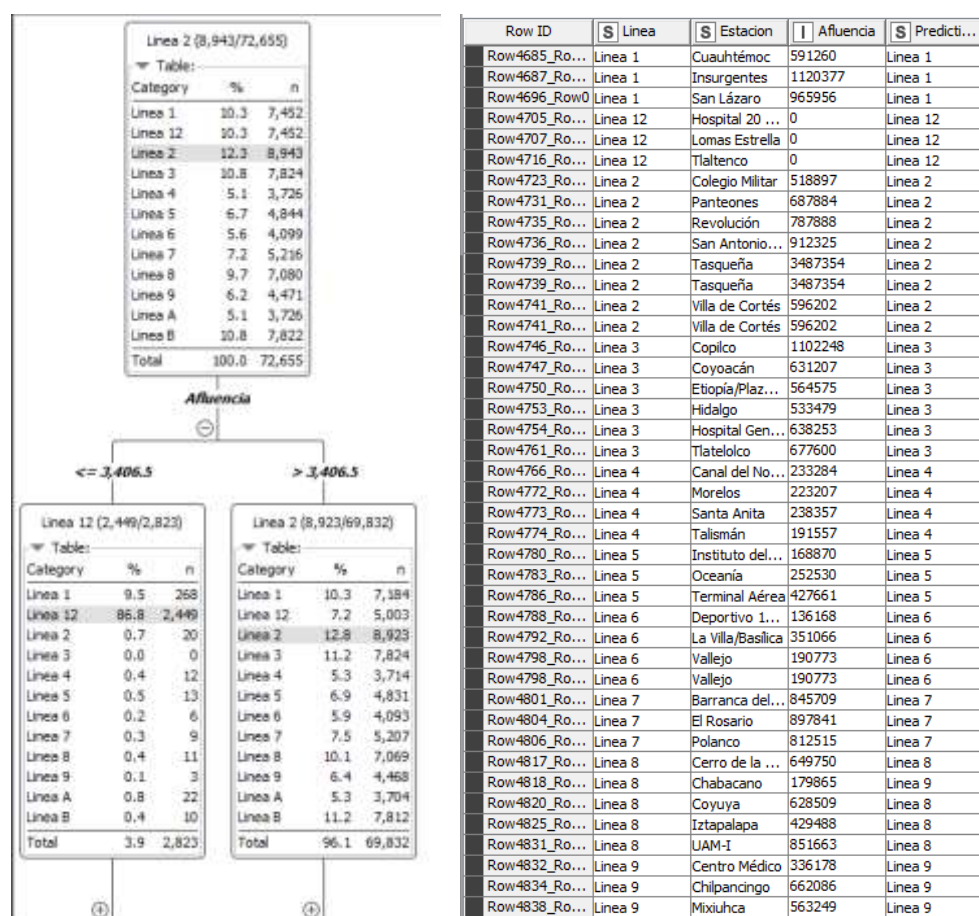
CART



Como se puede observar, el árbol generado es muy grande por lo que se mostrará una imagen parcial de lo que se generó en los resultados.

Statistics - 3:16 - Numeric Scorer	
File	
Can't calculate Mean Absolute Percentage error: target value is 0! Row136	
R ² :	0.73
Mean absolute error:	6,644,481.261
Mean squared error:	104,845,477,214,711.05
Root mean squared error:	10,239,408.05
Mean signed difference:	19,439.341
Mean absolute percentage error:	NaN
Adjusted R ² :	0.73

C4.5



Como se puede observar, el árbol generado es muy grande por lo que se mostrará una imagen parcial de lo que se generó en los resultados y además agregamos una captura de la predicción de datos.

b) Medidas Obtenidas

CART

Statistics - 3:16 - Numeric Scorer	
File	
Can't calculate Mean Absolute Percentage error: target value is 0! Row136	
R ² :	0.73
Mean absolute error:	6,644,481.261
Mean squared error:	104,845,477,214,711.05
Root mean squared error:	10,239,408.05
Mean signed difference:	19,439.341
Mean absolute percentage error:	NaN
Adjusted R ² :	0.73

C4.5

Row ID	\$ Rule	D Record count	D Number of correct
Row1	\$Estación = "Beluena" AND \$Afiliación ≤ 3406.5 => "Línea 1"	21	21
Row2	\$Estación = "Balderrá" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row3	\$Estación = "Boulevard Puerto Aéreo" AND \$Afiliación ≤ 3406.5 => "Línea 1"	23	23
Row4	\$Estación = "Candelaria" AND \$Afiliación ≤ 3406.5 => "Línea 1"	23	23
Row5	\$Estación = "Chapultepec" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row6	\$Estación = "Cuauhtémoc" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row7	\$Estación = "Gómez Parias" AND \$Afiliación ≤ 3406.5 => "Línea 1"	22	22
Row8	\$Estación = "Insurgentes" AND \$Afiliación ≤ 3406.5 => "Línea 1"	2	2
Row9	\$Estación = "Isabel la Católica" AND \$Afiliación ≤ 3406.5 => "Línea 1"	24	24
Row10	\$Estación = "Juarocadán" AND \$Afiliación ≤ 3406.5 => "Línea 1"	3	3
Row11	\$Estación = "Merced" AND \$Afiliación ≤ 3406.5 => "Línea 1"	26	26
Row12	\$Estación = "Moctezuma" AND \$Afiliación ≤ 3406.5 => "Línea 1"	21	21
Row13	\$Estación = "Observatorio" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row14	\$Estación = "Parotlán" AND \$Afiliación ≤ 3406.5 => "Línea 1"	18	18
Row15	\$Estación = "Pino Suárez" AND \$Afiliación ≤ 3406.5 => "Línea 1"	19	19
Row16	\$Estación = "Salto del Agua" AND \$Afiliación ≤ 3406.5 => "Línea 1"	24	24
Row17	\$Estación = "San Lázaro" AND \$Afiliación ≤ 3406.5 => "Línea 1"	20	20
Row18	\$Estación = "Sevilla" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row19	\$Estación = "Tacubaya" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0
Row20	\$Estación = "Zaragoza" AND \$Afiliación ≤ 3406.5 => "Línea 1"	21	21
Row21	\$Estación = "Atilaco" AND \$Afiliación ≤ 3406.5 => "Línea 12"	93	93
Row22	\$Estación = "Calle 11" AND \$Afiliación ≤ 3406.5 => "Línea 12"	141	141
Row23	\$Estación = "Culhuacán" AND \$Afiliación ≤ 3406.5 => "Línea 12"	153	153
Row24	\$Estación = "Eje Central" AND \$Afiliación ≤ 3406.5 => "Línea 12"	89	89
Row25	\$Estación = "Ermita" AND \$Afiliación ≤ 3406.5 => "Línea 12"	92	92
Row26	\$Estación = "Hospital 20 de Noviembre" AND \$Afiliación ≤ 3406.5 => "Línea 12"	92	92
Row27	\$Estación = "Insurgentes Sur" AND \$Afiliación ≤ 3406.5 => "Línea 12"	89	89
Row28	\$Estación = "Lomas Estrella" AND \$Afiliación ≤ 3406.5 => "Línea 12"	142	142
Row29	\$Estación = "Mexicaltzingo" AND \$Afiliación ≤ 3406.5 => "Línea 12"	92	92
Row30	\$Estación = "Mixcoac" AND \$Afiliación ≤ 3406.5 => "Línea 12"	98	98
Row31	\$Estación = "Nepetla" AND \$Afiliación ≤ 3406.5 => "Línea 12"	150	150
Row32	\$Estación = "Olivos" AND \$Afiliación ≤ 3406.5 => "Línea 12"	148	148
Row33	\$Estación = "Parque de los Venados" AND \$Afiliación ≤ 3406.5 => "Línea 12"	95	95
Row34	\$Estación = "Perifoneo Oriente" AND \$Afiliación ≤ 3406.5 => "Línea 12"	138	138
Row35	\$Estación = "San Andrés Tonalá" AND \$Afiliación ≤ 3406.5 => "Línea 12"	139	139
Row36	\$Estación = "Tresasco" AND \$Afiliación ≤ 3406.5 => "Línea 12"	153	153
Row37	\$Estación = "Tlalisco" AND \$Afiliación ≤ 3406.5 => "Línea 12"	158	158
Row38	\$Estación = "Tláhuac" AND \$Afiliación ≤ 3406.5 => "Línea 12"	148	148
Row39	\$Estación = "Zapate" AND \$Afiliación ≤ 3406.5 => "Línea 12"	92	92
Row40	\$Estación = "Zapotlán" AND \$Afiliación ≤ 3406.5 => "Línea 12"	147	147
Row41	\$Estación = "Alameda" AND \$Afiliación ≤ 3406.5 => "Línea 2"	9	9
Row42	\$Estación = "Bellas Artes" AND \$Afiliación ≤ 3406.5 => "Línea 1"	0	0

c) Descripción de las características de los resultados generados

CART

En este caso, encontramos que la precisión en la regresión es un tanto alta, por lo que se puede entender que, a pesar de que existen algunos errores, podemos decir que la confianza en la predicción del árbol es alta.

C4.5

Encontramos un error mínimo, no mayor al 4% en el árbol, por lo que podemos concluir que existe una gran confianza en la precisión del árbol y que por lo tanto una forma de conocer a que tipo de línea pertenecen las estaciones especialmente las que cuentan con múltiples correspondencias pueden ser aplicadas a través del árbol C4.5.

d) Tipo de muestra que utilizó para prueba y entrenamiento

CART

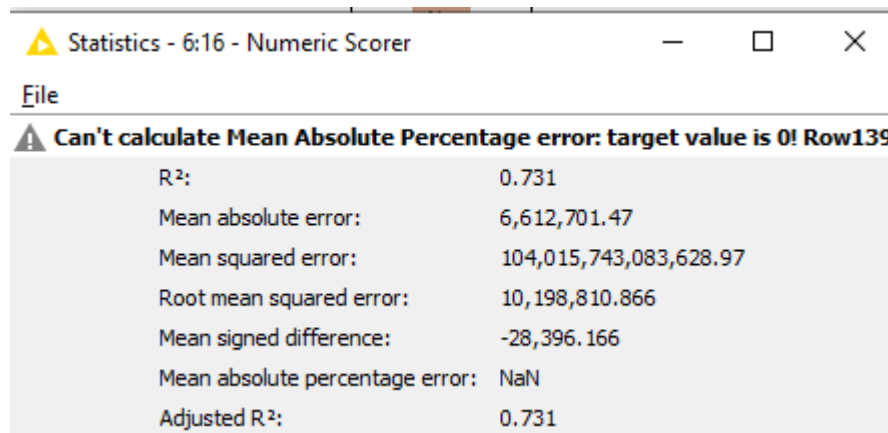
Para este caso en particular no se realiza un particionamiento para prueba y entrenamiento ya que no se realiza algún tipo de configuración como tal. Sin embargo, configuramos el árbol para que partir de cierto número de datos ingresados se puedan hacer las particiones para su predicción.

C4.5

En el caso del árbol C4.5 se asigna un valor para el número mínimo de particiones con el cual, nosotros asignamos un valor mínimo de 10 datos.

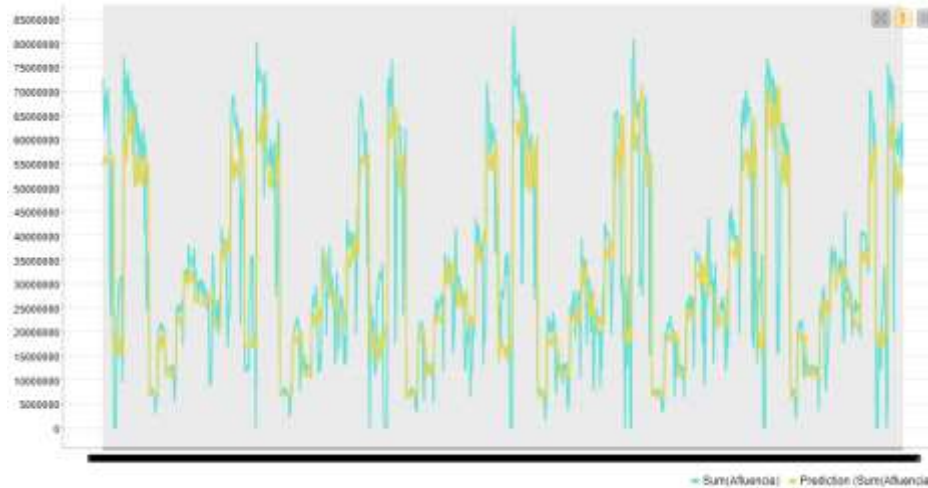
I.4 Análisis de los resultados

CART



R ² :	0.731
Mean absolute error:	6,612,701.47
Mean squared error:	104,015,743,083,628.97
Root mean squared error:	10,198,810.866
Mean signed difference:	-28,396.166
Mean absolute percentage error:	NaN
Adjusted R ² :	0.731

Con base en el análisis previo, se puede llegar a la conclusión de que el coeficiente de determinación (r^2) representa el nivel de precisión de nuestro árbol de decisiones. En este caso, se observa que el árbol presenta una precisión moderada, dado que su valor de estimación es de 0.731, el cual se encuentra muy próximo a uno. Para una mejor comprensión, se adjunta a continuación una gráfica que compara los valores originales con aquellos predichos por nuestro árbol.



Es importante destacar que este árbol de decisiones tiene la capacidad de determinar la afluencia generada por una línea de metro en un mes específico del año. Esto significa que el modelo es capaz de predecir la cantidad de pasajeros que se espera utilizarán dicha línea durante ese período. Esta información resulta valiosa para la planificación y gestión eficiente de los recursos y servicios del sistema de transporte.

C4.5

En este ejercicio podemos concluir que, a través de la predicción del árbol, hallamos una alta precisión en la predicción de la pertenencia de una estación con respecto a su línea (en especial las de más de una correspondencia).

Linea \ Pre...	Linea 1	Linea 6	Linea 9	Linea 8	Linea 5	Linea 7	Linea 3	Linea 4	Linea 2	Linea 8	Linea 12	Linea A
Linea 1	7896	0	76	115	30	9	18	19	49	67	0	21
Linea 6	0	4342	0	0	37	42	0	133	0	0	0	0
Linea 9	98	0	4599	18	135	0	0	86	6	0	0	26
Linea 8	54	0	22	7612	0	0	0	34	25	36	83	0
Linea 5	78	83	70	0	5044	0	0	64	0	21	0	22
Linea 7	0	83	0	0	0	5661	0	0	0	0	52	0
Linea 3	56	3	120	0	15	0	8411	0	10	40	39	0
Linea 4	0	30	0	12	95	0	0	3910	0	53	0	0
Linea 2	100	0	13	147	0	3	55	0	9547	0	71	0
Linea 8	6	0	0	12	124	0	2	0	0	8547	0	0
Linea 12	0	0	0	26	0	11	0	0	2	0	8241	0
Linea A	20	0	39	0	73	0	0	0	0	0	0	4008
Correct classified: 77,818												
Accuracy: 96.396%												
Cohen's kappa (κ): 0.96%												
Wrong classified: 2,909												
Error: 3.604%												

Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
Linea 1	7896	412	72035	384	0.954	0.95	0.954	0.994	0.952	?	?
Linea 6	4342	199	75974	212	0.953	0.956	0.953	0.997	0.955	?	?
Linea 9	4599	340	75419	369	0.926	0.931	0.926	0.996	0.928	?	?
Linea 8	7612	330	72531	254	0.968	0.958	0.968	0.995	0.963	?	?
Linea 5	5044	489	74856	338	0.937	0.912	0.937	0.994	0.924	?	?
Linea 7	5661	65	74866	135	0.977	0.989	0.977	0.999	0.983	?	?
Linea 3	8411	75	71958	283	0.967	0.991	0.967	0.999	0.979	?	?
Linea 4	3910	336	76251	230	0.944	0.921	0.944	0.996	0.933	?	?
Linea 2	9547	92	70699	389	0.961	0.99	0.961	0.999	0.975	?	?
Linea 8	8547	257	71779	144	0.983	0.971	0.983	0.996	0.977	?	?
Linea 12	8241	245	72202	39	0.995	0.971	0.995	0.997	0.983	?	?
Linea A	4008	69	76518	112	0.968	0.983	0.968	0.999	0.976	?	?
Overall	?	?	?	?	?	?	?	?	?	0.964	0.96

II. Multi Clasificación.

Bagging y Boosting

Equipo 4

Data Mining

3CV15

II. Multi Clasificación. Bagging y Boosting

II.1 Descripción del ejercicio

Random Forest

Los modelos Random Forest están formados por un conjunto de árboles de decisión individuales, cada uno entrenado con una muestra ligeramente distinta de los datos de entrenamiento generados mediante bootstrapping.

La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

Boosting

También llamada potenciación del gradiente es una técnica de machine learning utilizada para el análisis de la regresión y en la clasificación estadística; este produce un modelo predictivo; Al igual que otros modelos de boosting, va construyendo nuevos modelos considerando en cada nueva iteración los errores cometidos anteriormente.

II.2 Diccionario de datos

II.2.1 Diccionario de datos Random Forest.

Nombre	Significado	Tipo	Dominio
Línea	Número de línea del sistema del transporte público.	Categórico	Línea 1 – Línea 12 Línea A Línea B
Estacion	Nombre de la estación donde se encuentra la afluencia.	Categórico	Nombre de la estación.
Afluencia	Número de personas que utilizaron esa estación en particular en el día especificado.	Numérico	Numero entero Positivo.
Año	La fecha en la que se registró la afluencia.	Numérico	2010 – 2023
Mes	El mes correspondiente a la fecha.	Categórico	Enero – Diciembre

El conjunto de datos utilizado en este estudio está compuesto por los conjuntos de ingresos y afluencia. Para llevar a cabo el análisis, se aplicó una técnica de filtrado de columnas (Column Filter) con el fin de seleccionar y separar los atributos relevantes para el estudio, que incluyen la línea del

metro, la estación, la afluencia, así como los campos relacionados con los años y meses. Esta acción de filtrado permitió aislar y enfocar los datos necesarios para realizar el análisis específico requerido, descartando aquellas variables que no eran reelevantes para el ejercicio en cuestión.

II.2.3 Diccionario de datos Gradient Boost.

Nombre	Significado	Tipo	Dominio
Genders	Genero de la persona	Categórico	Masculino Femenino
Married	Estado civil	Categórico	Soltero/ divorciado Casado
BankCustomer	Cliente del banco	Categórico	No tiene cuenta Tiene cuenta
Industry	Sector de la industria en el que trabaja	Categórico	Nombre del sector
Ethnicity	Etnia del solicitante	Categórico	Asiático, Latino, Blanco, Negro, etc.
PriorDefault	Valor predeterminado con anterioridad.	Categórico	Sin valor Con valor
Employed	Estado laboral	Categórico	Con empleo Desempleado
DriversLicense	Licencia de conducir	Categórico	No tiene licencia Tiene licencia
Citizen	Ciudadanía del solicitante	Categórico	Por nacimiento Por otras vías
Approved	Aprobación de expedición de tarjeta	Categórico	No aprobado Aprobado

II.3. Resultados.

a) Diagrama generado

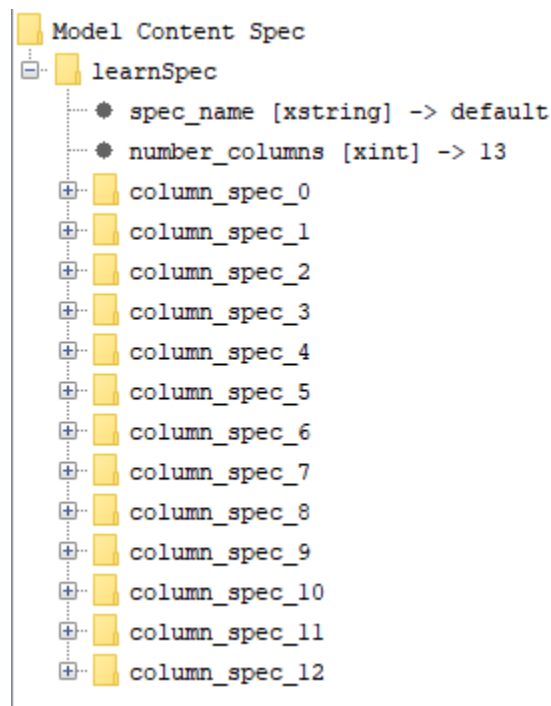
Random Forest

Aquí tenemos uno de los 100 modelos que se generan.



R ² :	0.569
Mean absolute error:	268,680.41
Mean squared error:	137,976,742,572.67
Root mean squared error:	371,452.208
Mean signed difference:	3,397.829
Mean absolute percentage error:	NaN
Adjusted R ² :	0.569

Gradient Boost



b) Medidas obtenidas

Random Forest

Statistics - 6:25 - Numeric Scorer

File Edit Hilite Navigation View

Table "Scores" - Rows: 7 Spec - Column: 1 Properties Flow Variables

Row ID	D Predicti...
R^2	0.559
mean absolut...	268,180.983
mean square...	138,975,62...
root mean sq...	372,794.345
mean signed ...	6,148.226
mean absolut...	NaN
adjusted R^2	0.559

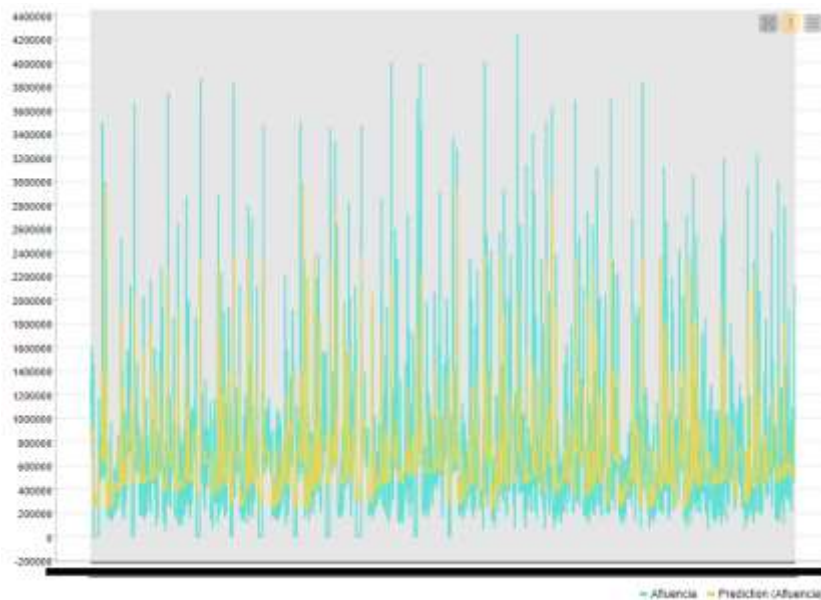
Gradient Boost

Debido a que knime no genera el modelo de clasificación no es posible colocar el diagrama y por ende tampoco es posible obtener medidas, ni describir resultados. Es por eso por lo que solo mostramos captura de las carpetas que generan el modelo en la sección del diagrama

c) Descripción de las características

Random Forest

Al tener un valor de r^2 con porcentajes cercanos al 50% de precisión podemos observar que aún existen errores en que pueden afectar en la predicción del modelo y que aún faltarían por analizar más información para generar un modelo con una mayor precisión y confianza.



Gradient Boost

En este caso percibimos que existe un porcentaje de precisión alto y por lo tanto se puede tener una gran confianza en que la predicción para saber si una tarjeta de crédito es aceptada o no.

Approved ...	APPROVED	NOT APPR...
APPROVED	51	10
NOT APPRO...	5	72

Correct classified: 123	Wrong classified: 15
Accuracy: 89.13%	Error: 10.87%
Cohen's kappa (κ): 0.778%	

d) Tipo de muestra que se utilizó para prueba y entrenamiento

Random Forest

First partition | Flow Variables | Job Manager Selection | Memory Policy

Choose size of first partition

☐ Absolute 100

☒ Relative[%] 80

☐ Take from top

☐ Linear sampling

☒ Draw randomly

☐ Stratified sampling S Mes

☐ Use random seed 1,687,236,246

OK Apply Cancel ?

Se asignó un porcentaje de 80% para entrenamiento y un 20% para prueba

Gradient Boost

First partition | Flow Variables | Job Manager Selection | Memory Policy

Choose size of first partition

☐ Absolute 100

☒ Relative[%] 80

☐ Take from top

☐ Linear sampling

☐ Draw randomly

☒ Stratified sampling S Approved

☐ Use random seed 1,687,236,342

OK Apply Cancel ?

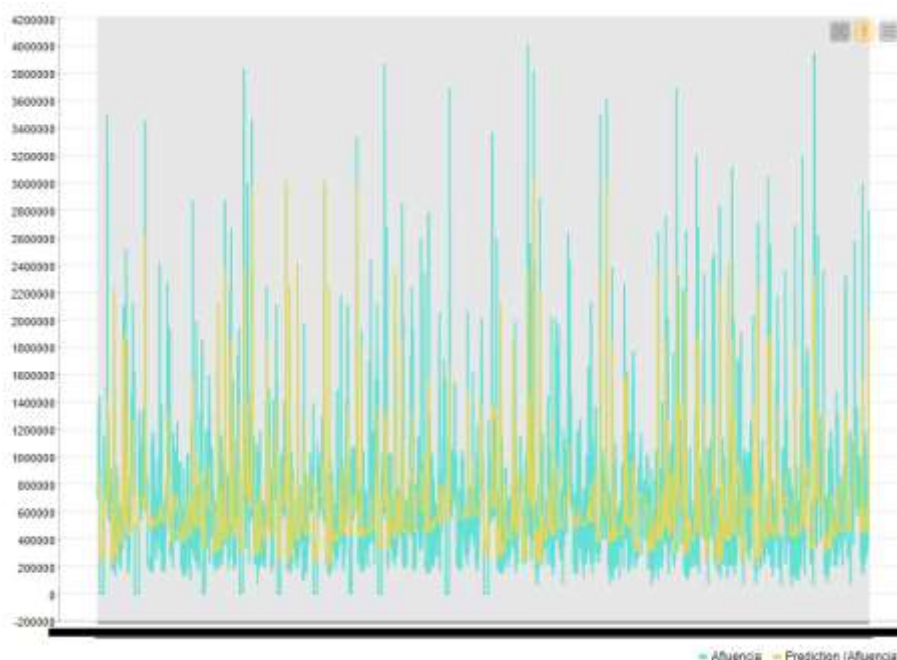
Se asignó un porcentaje de 80% para entrenamiento y un 20% para prueba

II.4. Análisis de los resultados.

Random Forest

Statistics - 6:25 - Numeric Scorer	
File	
Can't calculate Mean Absolute Percentage error: target value is 0! Row470...	
R ² :	0.559
Mean absolute error:	268,180.983
Mean squared error:	138,975,623,479.899
Root mean squared error:	372,794.345
Mean signed difference:	6,148.226
Mean absolute percentage error:	NaN
Adjusted R ² :	0.559

Con base en el análisis previo, se puede llegar a la conclusión de que el coeficiente de determinación (r^2) representa el nivel de precisión de nuestro árbol de decisiones. En este caso, se observa que el árbol presenta una precisión mediana, dado que su valor de estimación es de 0.559, es decir poco más de la mitad. Para una mejor comprensión, se adjunta a continuación una gráfica que compara los valores originales con aquellos predichos por nuestro árbol.



Al analizar los resultados obtenidos, se puede observar que nuestras predicciones presentan una exactitud que oscila entre aproximadamente entre la mitad y un poco más de la mitad de la precisión de nuestros datos. Es relevante destacar que nuestro modelo de Random Forest ha sido aplicado para calcular la afluencia de cada estación del sistema de metro en un mes específico. Esta técnica nos permite realizar estimaciones y pronósticos respecto a la cantidad de personas que utilizarán cada estación durante ese período.

Gradient Boost

Approved ...	APPROVED	NOT APPR...
APPROVED	51	10
NOT APPRO...	5	72

Correct classified: 123	Wrong classified: 15
Accuracy: 89.13%	Error: 10.87%
Cohen's kappa (κ):	

Por otro lado, con relación al algoritmo de Boosting, se ha aplicado para predecir dadas las características de algunas personas que pidieron algún tipo de tarjeta de crédito y saber si es que se les fueron aprobadas o no. Al analizar los resultados obtenidos en este ejercicio, se puede observar que la precisión del modelo es del 89.13%. Esta alta precisión indica que el modelo es capaz de realizar predicciones con un nivel muy elevado de exactitud y confiabilidad.

Además, es importante destacar que el error del modelo es un tanto menor, aproximadamente del 10.87%. Esta baja tasa de error nos indica que el modelo es muy preciso y confiable en la asignación de las aprobaciones y negaciones de tarjetas de crédito. En consecuencia, podemos afirmar que el modelo es capaz de realizar predicciones con un alto nivel de exactitud y confiabilidad, lo cual es fundamental para la toma de decisiones y la solución de problemas.

III. Agrupamiento

(Jerárquico y No Jerárquico)

Equipo 4
Data Mining
3CV15

III. Agrupamiento (Jerárquico y No Jerárquico)

III.1 Descripción del ejercicio

Para llevar a cabo el análisis de agrupamiento jerárquico, utilizaremos el conjunto de Fifa World Cup. Este conjunto será sometido a un proceso de agrupamiento jerárquico con el objetivo de identificar estructuras y relaciones entre los elementos. Por otro lado, para el análisis de agrupamiento no jerárquico, continuaremos empleando el conjunto de datos con el que hemos trabajado previamente. Este conjunto se utilizará para aplicar técnicas de agrupamiento no jerárquico con el fin de descubrir patrones y relaciones entre los elementos, sin una estructura jerárquica específica.

III.2. Diccionario de Datos

III.2.1 Diccionario de Datos Jerárquico

Conjunto de datos FIFA World Cup Attendance 1930-2022
Mediante la utilización del nodo Column Filter, procederemos a aplicar un filtrado de datos con el objetivo de reducir y seleccionar el conjunto de datos de manera precisa. Esta operación permitirá obtener un subconjunto de datos que se ajuste a los criterios y atributos deseados para nuestro análisis.

De esta manera, garantizaremos que el conjunto de datos final utilizado en nuestro trabajo se encuentre adecuadamente filtrado y optimizado para cumplir con los objetivos y requerimientos específicos del estudio.

Nombre	Significado	Tipo	Dominio
Host	Anfitrión de la Copa del Mundo	Categorico	Sede elegida por la FIFA
Total_Attendance	Asistencia total de personas	Numérico	Numero entero Positivo
Year	El año que se disputo el mundial	Numérico	1930 – 2022
Average_Attendance	Promedio de personas con asistencia en vivo	Numérico	Numero entero Positivo
Matches	Número total de partidos Jugados	Numerico	Numero entero Positivo

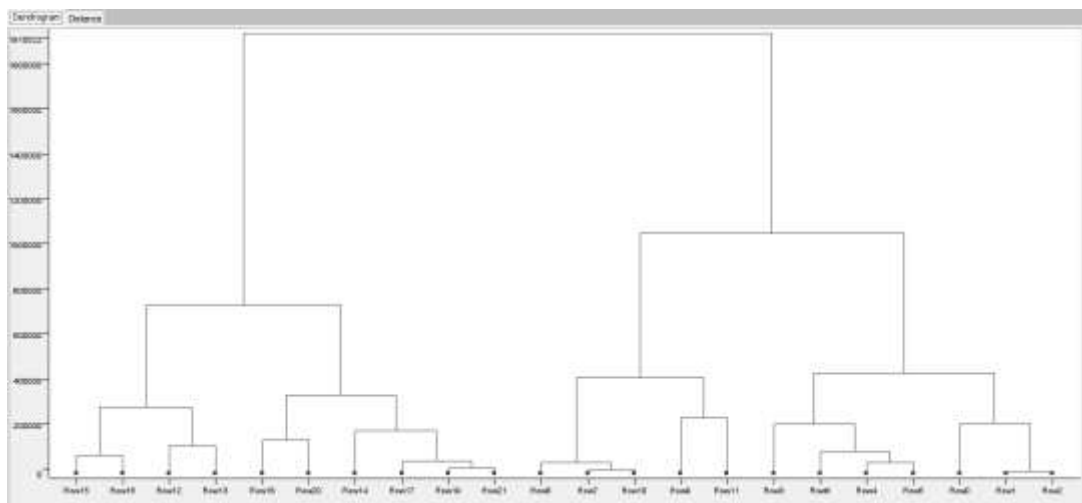
III.2.2 Diccionario No Jerárquico

Conjunto de datos Ingresos y Afluencia.

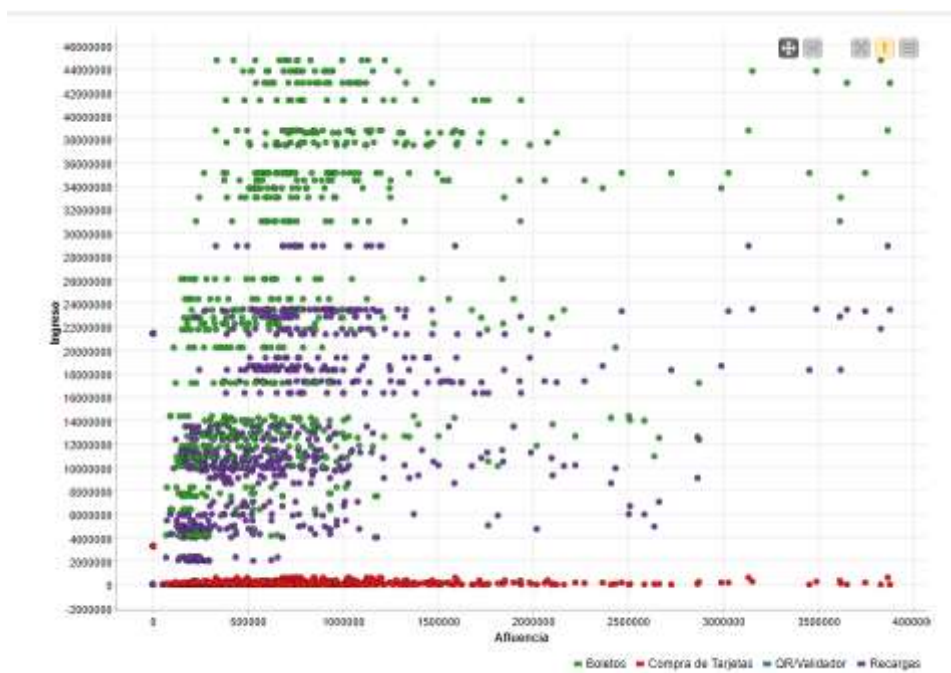
Nombre	Significado	Tipo	Dominio
Línea	Número de línea del sistema del transporte público.	Categórico	Línea 1 – Línea 12 Línea A Línea B
Estacion	Nombre de la estación donde se encuentra la afluencia.	Categórico	Nombre de la estación.
Afluencia	Número de personas que utilizaron esa estación en particular en el día especificado.	Numérico	Numero entero Positivo.
TipoIngreso	El valor numérico que representa la cantidad de ingresos registrados para esa línea en particular en el día especificado.	Categórico	Forma de pago.
Año	La fecha en la que se registró la afluencia.	Numérico	2010 – 2023
Mes	El mes correspondiente a la fecha.	Categórico	Enero – Diciembre
Ingreso	El valor numérico que representa la cantidad de ingresos registrados para esa línea en particular en el día especificado.	Numérico	Numero entero positivo.

III.3. Resultados.

a) Diagrama generado

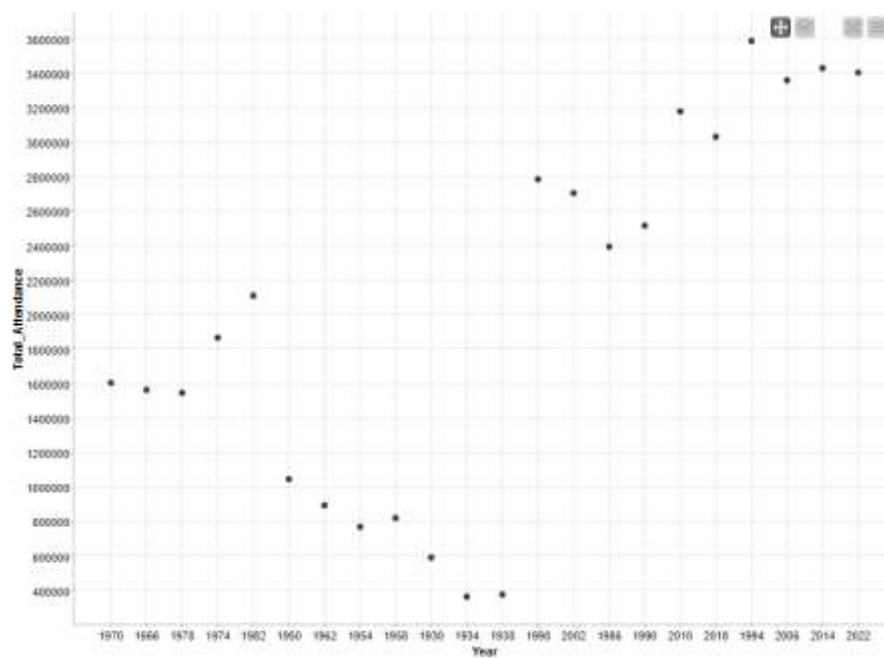
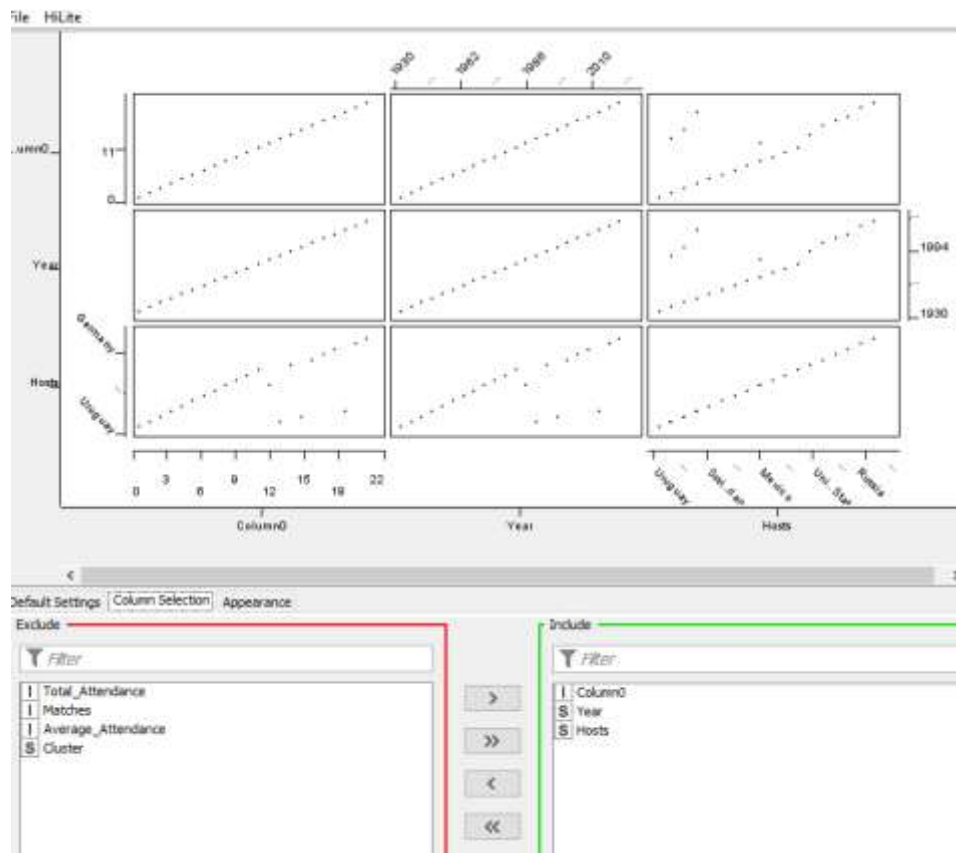
Jerárquico

No Jerárquico

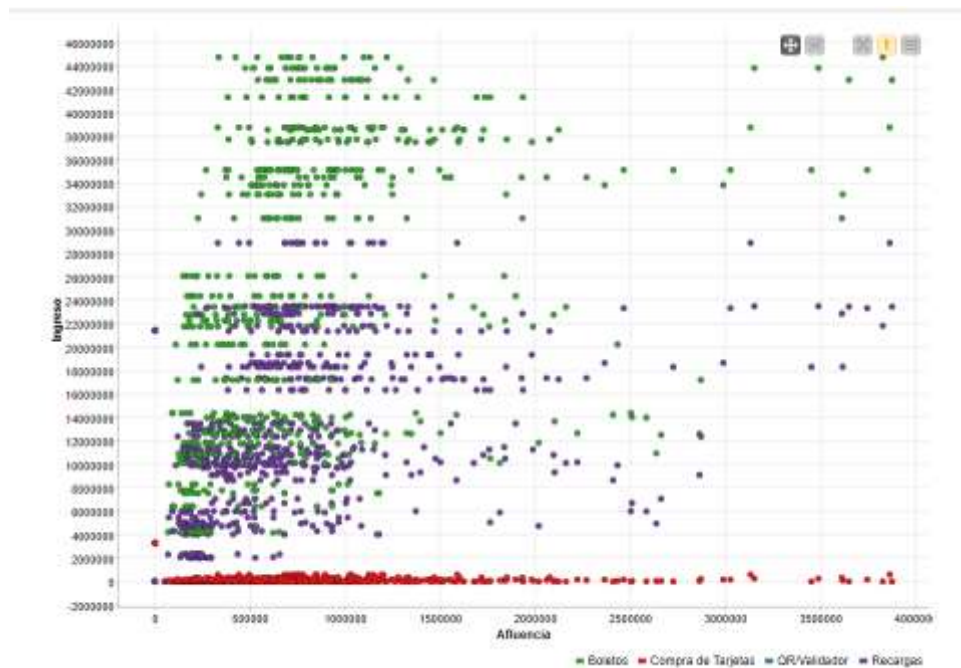
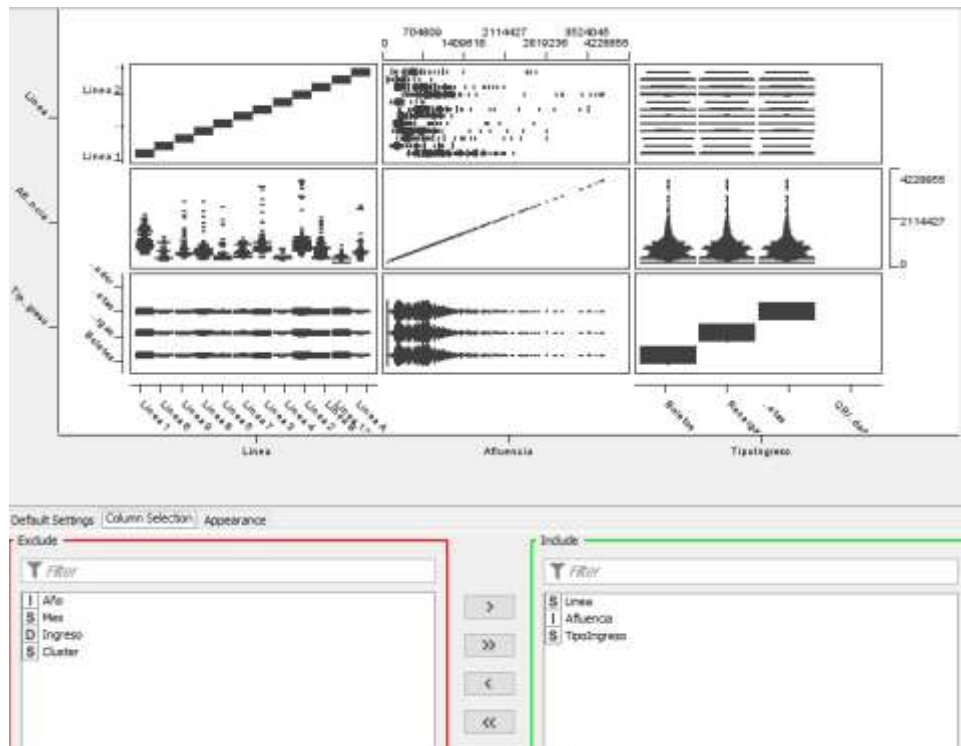


b) Medidas obtenidas.Silueta.

Silueta modelo Jerárquico



Silueta modelo No Jerarquico

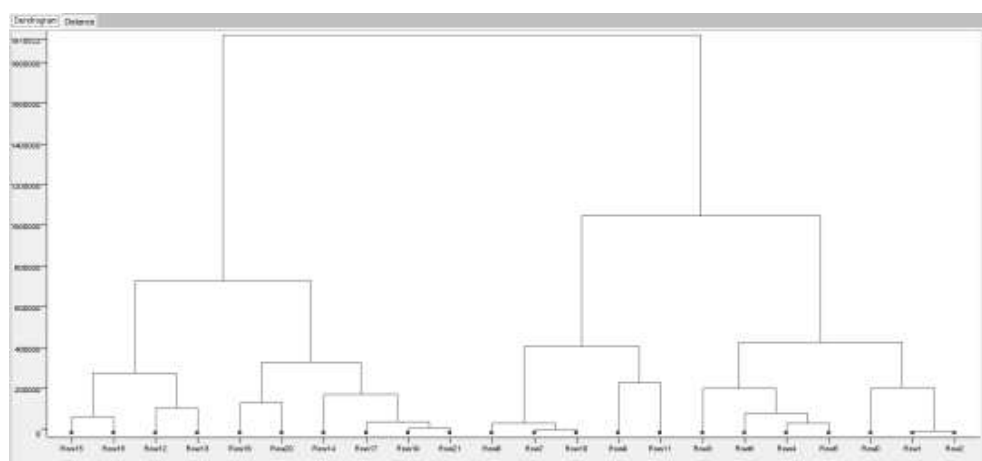


c) Descripción de las características de los resultados generados

Dendrograma (Jerárquico)

El dendrograma nos brinda la posibilidad de visualizar las variaciones en la asistencia de público entre las distintas ediciones. Esta representación gráfica nos permite analizar y comparar de manera sistemática la diferencia en términos de afluencia entre cada Copa Mundial.

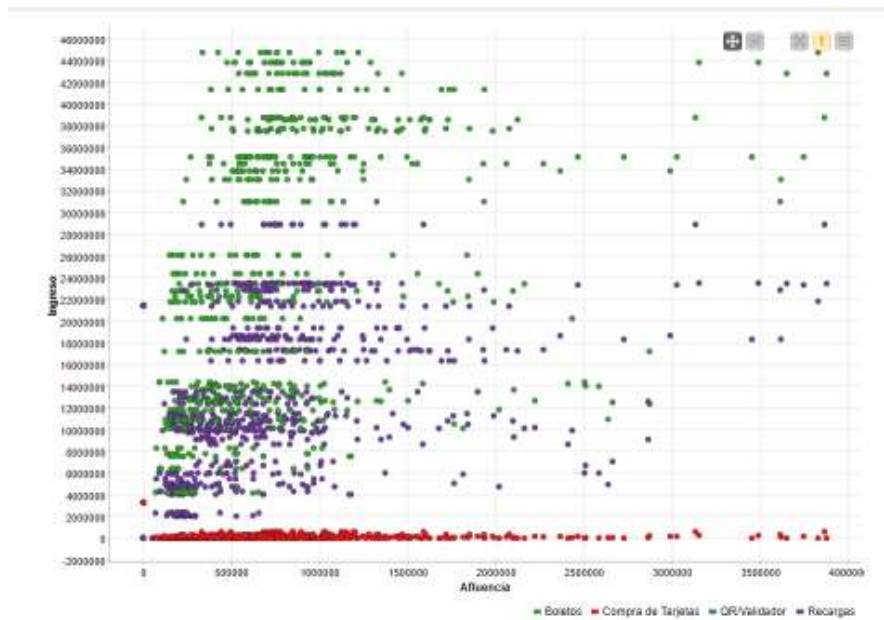
En el dendrograma, se pueden identificar los diferentes niveles de agrupamiento y la distancia relativa entre ellos. Estas agrupaciones reflejan similitudes o disparidades en las asistencias registradas en cada edición del torneo.



No Jerárquico (Scatter Plot)

En el Scatter Plot presente, se muestra la relación entre la afluencia en el sistema de transporte público del metro y los ingresos percibidos. Asimismo, se pueden identificar los distintos grupos de pago generados en función de dicha afluencia.

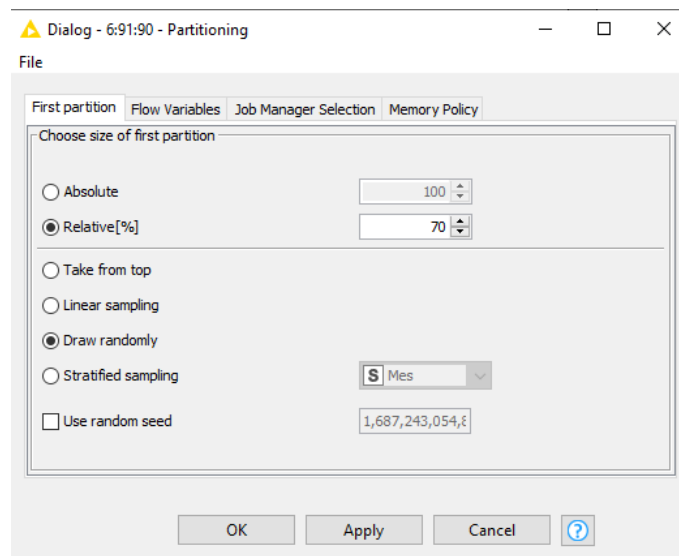
El gráfico permite visualizar la relación entre estas dos variables, lo cual nos proporciona información sobre cómo la afluencia en el metro puede influir en los ingresos generados. Al observar el gráfico, es posible identificar patrones y tendencias en la distribución de los grupos de pago en relación con la afluencia registrada.



d) Tipo de muestra que utilizó para prueba y entrenamiento.

No Jerárquico

Con relación al conjunto de datos no jerárquico, se realizó una partición de los datos con el propósito de llevar a cabo un proceso de entrenamiento y evaluación del modelo. Específicamente, se asignó el 70% de los datos para el entrenamiento del modelo, con el objetivo de obtener resultados más precisos y robustos. El 30% restante de los datos se reservó para realizar pruebas y evaluar el rendimiento del modelo en datos no vistos previamente. Esta estrategia de partición de datos se llevó a cabo con el fin de evitar el sobreajuste del modelo y evaluar su capacidad para generalizar y realizar predicciones precisas en datos nuevos.



III.4 Análisis de Resultados

No Jerárquico

A través del análisis anterior, podemos observar que nuestro enfoque de agrupamiento no jerárquico revela la cantidad de ingresos percibidos en función de la afluencia, y los clasifica en grupos según su forma de pago. Por consiguiente, podemos inferir que la forma de pago más lucrativa en el Metro de la ciudad es a través de la venta de boletos. Esto nos lleva a deducir que mucha gente confía más en este tipo de pago, ya que los ingresos generados por la compra de tarjetas y otros métodos de pago no son equivalentes a los obtenidos por la venta de boletos.

Jerarquico

A través de análisis previo podemos inferir que el análisis jerárquico nos permite examinar el nivel de asistencia en cada Copa Mundial de la FIFA. Este enfoque nos revela patrones o tendencias en términos de la cantidad de espectadores presentes en cada edición del torneo, identificando picos de asistencia en eventos específicos o variaciones en la participación a lo largo del tiempo.

IV. Reglas de Asociación

Equipo 4

Data Mining

3CV15

IV. Reglas de Asociación

IV.1 Descripción del ejercicio

Las reglas de asociación son un conjunto de técnicas que permiten establecer o encontrar relaciones de nuestro interés en un conjunto de datos, estas son utilizadas en su mayoría en áreas mercadológicas para recomendar algún producto o servicio que pueda ser de interés del comprador.

A través de este estudio se pretende encontrara las reglas de asociación fuertes para poder hallar las reglas para poder determinar si el banco aprobara o no la expedición de una tarjeta de crédito a los solicitantes, mostrando todos los antecedentes que conllevan a la respuesta de nuestro algoritmo.

IV.2 Diccionario de Datos

Nombre	Significado	Tipo	Dominio
Genders	Genero de la persona	Categórico	Masculino Femenino
Married	Estado civil	Categórico	Soltero/ divorciado Casado
BankCustomer	Cliente del banco	Categórico	No tiene cuenta Tiene cuenta
Industry	Sector de la industria en el que trabaja	Categórico	Nombre del sector
Ethnicity	Etnia del solicitante	Categórico	Asiático, Latino, Blanco, Negro, etc.
PriorDefault	Valor predeterminado con anterioridad.	Categórico	Sin valor Con valor
Employed	Estado laboral	Categórico	Con empleo Desempleado
DriversLicense	Licencia de conducir	Categórico	No tiene licencia Tiene licencia
Citizen	Ciudadanía del solicitante	Categórico	Por nacimiento Por otras vías
Approved	Aprobación de expedición de tarjeta	Categórico	No aprobado Aprobado

IV.3 Resultados

a) Diagrama generado

Row ID	S combined string
Row0	"Male", "Married", "has a bank account", "Industrials", "White", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row1	"Female", "Married", "has a bank account", "Materials", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row2	"Female", "Married", "has a bank account", "Materials", "Black", "prior default", "not employed", "no license", "ByBirth", "APPROVED"
Row3	"Male", "Married", "has a bank account", "Industrials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row4	"Male", "Married", "has a bank account", "Industrials", "White", "prior default", "not employed", "no license", "ByOtherMeans", "APPROVED"
Row5	"Male", "Married", "has a bank account", "CommunicationServices", "White", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row6	"Male", "Married", "has a bank account", "Transport", "Black", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row7	"Female", "Married", "has a bank account", "InformationTechnology", "White", "prior default", "not employed", "no license", "ByBirth", "APPROVED"
Row8	"Male", "Single/Divorced/etc", "does not have a bank account", "Financials", "Black", "prior default", "not employed", "no license", "ByBirth", "APPROVED"
Row9	"Male", "Single/Divorced/etc", "does not have a bank account", "Industrials", "White", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row10	"Male", "Married", "has a bank account", "Energy", "Black", "no prior defaults", "not employed", "has license", "ByBirth", "APPROVED"
Row11	"Male", "Married", "has a bank account", "Energy", "Black", "prior default", "not employed", "no license", "ByBirth", "APPROVED"
Row12	"Female", "Married", "has a bank account", "Financials", "White", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row13	"Male", "Married", "has a bank account", "Financials", "White", "no prior defaults", "not employed", "no license", "ByBirth", "APPROVED"
Row14	"Female", "Married", "has a bank account", "Materials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row15	"Male", "Single/Divorced/etc", "does not have a bank account", "Financials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row16	"Male", "Married", "has a bank account", "CommunicationServices", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row17	"Female", "Married", "has a bank account", "Materials", "White", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row18	"Male", "Married", "has a bank account", "Real Estate", "Black", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row19	"Female", "Married", "has a bank account", "InformationTechnology", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row20	"Male", "Married", "has a bank account", "Energy", "White", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row21	"Male", "Married", "has a bank account", "Energy", "White", "prior default", "not employed", "no license", "ByOtherMeans", "APPROVED"
Row22	"Female", "Married", "has a bank account", "Energy", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row23	"Female", "Married", "has a bank account", "Utilities", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row24	"Female", "Married", "has a bank account", "Materials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row25	"Female", "Married", "has a bank account", "Energy", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row26	"Female", "Married", "has a bank account", "ConsumerDiscretionary", "Asian", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row27	"Male", "Married", "has a bank account", "Real Estate", "Asian", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row28	"Male", "Married", "has a bank account", "Education", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row29	"Male", "Married", "has a bank account", "Industrials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row30	"Male", "Married", "has a bank account", "ConsumerStaples", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row31	"Male", "Married", "has a bank account", "Utilities", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row32	"Male", "Married", "has a bank account", "ConsumerDiscretionary", "Asian", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row33	"Female", "Married", "has a bank account", "Education", "White", "prior default", "not employed", "has license", "ByBirth", "APPROVED"
Row34	"Female", "Married", "has a bank account", "Materials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row35	"Male", "Married", "has a bank account", "Industrials", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row36	"Male", "Married", "has a bank account", "InformationTechnology", "Black", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row37	"Female", "Married", "has a bank account", "Utilities", "Black", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row38	"Male", "Single/Divorced/etc", "does not have a bank account", "InformationTechnology", "White", "prior default", "employed", "no license", "ByBirth", "APPROVED"
Row39	"Male", "Married", "has a bank account", "Healthcare", "Latino", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row40	"Male", "Married", "has a bank account", "Energy", "White", "prior default", "employed", "has license", "ByBirth", "APPROVED"
Row41	"Male", "Married", "has a bank account", "Energy", "Black", "prior default", "employed", "no license", "ByBirth", "APPROVED"

b) Medidas obtenidas

S	Consequent	Antecedent	I	ItemSupport	D	RelativeItemSet	D	RuleConfidence%	D	AbsoluteRuleConf.	D	RelativeRuleConf.	D	RuleFth	D	RuleFth%	D	AbsoluteRule...	D	...
Total Estate	Has license, "no prior defaults", "NOT APPROVED", ...		8	1.159	10.7	75	10.9	2.453	245.23	30	4.346									
Total Estate	Has license, "no prior defaults", "NOT APPROVED", ...		30	1.449	10.3	97	14.1	2.371	237.11	30	4.346									
Total Estate	Has license, "no prior defaults", "not employed", ...		9	1.304	10.8	83	12	2.494	249.4	30	4.346									
Total Estate	Has license, "no prior defaults", "not employed", ...		9	1.304	12.9	70	10.1	2.957	295.71	30	4.346									
Total Estate	Has license, "no prior defaults", "not employed", ...		9	1.304	12.9	70	10.1	2.957	295.71	30	4.346									
Total Estate	Has license, "no prior defaults", "not employed", ...		9	1.304	12.5	72	10.4	2.875	287.5	30	4.346									
Total Estate	Has license, "no prior defaults", "not employed", ...		11	1.594	10.4	106	15.4	2.307	230.68	30	4.346									
Total Estate	Has license, "no prior defaults", "White", ...		9	1.304	10.4	95	12.1	2.435	243.55	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		30	1.449	12.5	80	11.8	2.875	287.5	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		30	1.449	10.2	96	14.1	2.347	234.68	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		8	1.159	11.4	70	10.1	2.629	262.06	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		12	1.739	13.5	99	12.9	2.901	290.11	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		8	1.159	11.4	70	10.1	2.629	262.06	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		12	1.739	13.5	99	12.9	2.901	290.11	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		8	1.159	11.4	70	10.1	2.629	262.06	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		12	1.739	13.5	99	12.9	2.901	290.11	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		13	1.894	10.1	129	10.7	2.318	231.78	30	4.346									
Total Estate	Has license, "NOT APPROVED", "White", ...		9	1.304	11.4	79	11.4	2.62	262.03	30	4.346									
Total Estate	Has license, "NOT APPROVED", "White", ...		9	1.304	11.4	79	11.4	2.62	262.03	30	4.346									
Total Estate	Has license, "NOT APPROVED", "White", ...		9	1.304	11.4	79	11.4	2.62	262.03	30	4.346									
Total Estate	Has license, "NOT APPROVED", "Male", ...		9	1.304	10.8	83	12	2.494	249.4	30	4.346									
Total Estate	Has license, "NOT APPROVED", "Male", ...		9	1.304	10.8	83	12	2.494	249.4	30	4.346									
Total Estate	Has license, "NOT APPROVED", "Male", ...		9	1.304	10.8	83	12	2.494	249.4	30	4.346									
Total Estate	Has license, "not employed", "White", ...		9	1.304	11.1	91	11.7	2.556	255.56	30	4.346									
Total Estate	Has license, "not employed", "White", ...		9	1.304	12.2	74	10.7	2.797	279.77	30	4.346									
Total Estate	Has license, "not employed", "White", ...		9	1.304	12.2	74	10.7	2.797	279.77	30	4.346									
Total Estate	Has license, "not employed", "White", ...		9	1.304	12.2	74	10.7	2.797	279.77	30	4.346									
Total Estate	Has license, "not employed", "White", ...		12	1.739	11.4	105	10.2	2.629	262.06	30	4.346									
Total Estate	Has license, "not employed", "Male", ...		10	1.449	10.5	95	13.8	2.421	242.11	30	4.346									
Total Estate	Has license, "not employed", "Male", ...		10	1.449	10.5	95	13.8	2.421	242.11	30	4.346									
Total Estate	Has license, "not employed", "Male", ...		10	1.449	10.4	96	13.9	2.396	239.58	30	4.346									
Total Estate	Has license, "not employed", "Married", ...		13	1.894	10.3	126	10.3	2.373	237.3	30	4.346									
Total Estate	Has license, "not employed", "Married", ...		13	1.894	10.3	126	10.3	2.373	237.3	30	4.346									
Total Estate	Has license, "not employed", "has a bank account", ...		13	1.894	10.2	128	10.6	2.336	233.59	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		9	1.304	11.5	78	11.3	2.654	265.38	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		9	1.304	11.5	78	11.3	2.654	265.38	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		11	1.594	12.2	90	13	2.811	281.11	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		13	1.894	12.3	106	15.4	2.821	282.08	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		11	1.594	12.2	90	13	2.811	281.11	30	4.346									
Total Estate	Has license, "NOT APPROVED", "not employed", ...		13	1.894	12.3	106	15.4	2.821	282.08	30	4.346									

c) Descripción de las características de los resultados generados

Table "default" - Rows: 10			Spec - Columns: 2		
Row ID	\$ Conseq...	\$ Antece...	Row ID	\$ Consequent	\$ Antecedent
Row20_1	"APPROVED"	"Female"	Row11_1	"NOT APPROVED"	"does not have a bank account"
Row20_2	"APPROVED"	"Married"	Row11_2	"NOT APPROVED"	"Single/Divorced/etc"
Row20_3	"APPROVED"	"has a bank ..."	Row11_3	"NOT APPROVED"	"not employed"
Row20_4	"APPROVED"	"ByBirth"	Row11_4	"NOT APPROVED"	"ByBirth"
Row50_1	"APPROVED"	"employed"	Row14_3	"NOT APPROVED"	"Male"
Row50_2	"APPROVED"	"has license"	Row24_1	"NOT APPROVED"	"Female"
Row50_3	"APPROVED"	"prior default"	Row24_2	"NOT APPROVED"	"Married"
Row87_3	"APPROVED"	"no license"	Row24_3	"NOT APPROVED"	"has a bank account"
Row125_3	"APPROVED"	"White"	Row913_1	"NOT APPROVED"	"has license"
Row156_3	"APPROVED"	"Male"	Row932_2	"NOT APPROVED"	"White"
			Row1042_1	"NOT APPROVED"	"no prior defaults"
			Row1042_2	"NOT APPROVED"	"no license"

Tarjetas aprobadas

/

Tarjetas no aprobadas

d) Tipo de muestra que se utilizó para prueba y entrenamiento

Options Advanced Settings Flow Variables Job Manager Selection Memory Policy

Item column: [...] combined string_SplitResultSet

Item set settings

Minimum set size: 5

☒ Limit set size

Maximum set size: 10

Minimum support: 15.0

☐ Absolute number ☒ Percentage

Minimum rule confidence: 40.0

☐ Sort antecedent list

Se configura la regla para que cuente con un soporte de 15 y una confianza de 40, todo esto con valores mínimos, así como establecer el valor mínimo para poder generar un tamaño mínimo de 5.

IV.4 Análisis de Resultados

Para este caso, encontramos que existen características distintas en cuanto a la aprobación y no aprobación de las tarjetas de crédito. A pesar de que existen datos que tienen en común como lo son el género o la etnia, es claro que se refleja que las aprobaciones tienden a ir más hacia personas que cuentan con empleo, así como las personas las cuales ya son clientes del banco o se encuentran en un estado civil de "Casados".

V. Regresión Lineal

Equipo 4

Data Mining

3CV15

V. Regresión Lineal

V.1 Descripción del ejercicio

Tan, Steinbach y Kumar (2013), definen a la regresión lineal como “una técnica de modelado predictivo”. Además, Carollo (2012) establece que “El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente y un conjunto de variables independientes”.

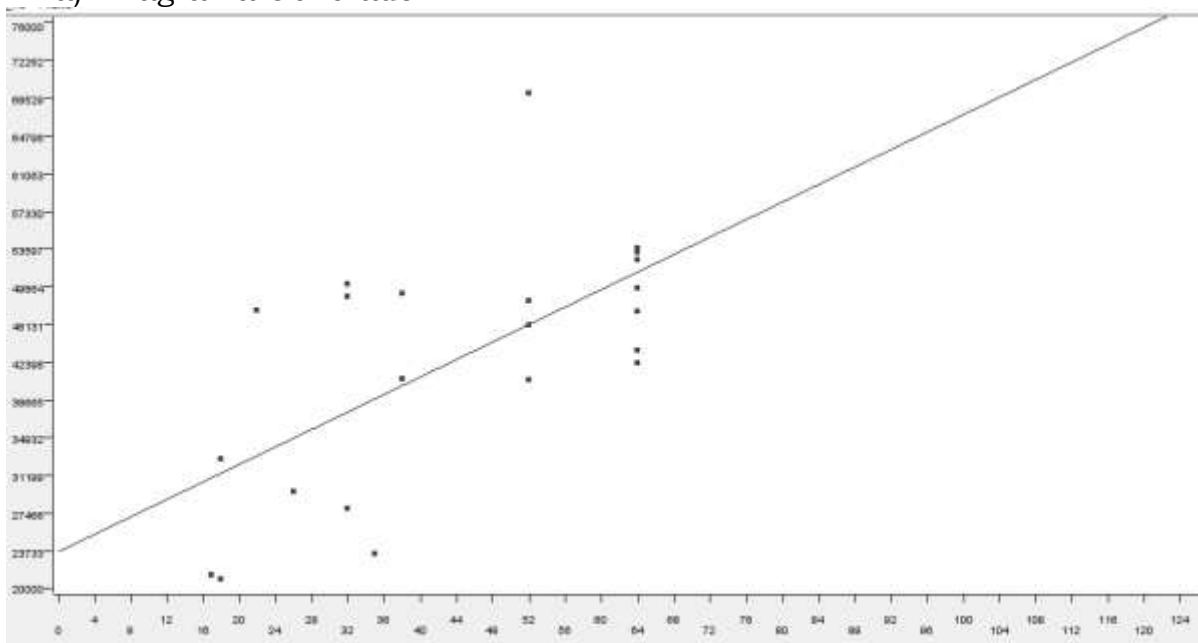
Con lo anterior, crearemos un modelo para tratar de predecir el promedio de asistencia a las copas del mundo según los datos brindados en el dataset. Teniendo en cuenta los partidos que se juegan y la sede en la que se encuentra cada evento.

V.2 Diccionario de Datos.

Nombre	Significado	Tipo	Dominio
Avarage_Attendace	Promedio de personas con asistencia en vivo	Numérico	Numero entero Positivo
Matches	Número total de partidos Jugados	Numerico	Numero entero Positivo

V.3 Resultados

a) Diagrama Generado



b) Medidas obtenidas

Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
Matches	432.3442	119.6202	3.6143	0.0017
Intercept	23,704.6924	5,632.0908	4.2089	0.0004
R-Squared: 0.3951				
Adjusted R-Squared: 0.3649				

c) Descripción de las características de los resultados generados

En el punto anterior podemos observar algunos datos, como por ejemplo el coeficiente de nuestra ecuación, es positivo por lo que nuestra recta tiene una relación lineal positiva. También tenemos otro dato como por ejemplo el punto donde intercepta en nuestra coordenada Y, con un valor de 23,704.6924 y coeficiente de determinación igual a 0.3951, entre otros datos.

d) Tipo de muestra que se utilizó para prueba y entrenamiento

En este punto no se coloca algún tipo de configuración para particionar los datos para prueba y entrenamiento, sino que como el mismo tema lo dice, se trata de asignar una variable dependiente a una independiente, por la cual esta generará una ecuación que explique el comportamiento lineal de los datos brindados.

V.4 Análisis de Resultados

Con los resultados anteriores observamos que contamos con un valor de R^2 un tanto bajo, con lo que se puede concluir que el porcentaje de certeza de nuestra ecuación $Y = 432.3442X + 23,704.6924$ cuenta con un 39.51% de precisión por lo que se concluye que a pesar de que los datos que predijo nuestro algoritmo no fueron desatinados, si se necesitan más datos para poder aumentar el porcentaje de precisión de este mismo con respecto a los datos reales.

Row ID	S Year	S Hosts	I Total_A...	I Matches	I Averag...	D Predict...
0	1930	Uruguay	590549	18	32808	31,486.887
1	1934	Italy	363000	17	21353	31,054.543
2	1938	France	375700	18	20872	31,486.887
3	1950	Brazil	1045246	22	47511	33,216.264
4	1954	Switzerland	768607	26	29562	34,945.64
5	1958	Sweden	819810	35	23423	38,836.738
6	1962	Chile	893172	32	27912	37,539.705
7	1966	England	1563135	32	48848	37,539.705
8	1970	Mexico	1603975	32	50124	37,539.705
9	1974	West Germany	1855753	38	49099	40,133.77
10	1978	Argentina	1545791	38	40679	40,133.77
11	1982	Spain	2109723	52	40572	46,186.589
12	1986	Mexico	2394031	52	46039	46,186.589
13	1990	Italy	2516215	52	48389	46,186.589
14	1994	United States	3587538	52	68991	46,186.589
15	1998	France	2785100	64	43517	51,374.718
16	2002	South Korea...	2705197	64	42269	51,374.718
17	2006	Germany	3359439	64	52491	51,374.718
18	2010	South Africa	3178856	64	49670	51,374.718
19	2014	Brazil	3429873	64	53592	51,374.718
20	2018	Russia	3031768	64	47371	51,374.718
21	2022	Qatar	3404252	64	53191	51,374.718