**Logotipo

Descripción generada automáticamenteImagen que contiene Logotipo

Descripción generada automáticamente**

**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](#_Toc138184504)

[i. FUENTES DE DATOS 1](#_Toc138184505)

[ii. DESCRIPCIÓN DE DATOS 2](#_Toc138184506)

[iii. TRATAMIENTO DE DATOS. 5](#_Toc138184507)

[I. CLASIFICACIÓN DE ARBOLES: 22](#_Toc138184508)

[I.1 Descripción del ejercicio 22](#_Toc138184509)

[I.2 Diccionarios de Datos. 22](#_Toc138184510)

[I.2.1 Diccionario de Datos CART 22](#_Toc138184511)

[I.2.2 Diccionario de datos C4.5 23](#_Toc138184512)

[I.3 Resultados 30](#_Toc138184513)

[I.4 Análisis de los resultados 33](#_Toc138184514)

[II. Multi Clasificación. Bagging y Boosting 36](#_Toc138184515)

[II.1 Descripción del ejercicio 36](#_Toc138184516)

[II.2 Diccionario de datos 36](#_Toc138184517)

[II.2.1 Diccionario de datos Random Forest. 36](#_Toc138184518)

[II.2.3 Diccionario de datos Gradient Boost. 37](#_Toc138184519)

[II.3. Resultados 38](#_Toc138184520)

[II.4. Análisis de los resultados. 42](#_Toc138184521)

[III. Agrupamiento (Jerárquico y No Jerárquico) 45](#_Toc138184522)

[III.1 Descripción del ejercicio 45](#_Toc138184523)

[III.2. Diccionario de Datos 45](#_Toc138184524)

[III.2.1 Diccionario de Datos Jerárquico 45](#_Toc138184525)

[III.2.2 Diccionario No Jerárquico 46](#_Toc138184526)

[III.3. Resultados. 47](#_Toc138184527)

[III.4 Análisis de Resultados 52](#_Toc138184528)

[IV. Reglas de Asociación 54](#_Toc138184529)

[IV.1 Descripción del ejercicio 54](#_Toc138184530)

[IV.2 Diccionario de Datos 54](#_Toc138184531)

[IV.3 Resultados 55](#_Toc138184532)

[IV.4 Análisis de Resultados 56](#_Toc138184533)

[V. Regresión Lineal 58](#_Toc138184534)

[V.1 Descripción del ejercicio 58](#_Toc138184535)

[V.2 Diccionario de Datos. 58](#_Toc138184536)

[V.3 Resultados 58](#_Toc138184537)

[V.4 Análisis de Resultados 59](#_Toc138184538)

Instituto Politécnico Nacional

Escuela Superior de Cómputo

Unidad de Aprendizaje: Minería de datos

Ciclo escolar: 2023–2

**Proyecto Final**

# CONJUNTOS DE DATOS

## 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)

1. **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

1. **Conjunto de datos 3.** Credit Cards Aprovals

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/samuelcortinhas/credit-card-approval-clean-data>

**Autor**: Samuel Cortinhas

## DESCRIPCIÓN DE DATOS

Conjunto de Afluencia

Este conjunto está formado por un total de 955,305registros; 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. | Categórico | Enero - Diciembre |
| Línea | La línea del sistema de transporte público | Categórico | Línea 1 – Línea 12 Linea A Linea B |
| Estación | El nombre de la estación donde se registró la afluencia | Categórico | 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 151,260registros; 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 | Categórico | 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. | Categórico | 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 | Categórico | Nombre del estadio sede |
| Number | Número de asistencia en el torneo | Numérico | Número no especificado |
| Host | Anfintrion de la Copa del Mundo | Categórico | 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 |
| Avarage\_Attendace | 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 Aprovals

|  |  |  |  |
| --- | --- | --- | --- |
| **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 | 1. Soltero, divorciado, etc. 2. Casado |
| BankCustomer | Cliente del banco | Numérico | 1. No tiene cuenta 2. 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 | 1. Sin valor 2. Con valor |
| Employed | Estado laboral | Numérico | 1. Con empleo 2. Desempleado |
| CreditScore | Puntaje de crédito (esta función se ha escalado) | Numérico | 0 – 99 |
| DriversLicense | Licencia de conducir | Numérico | 1. No tiene licencia 2. 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 | 1. No aprobado 2. Aprobado |

## TRATAMIENTO DE DATOS.

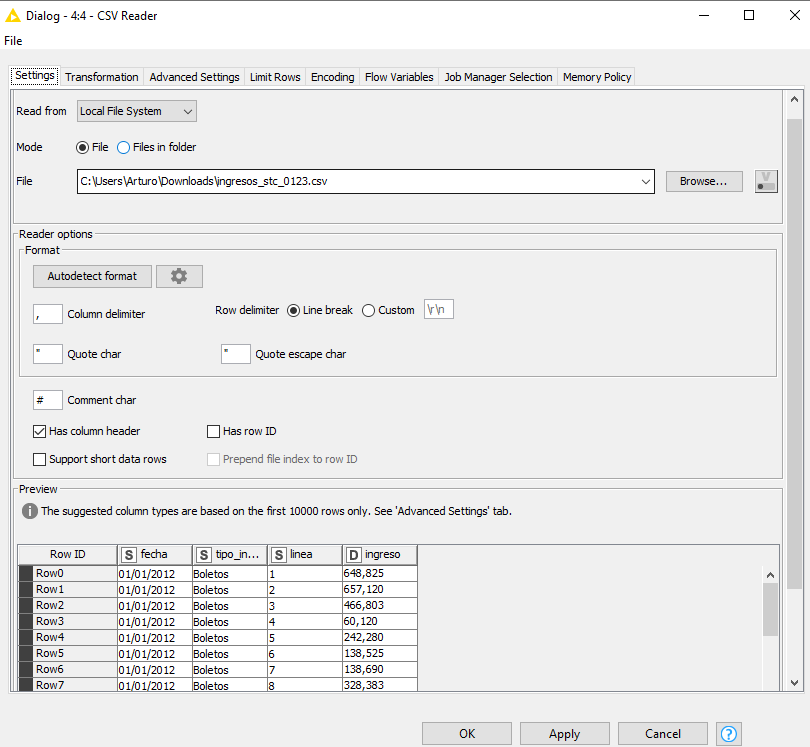
*Tratamiento de datos Ingresos*

A picture containing text, line, font, plot

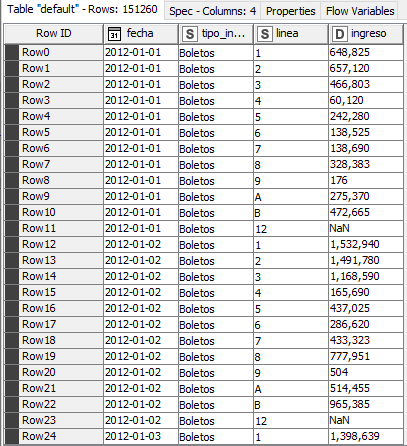
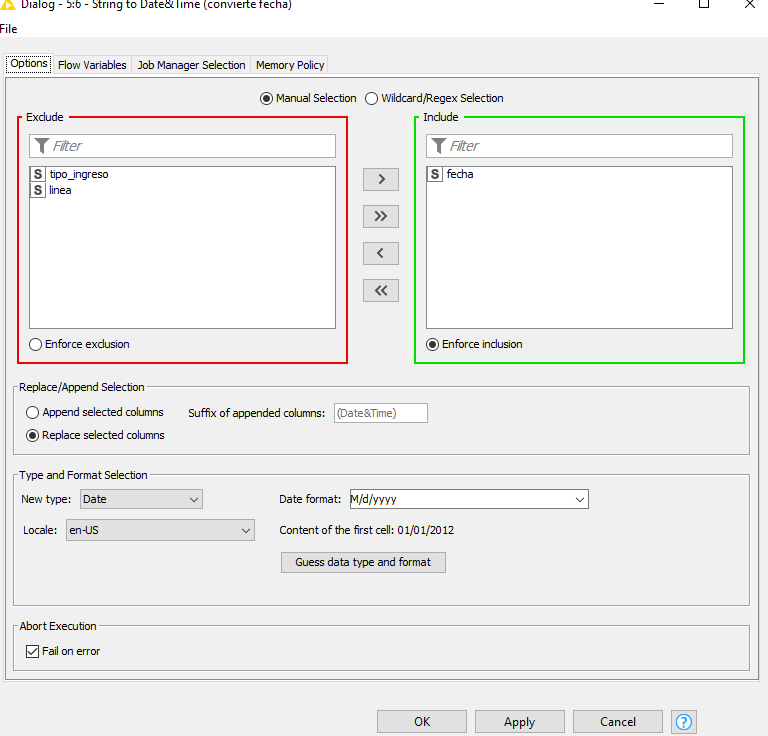
Description automatically generated

*Diagrama de Tratamiento de datos*

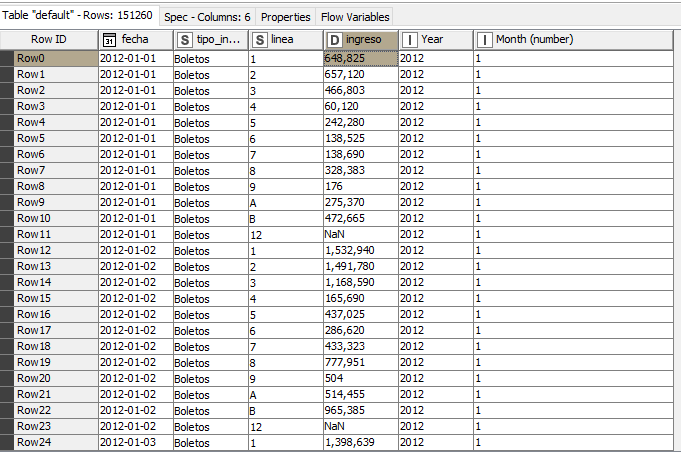
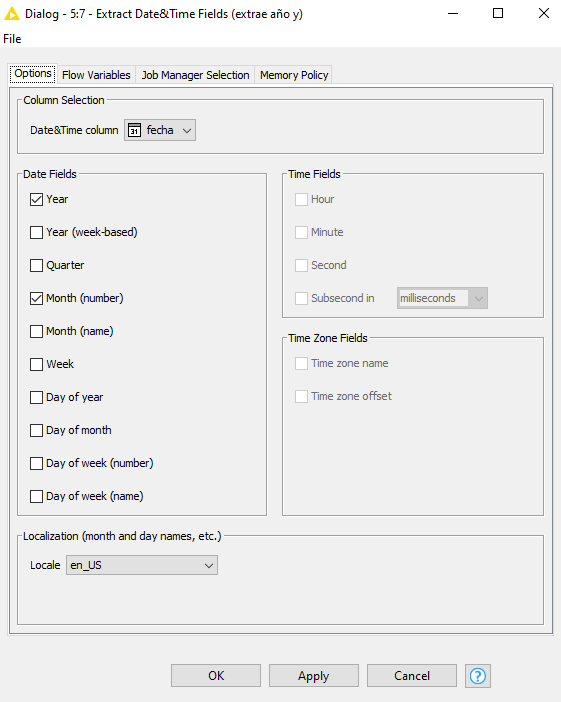
1. Cargaremos nuestro conjunto de datos (ingresos\_stc\_0123).



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

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

1. 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'**)**

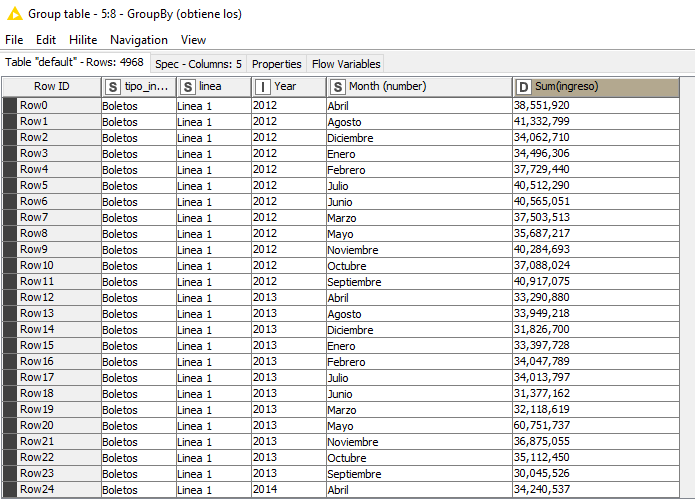
# --------------------- TRATAMIENTO DE DATOS ---------------------

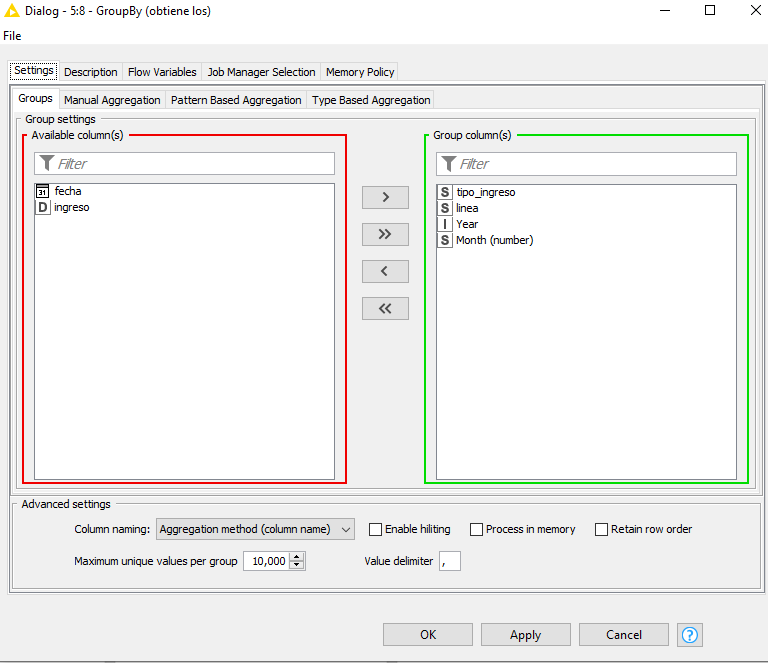
# Tabla de Salida

knio**.**output\_tables**[**0**]** **=** knio**.**Table**.**from\_pandas**(**ingreso**)**

1. Group by:

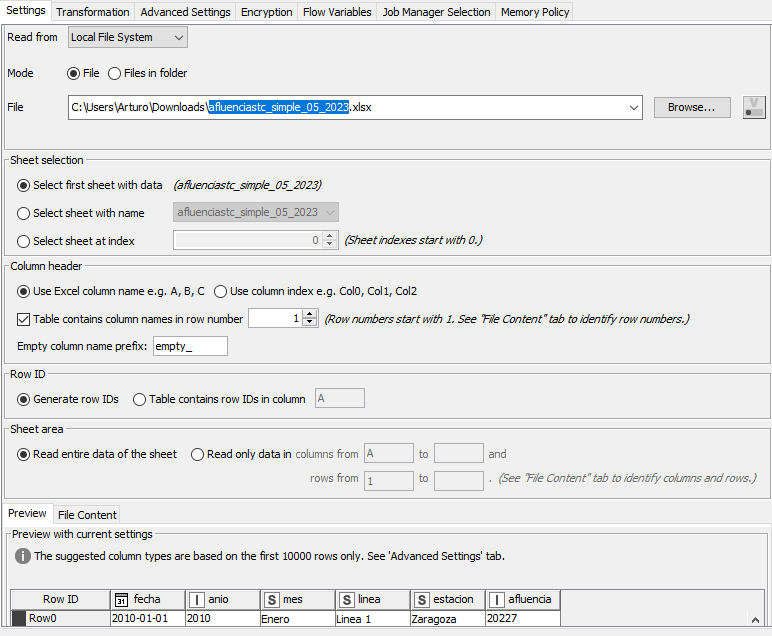
Este nodo nos permitira saber la suma los ingresos totales registrados por cada mes del año.





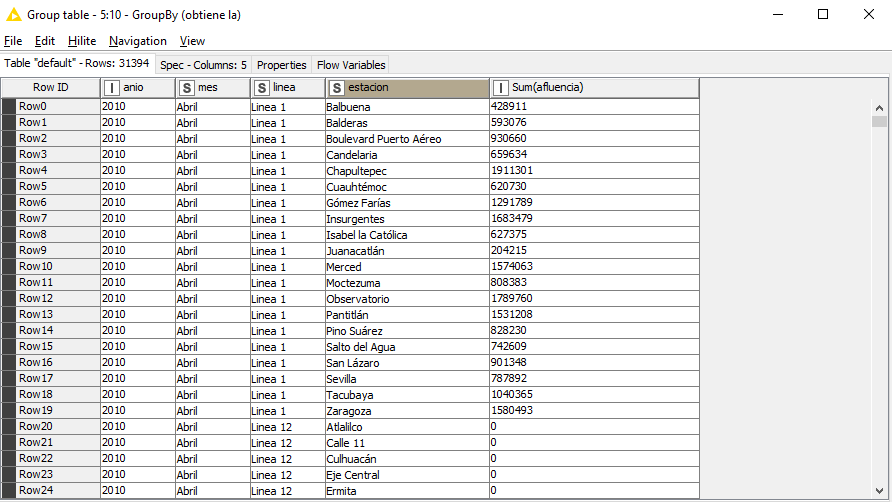
*Tratamiento de datos Afluencia de datos*

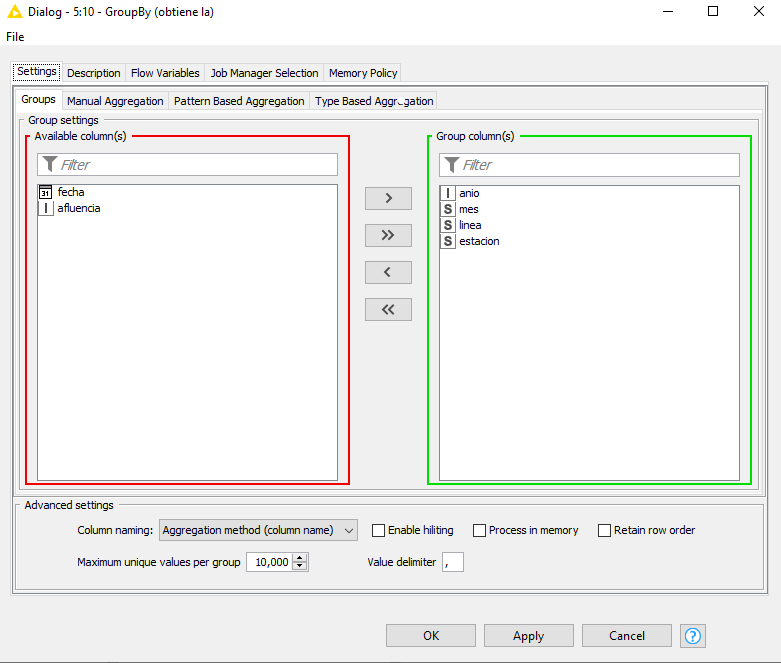
1. *Carga del conjunto de datos afluenciastc\_simple\_05\_2023*



1. Group by

Obtiene la Afluencia total que se registra por cada mes.

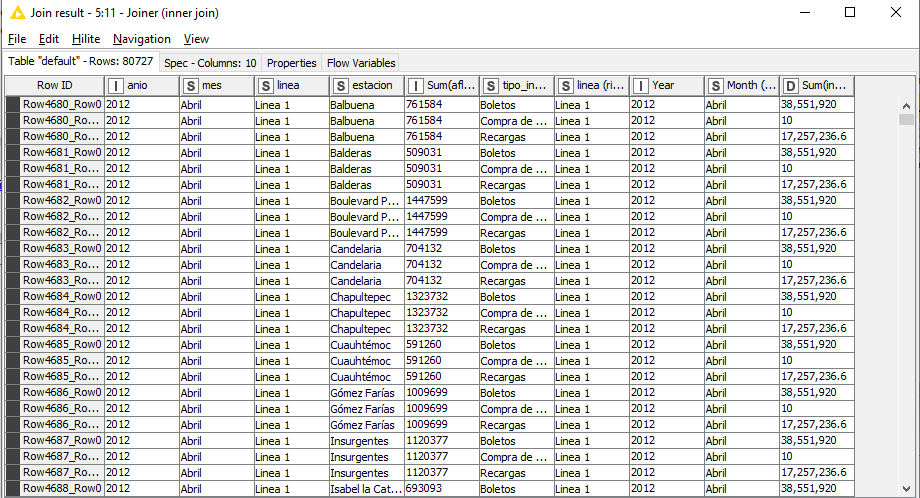


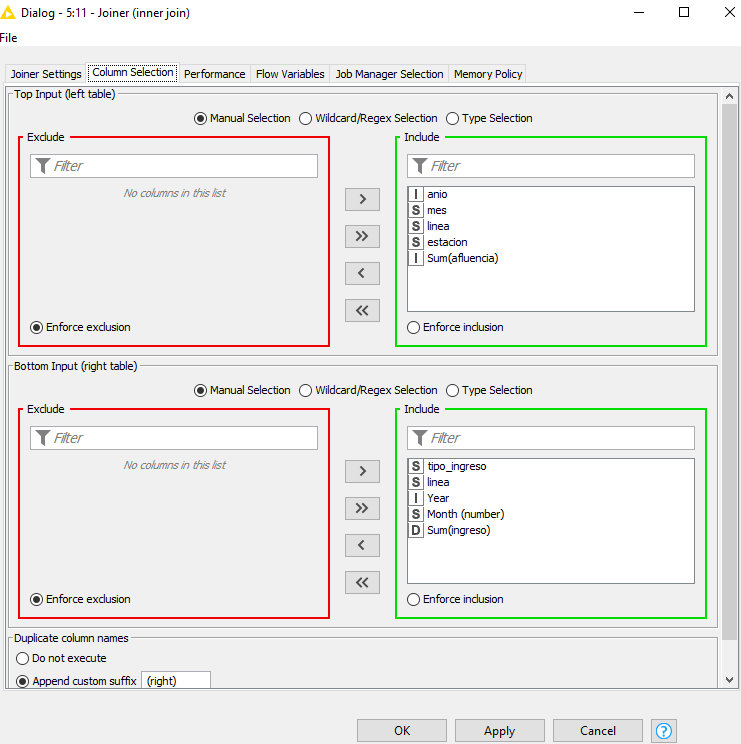


*Tratamiento de datos Ingresos y Afluencia de datos*

1. Inner join

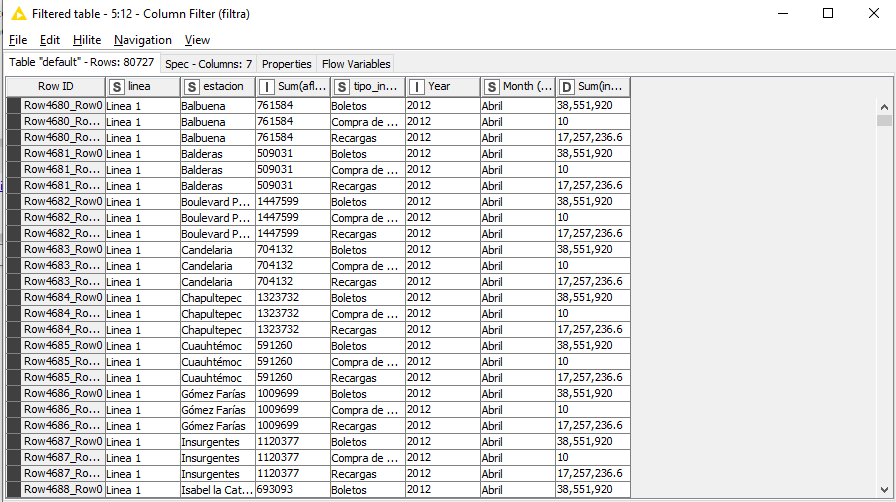
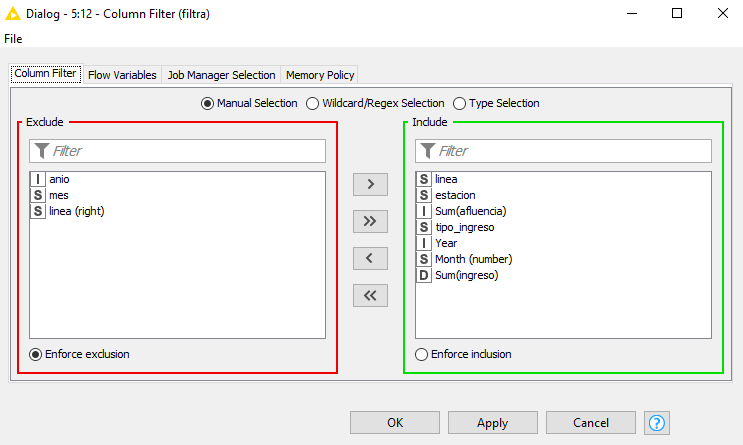
Se usa para juntar ambos conjuntos de datos con sus respectivos tratamientos hechos.





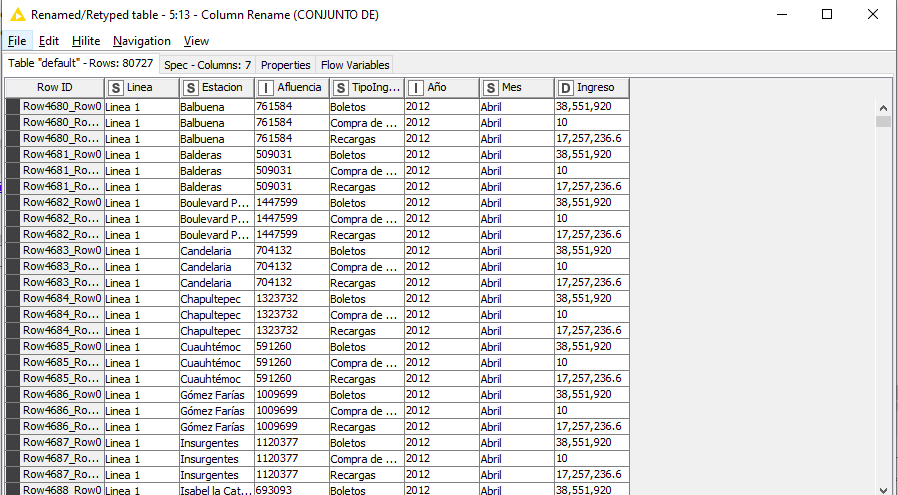
1. Colum filter

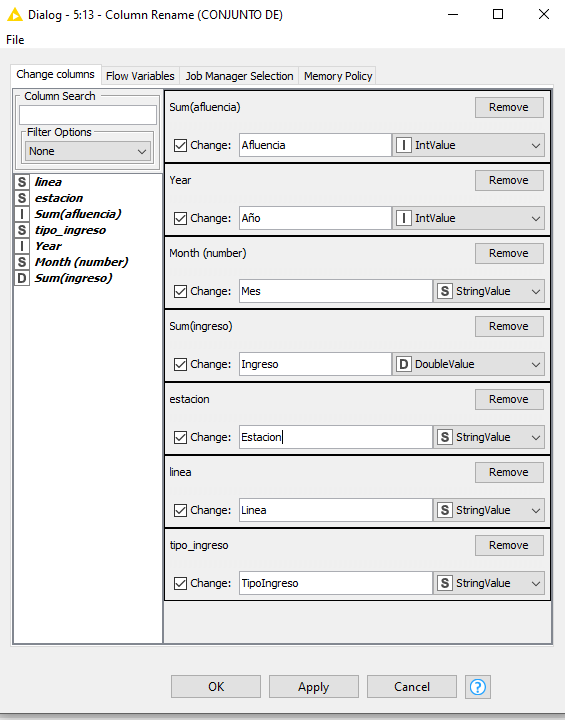
Se usa para eliminar datos que se encuentren repetidos.

1. Colum Rename

Se usa para renombrar el nombre de los atributos.





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

Descripción generada automáticamente

*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.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

1. Aplicación de Column Filter

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

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Tenemos una tabla ya tratada de la siguiente manera:

Tabla

Descripción generada automáticamente

*Tabla con el Tratamiento de Datos realizado*

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Significado** | **Tipo** | **Dominio** |
| Host | Anfintrion de la Copa del Mundo | Categórico | 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 |
| Avarage\_Attendace | 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 Aproval*

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**)**

Imagen que contiene ventana, edificio, grande, tabla

Descripción generada automáticamente

*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 |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Equipo 4**  Data Mining  3CV15 |

I. Clasificación. *Árboles*

# 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 predicir 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

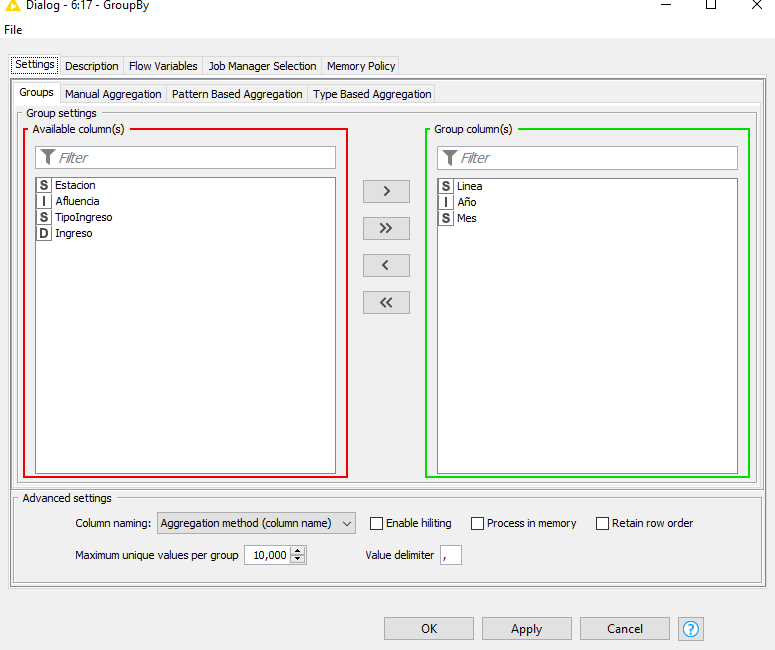
|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Significado** | **Tipo** | **Dominio** |
| Línea | Número de línea del sistema del transporte público. | Categórico | Línea 1 – Línea 12 Línea A Línea B |
| Afluencia | Número de personas que utilizaron esa estación en particular en el día especificado.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

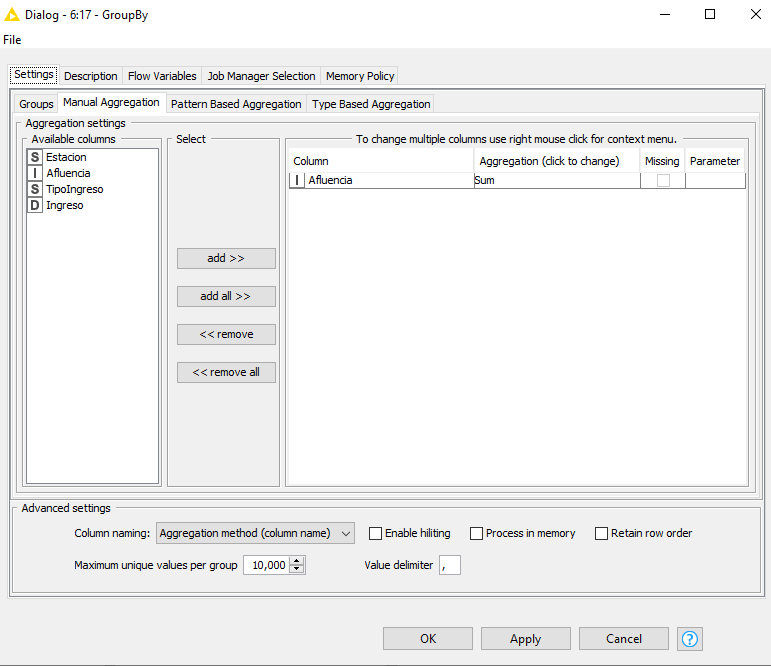
*CART*

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.

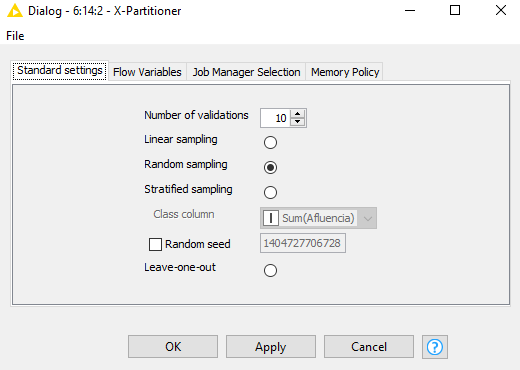
Tabla

Descripción generada automáticamente 



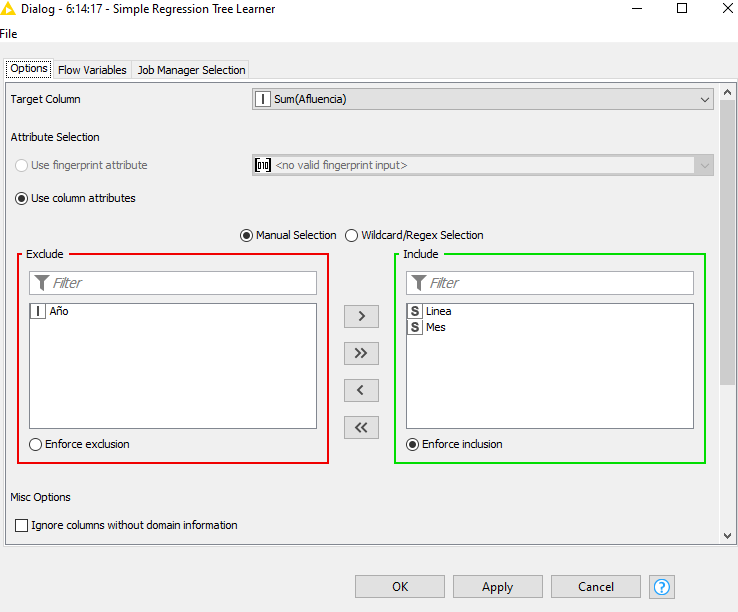
1. X-Partitioner

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.



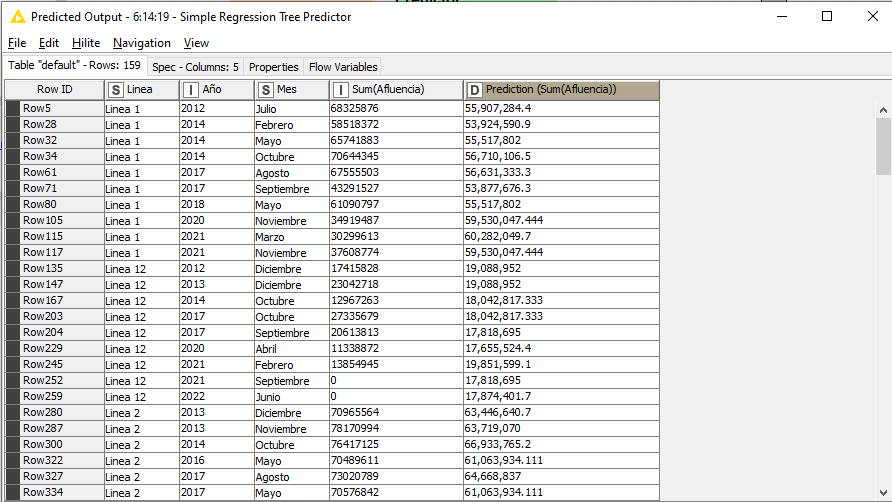
1. Simple Regression Tree Learner

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



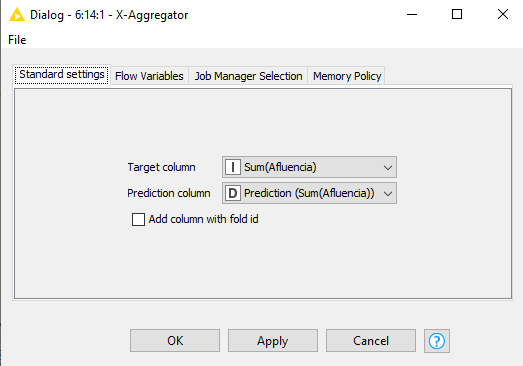
1. Simple Regression Tree Predictor

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.



1. 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.



1. 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.

A screenshot of a computer

Description automatically generated with medium confidence

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

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidencePara poder empezar con nuestro árbol de decisión, eliminamos las columnas de datos que resulten innecesarios.

1. 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.

A screenshot of a computer

Description automatically generated

1. Decision Tree Learner

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

A screenshot of a computer

Description automatically generated

1. 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.

A screenshot of a computer

Description automatically generated

1. 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.

A screenshot of a computer

Description automatically generated

1. 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%.

A screenshot of a computer

Description automatically generated with medium confidence

## I.3 Resultados

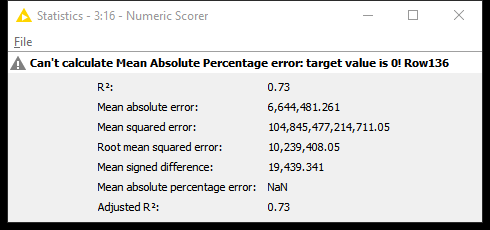
1. Diagrama generado

*CART*

Interfaz de usuario gráfica, Diagrama, Aplicación

Descripción generada automáticamente

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.



*C4.5*

Diagrama

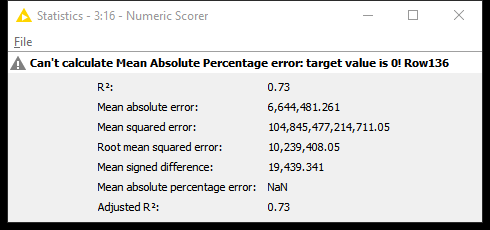
Descripción generada automáticamente Tabla

Descripción generada automáticamente

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.

1. Medidas Obtenidas

*CART*



*C4.5*

Texto

Descripción generada automáticamente con confianza baja

1. 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.

1. 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*

A screenshot of a computer error

Description automatically generated with low confidence

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.

A picture containing screenshot, text, plot, line

Description automatically generated

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).

Calendario

Descripción generada automáticamente

Tabla

Descripción generada automáticamente

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Equipo 4**  Data Mining  3CV15 |

II. Multi Clasificación. *Bagging y Boosting*

# 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.

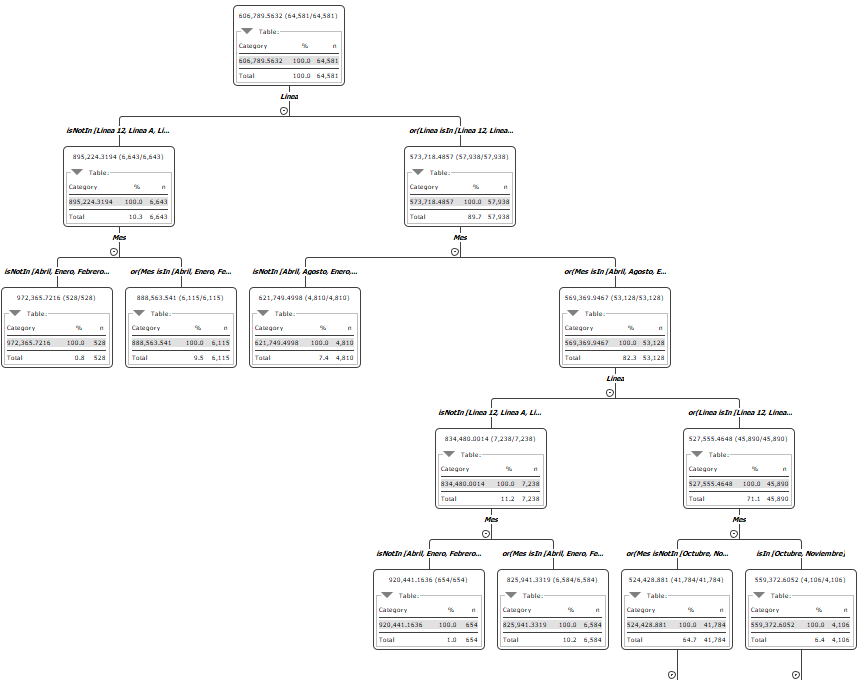
|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Significado** | **Tipo** | **Dominio** |
| Genders | Genero de la persona | Categórico | Masculino  Femenino |
| Married | Estado civil | Categórico | Soltero/ divorciado  Casado |
| BankCustomer | Cliente del banco | Categórico | No tiene cuenta  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.

1. Diagrama generado

*Random Forest*

Aquí tenemos uno de los 100 modelos que se generan.



Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

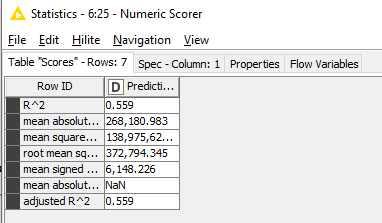
*Gradient Boost*

Tabla

Descripción generada automáticamente

1. Medidas obtenidas

*Random Forest*



*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

1. Descripción de las características

*Random Forest*

Al tener un valor de r2 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.

Tabla

Descripción generada automáticamente

1. Tipo de muestra que se utilizó para prueba y entrenamiento

*Random Forest*

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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

*Gradient Boost*

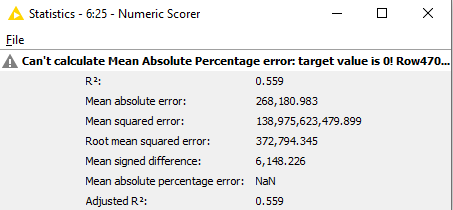
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

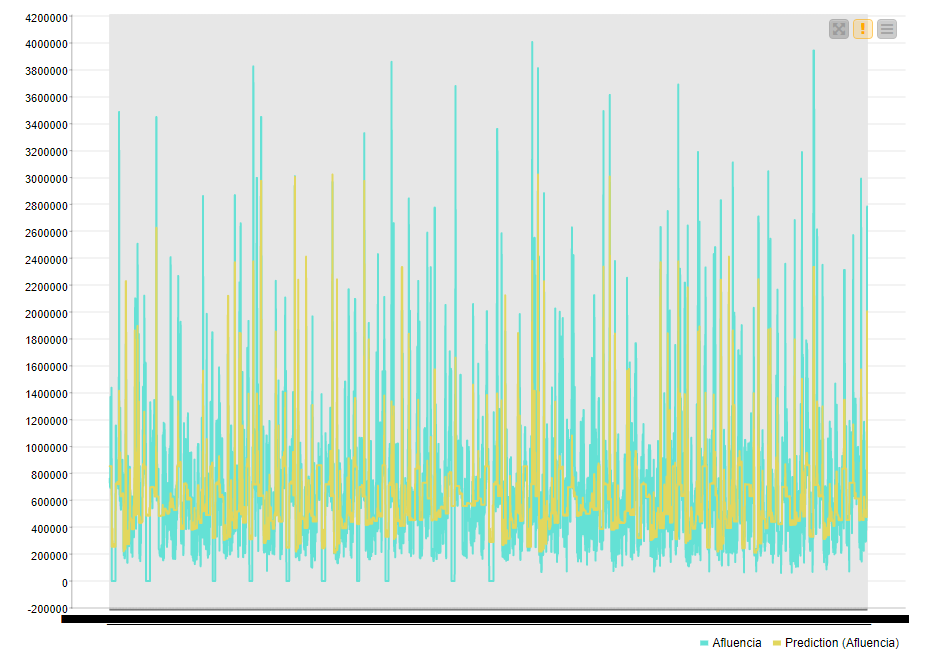
Se asignó un porcentaje de 80% para entrenamiento y un 20% para prueba

## II.4. Análisis de los resultados.

*Random Forest*



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*

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

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.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Equipo 4**  **Data Mining**  **3CV15** |

III. Agrupamiento *(Jerárquico y No Jerárquico)*

# 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 | Anfintrion de la Copa del Mundo | Categórico | 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 |
| Avarage\_Attendace | 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

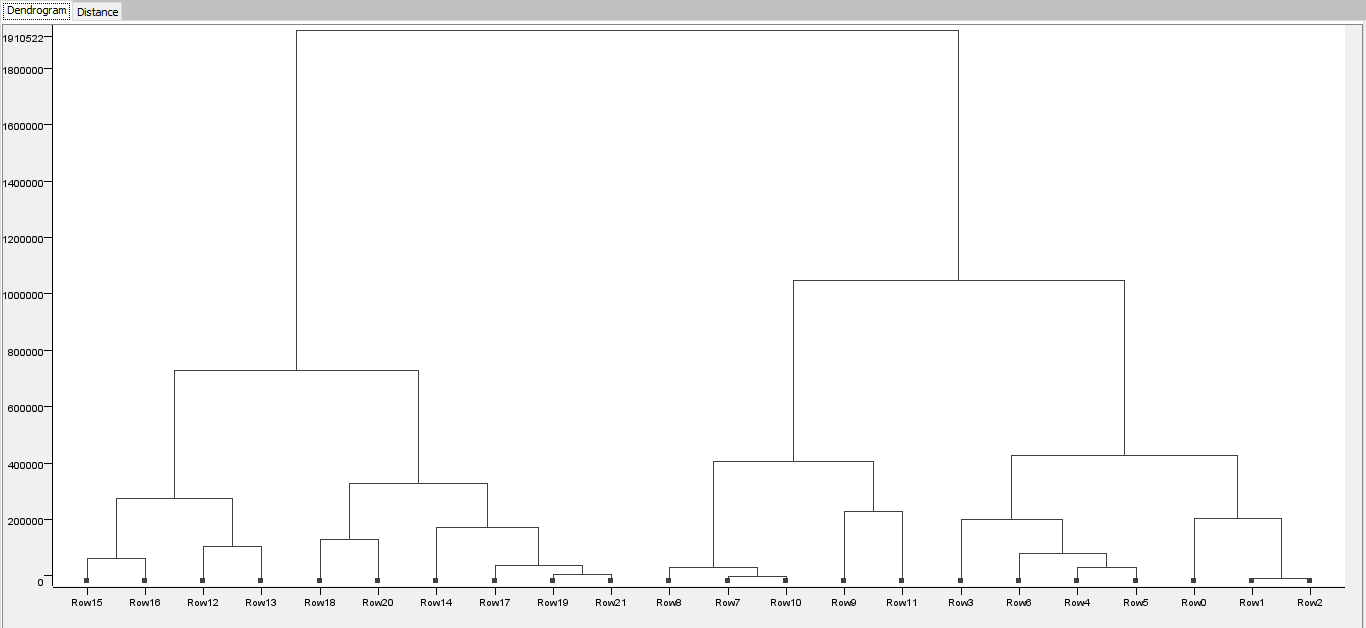
Conjunto de datos Ingresos y Afluencia.

|  |  |  |  |
| --- | --- | --- | --- |
| **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.

1. Diagrama generado

*Jerárquico*

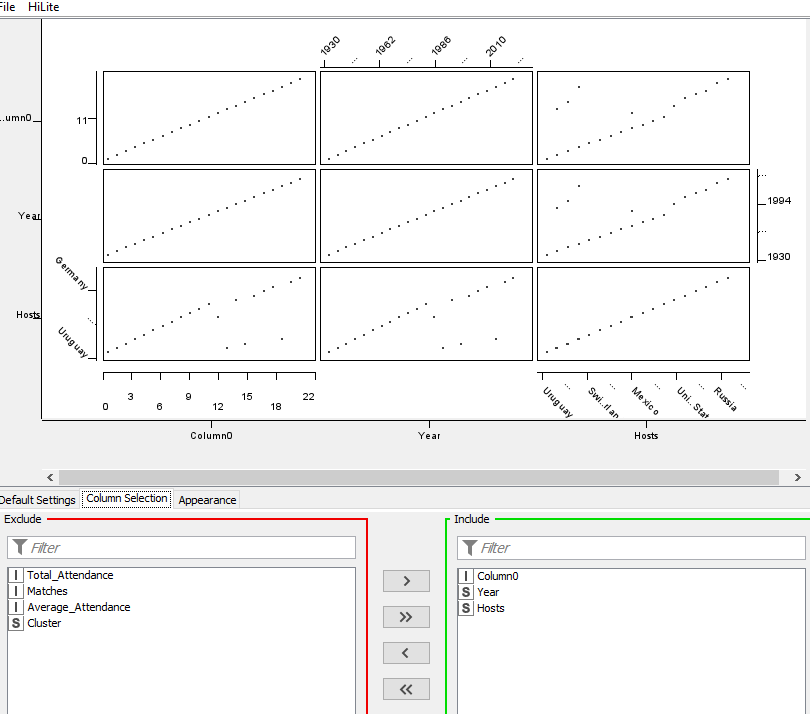


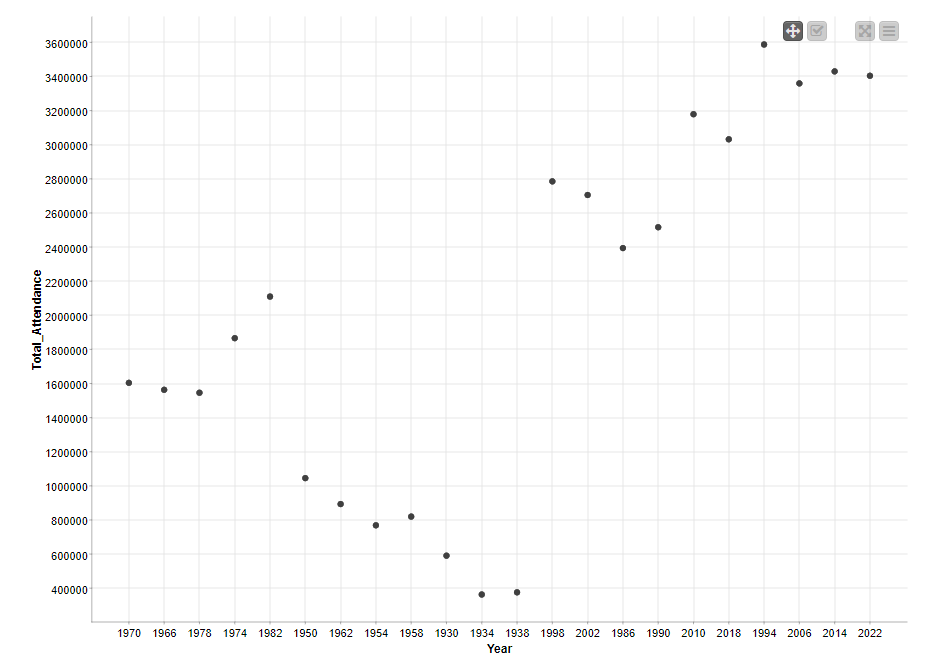
*No Jerárquico*



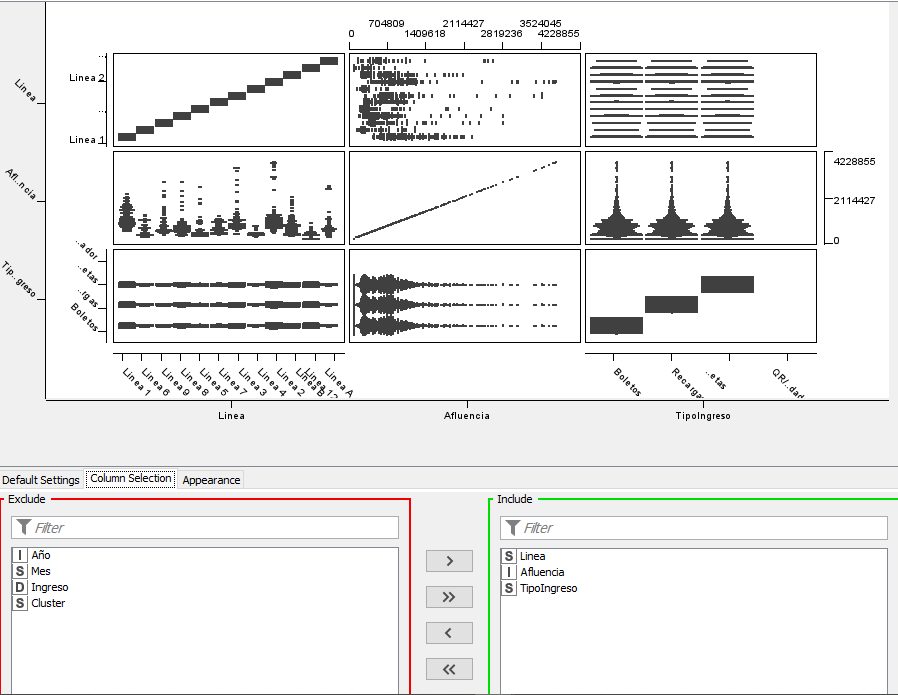
1. Medidas obtenidas.Silueta.

*Silueta modelo Jerárquico*





*Silueta modelo No Jerarquico*



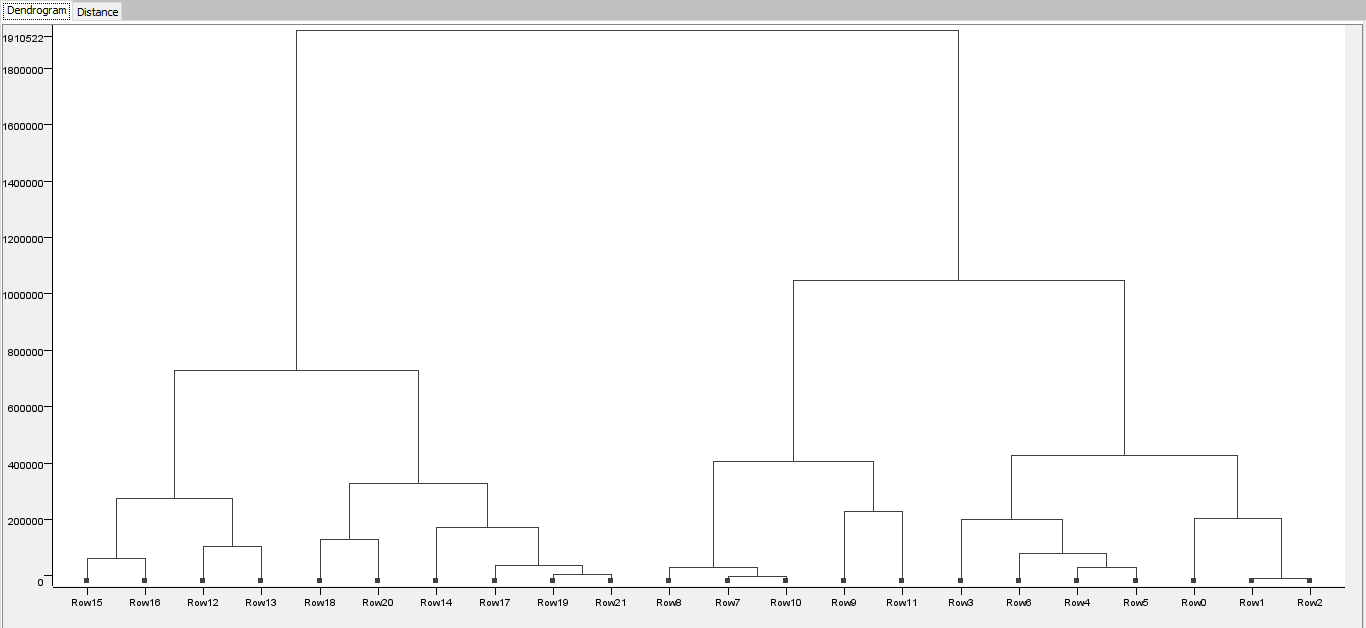


1. Descripción de las características de los resultados generados

*Dendograma (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.

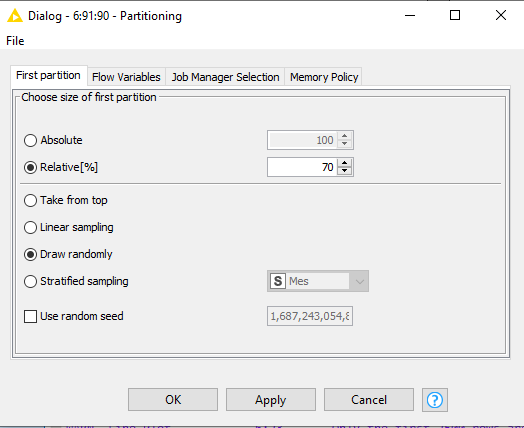


1. Tipo de muestra que utilizó para prueba y entrenamiento.

*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.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Equipo 4**  Data Mining  3CV15 |

IV. Reglas de Asociación

# 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

1. Diagrama generado

Texto

Descripción generada automáticamente

1. Medidas obtenidas

Imagen que contiene ventana, grande, cuarto

Descripción generada automáticamente

1. Descripción de las características de los resultados generados

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Tarjetas aprobadas / Tarjetas no aprobadas

1. Tipo de muestra que se utilizó para prueba y entrenamiento

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

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 hacía 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”.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  | **Equipo 4**  Data Mining  3CV15 |

V. Regresión Lineal

# 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”.

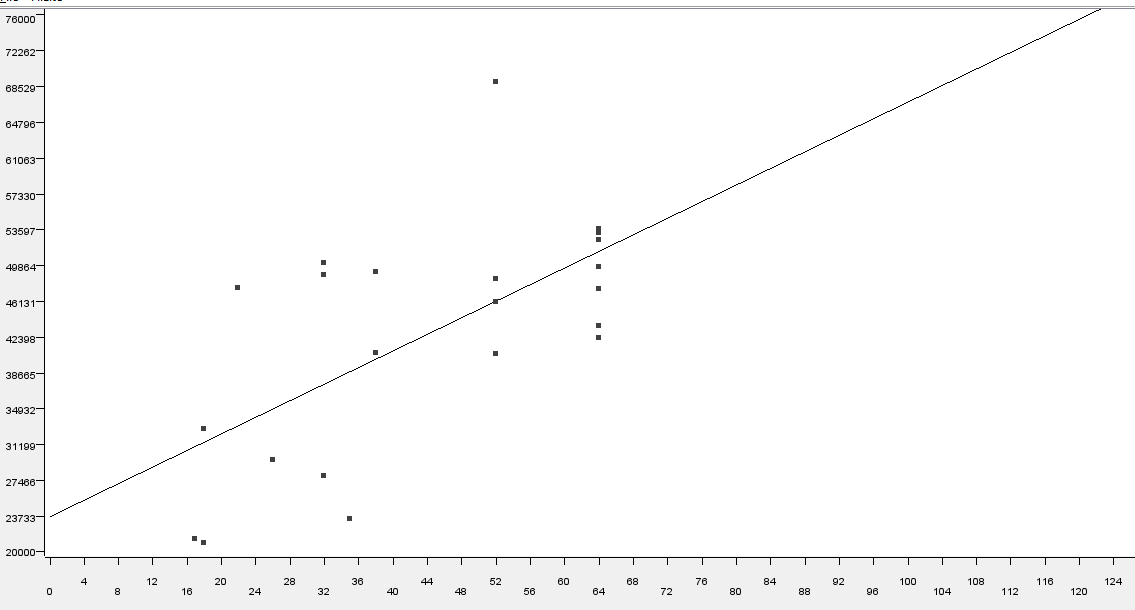
Con lo anterior, crearemos un modelo para tratar de predecir el promedio de asistencia a las copas del mundo según los datos brindados en el dataset. Teniendo en cuenta los partidos que se juegan y la sede en la que se encuentra cada evento.

## V.2 Diccionario de Datos.

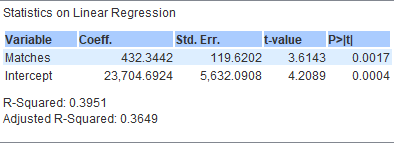
|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre** | **Significado** | **Tipo** | **Dominio** |
| Avarage\_Attendace | Promedio de personas con asistencia en vivo | Numérico | Numero entero Positivo |
| Matches | Número total de partidos Jugados | Numerico | Numero entero Positivo |

## V.3 Resultados

1. Diagrama Generado



1. Medidas obtenidas



1. 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.

1. 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 R2 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.

