INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

MINERÍA DE DATOS

Materia de Minería de datos. Periodo escolar: 2021-2

Nombre del alumno:
Domínguez Reyes Jesús Alejandro
Conde Francisco José Angel
Nambo Velázquez Carlos
Velasco Martínez Alan Alexis

Índice

1. Introducción

1.1 Descripción del conjunto de datos

Nombre	Heart 2020
Objetivo	Características de bienestar físico y mental de personas mayores de edad.
Créditos	Centros para el Control y Prevención de Enfermedades
URL	https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Originalmente, el conjunto de datos proviene de los CDC y es una parte importante del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS), que realiza encuestas telefónicas anuales para recopilar datos sobre el estado de salud de los residentes de los Estados Unidos. Como describen los CDC: "Establecido en 1984 con 15 estados, BRFSS ahora recopila datos en los 50 estados, así como en el Distrito de Columbia y tres territorios de los Estados Unidos. BRFSS completa más de 400,000 entrevistas a adultos cada año, lo que lo convierte en el sistema de encuestas de salud realizado continuamente más grande del mundo". El conjunto de datos más reciente (al 15 de febrero de 2022) incluye datos de 2020.

1.2 Diccionario de datos

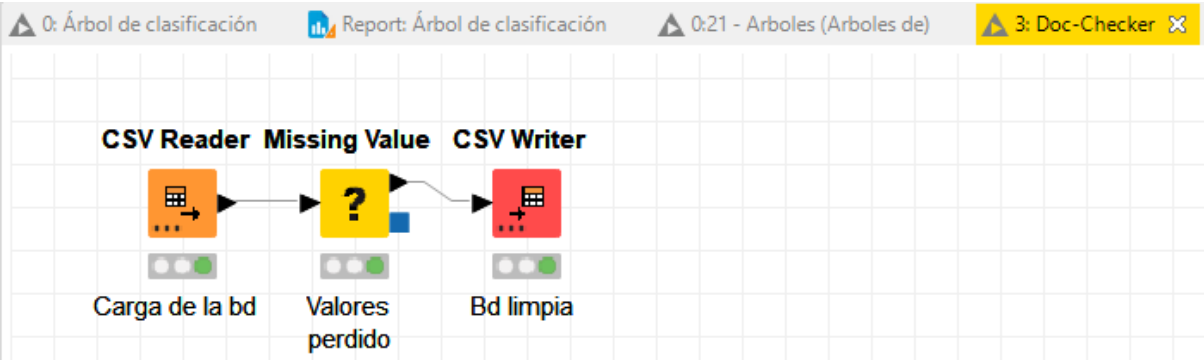
Nombre	Significado	Tipo	Dominio
HeartDisease	Encuestados que alguna vez han reportado tener enfermedad coronaria (CHD) o infarto de miocardio (IM)	Binario	Yes-Si No-No
BMI	Índice de masa corporal (IMC)	Numérico	12-94.8
AlcoholDrinking	Bebedores empedernidos (hombres adultos que beben más de 14 bebidas por semana y mujeres adultas que beben más	Binario	Yes-Si No-No
Smoke	¿Ha fumado al menos 100 cigarrillos en toda su vida? [Nota: 5 paquetes = 100 cigarrillos]	Binario	Yes-Si No-No

Stroke	(Alguna vez contado) (tuviste) un derrame cerebral?	Binario	Yes-Si No-No
PhysicalHealth	Pensando en su salud física, que incluye enfermedades físicas y lesiones, ¿durante cuántos días durante los últimos 30 días su salud física no fue buena?	Numérico	0-30 días
MentalHealth	Pensando en su salud mental, ¿durante cuántos días durante los últimos 30 días su salud mental no fue buena?	Numérico	0-30 días
DiffWalking	¿Tiene serias dificultades para caminar o subir escaleras?	Binario	Yes-Si No-No
Sex	¿Eres hombre o mujer?	Nominal	0 - ninguno, 1 - educación primaria (4º grado), 2 - 5º a 9º grado, 3 - educación secundaria o 4 - educación superior
AgeCategory	Categoría de edad de trece niveles	Nominal	18-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59 60-64 65-69 70-74 75-79 80 or older
Race	Valor de raza/etnia imputado	Nominal	White Hispanic American Indian/Alaskan Native
Diabetic	¿(Alguna vez contado) (tuviste) diabetes?	Nominal	Yes No No, borderline diabetes
PhysicalActivity	Adultos que informaron haber realizado actividad física o ejercicio durante los últimos 30	Binario	Yes-Si No-No

	días que no fueran su trabajo regular		
GenHealth	¿Dirías que en general tu salud es...	Nominal	Excellent Very good Good Fair Poor
SleepTime	En promedio, ¿cuántas horas de sueño obtienes en un período de 24 horas?	Numérico	1 - 24 horas
Asthma	¿(Alguna vez contado) (tuviste) asma?	Binario	Yes-Si No-No
KidneyDisease	Sin incluir cálculos renales, infección de la vejiga o incontinencia, ¿alguna vez le dijeron que tenía enfermedad renal?	Binario	Yes-Si No-No
SkinCancer	¿(Alguna vez contado) (tuviste) cáncer de piel?	Binario	Yes-Si No-No
respiratory disease	Encuestados que alguna vez han reportado tener enfermedad respiratoria	Binario	Yes-Si No-No
infectious disease	Encuestados que alguna vez han reportado tener enfermedad infecciosa	Binario	Yes-Si No-No

1.3 Tratamiento de datos

El conjunto de datos de los alumnos tenía datos perdidos en dos de sus atributos, específicamente, traveltime y studytime, se arregló ese problema con el nodo missing value, tomando como parámetro la media redondeada ya que el dato es entero, se cambian los datos que knime toma como numéricos y se pasan a ser nominales, en el primer meta nodo es la educación y trabajo de ambos padres y el segundo la clasificación de si algo era muy malo hasta muy bueno.



1.4 Técnicas de minería de datos

Técnica	Objetivo	Atributos
Árbol de clasificación	El objetivo del árbol de clasificación será el de encontrar si el paciente es probable que tenga o no una enfermedad cardiovascular, respiratoria, infecciosa **Atributos objetivo:** **HeartDisease** **respiratory disease** **infectious disease**	HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, Race, Diabetic, PhysicalActivity, GenHealth, Asthma, KidneyDisease, SkinCancer, cardiovascular disease, respiratory disease, infectious disease

2. Árbol de clasificación

2.1 Marco teórico

El árbol de clasificación ID3 ayudará a encontrar un patrón de características que nos diga si el paciente tiene o podría desarrollar alguna enfermedad.

2.2 Descripción del trabajo

2.2.1 Intención de la técnica empleada

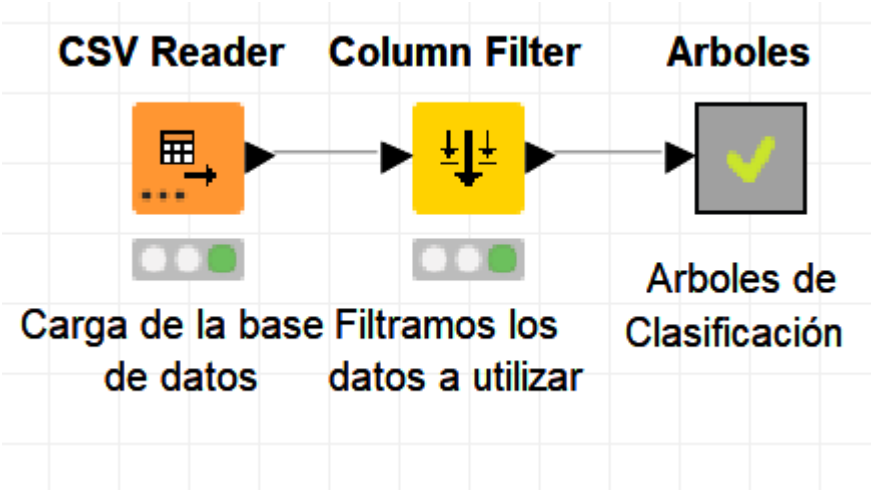
Encontrar si el paciente padece una enfermedad, ya sea una enfermedad infecciosa, respiratoria o cardiovascular, al usar el árbol de clasificación con las características del paciente, como por ejemplo si es que son fumadores, hacen ejercicio, si toman alcohol, etc.

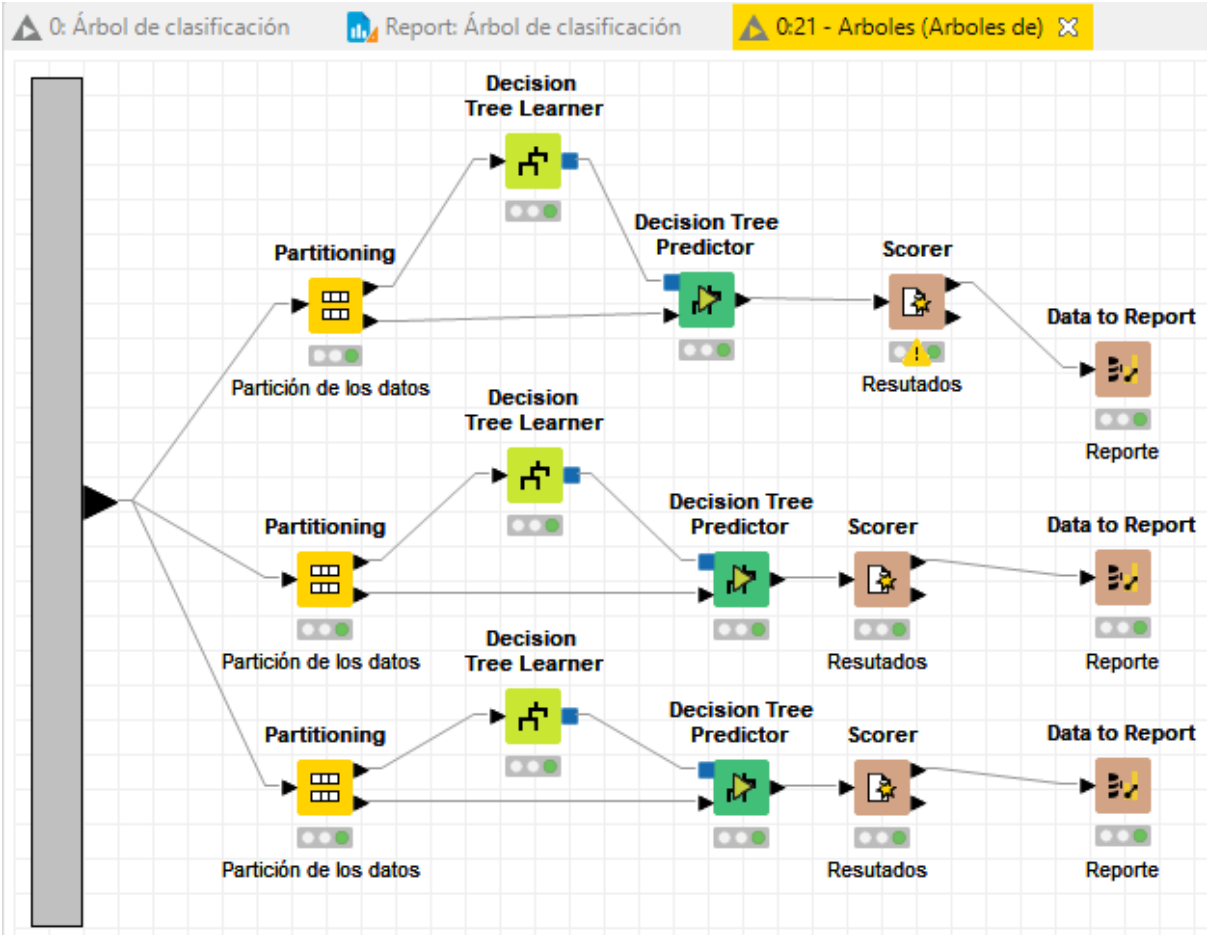
2.2.2 Atributos utilizados

- HeartDisease
- Smoking

- AlcoholDrinking
- Stroke
- DiffWalking
- Sex
- Race
- Diabetic
- PhysicalActivity
- GenHealth
- Asthma
- KidneyDisease
- SkinCancer
- cardiovascular disease
- respiratory disease
- infectious disease

2.2.3 Diagrama General





2.2.4 Evaluación de resultados

Confusion M...

File Hilite

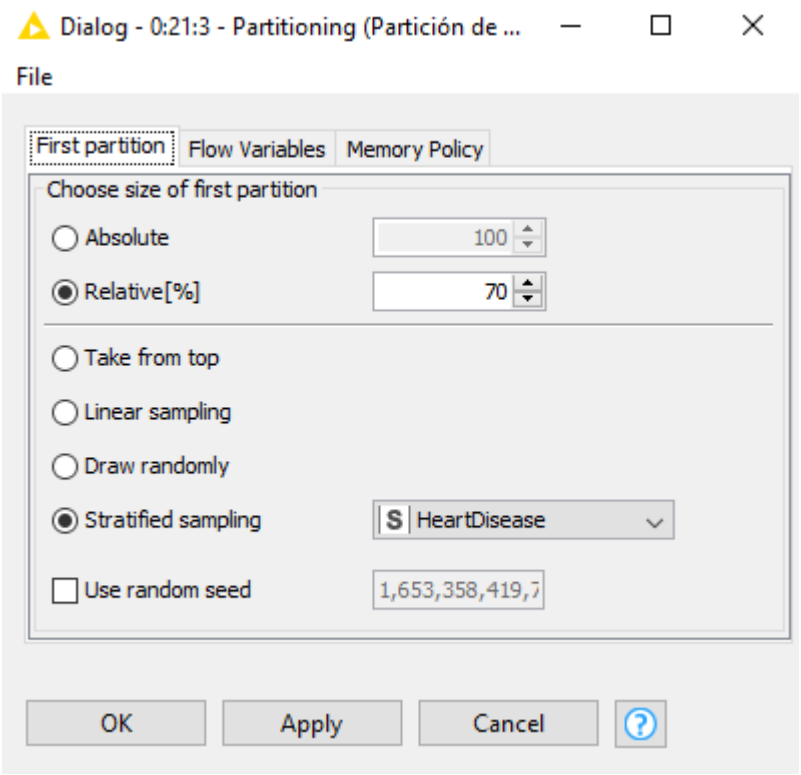
Pstatus \ Prediction (Pstatus)	A	T
A	89	7
T	14	669

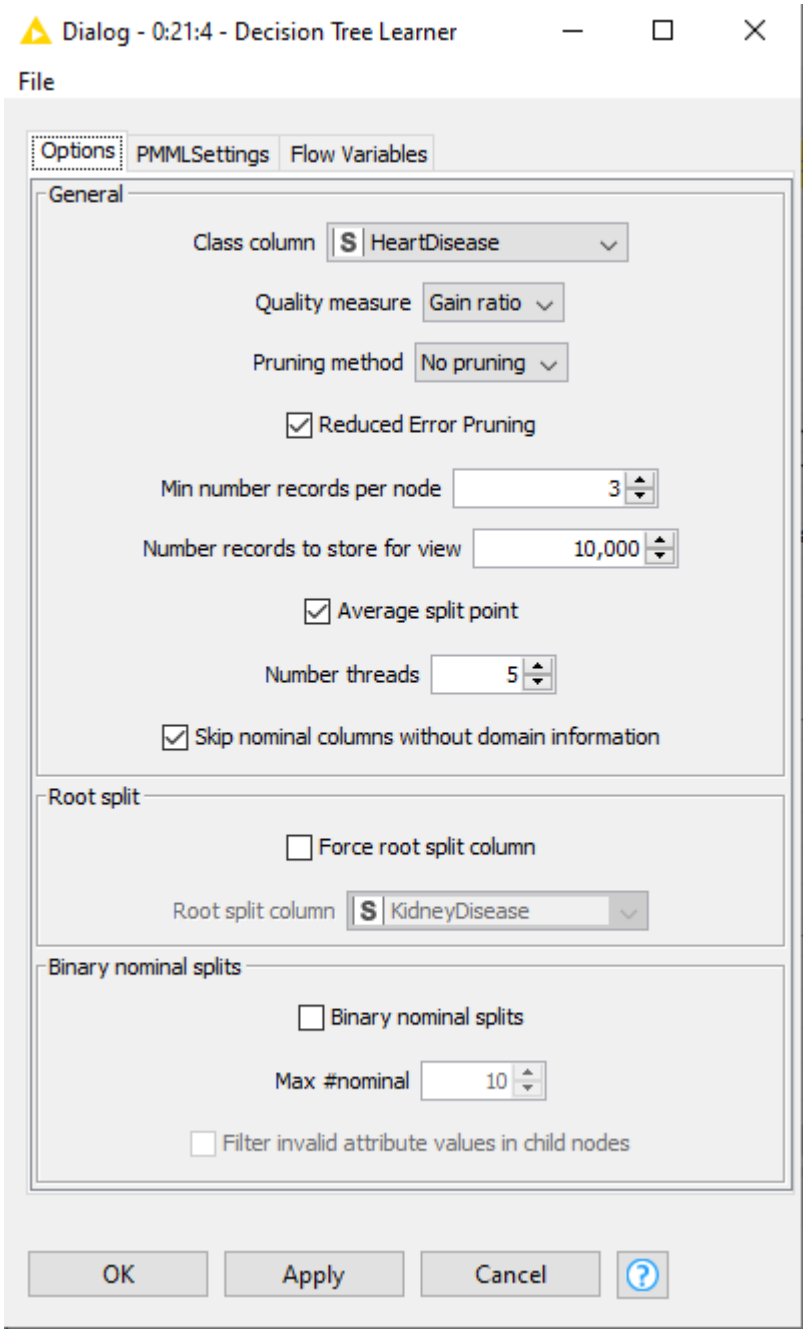
Correct classified: 758 Wrong classified: 21

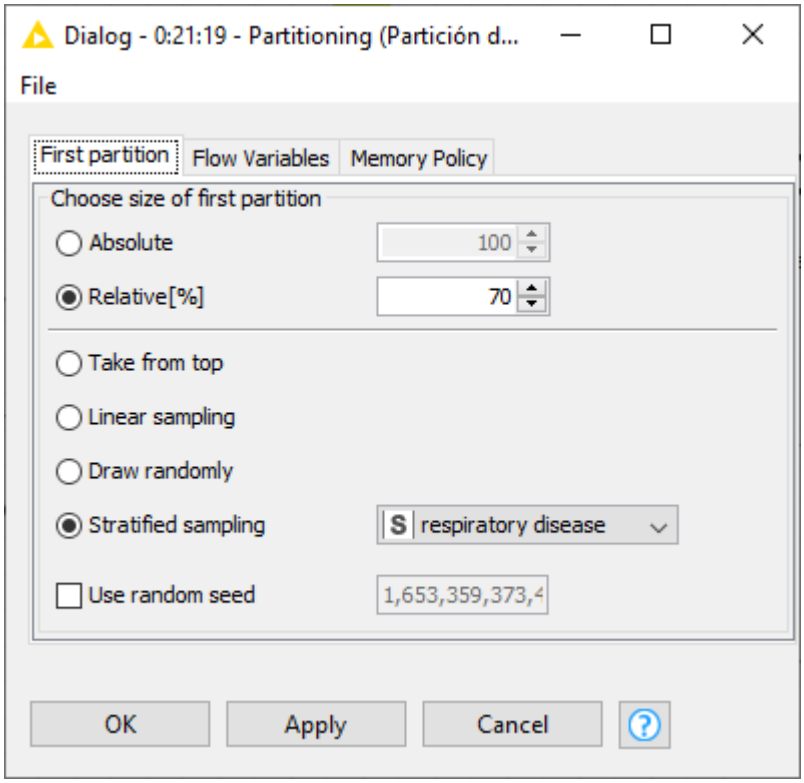
Accuracy: 97.304 % Error: 2.696 %

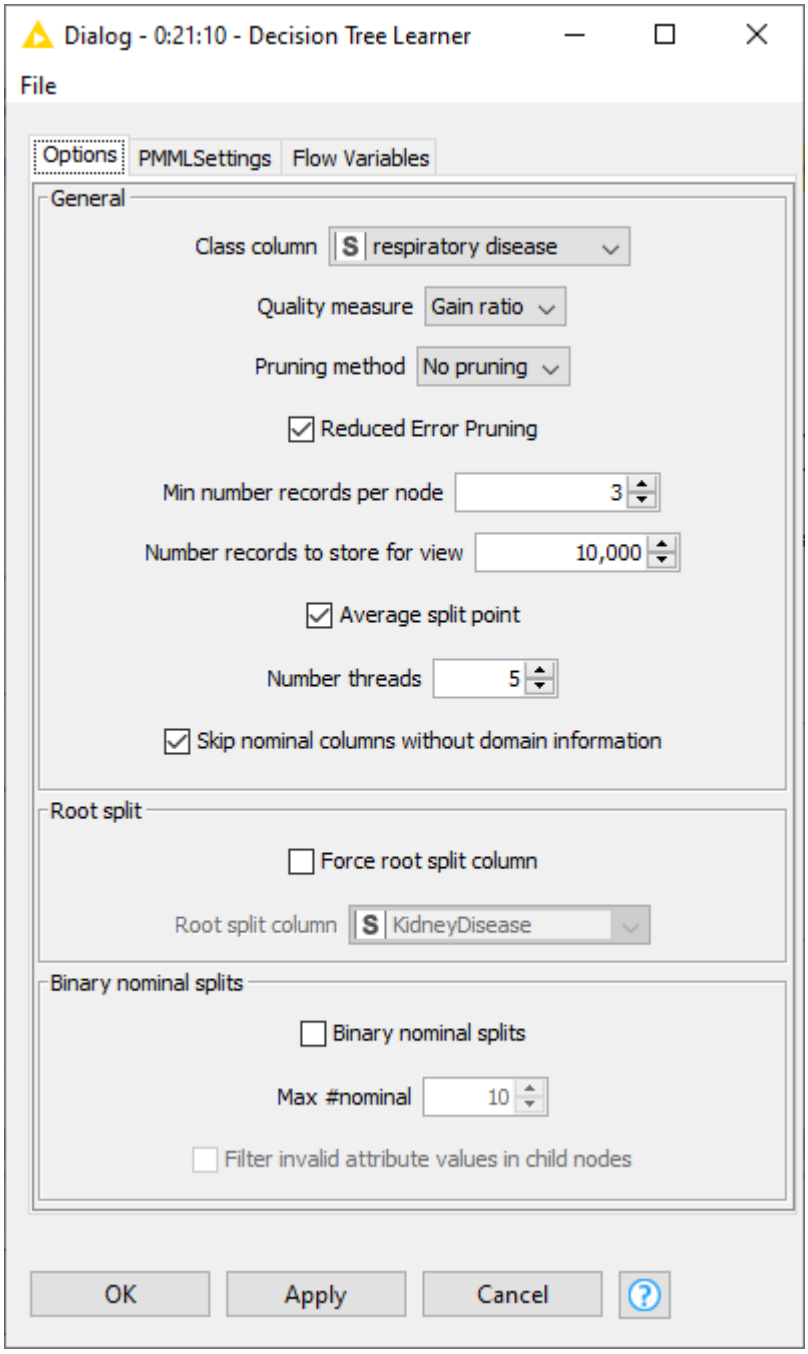
Cohen's kappa (κ) 0.879

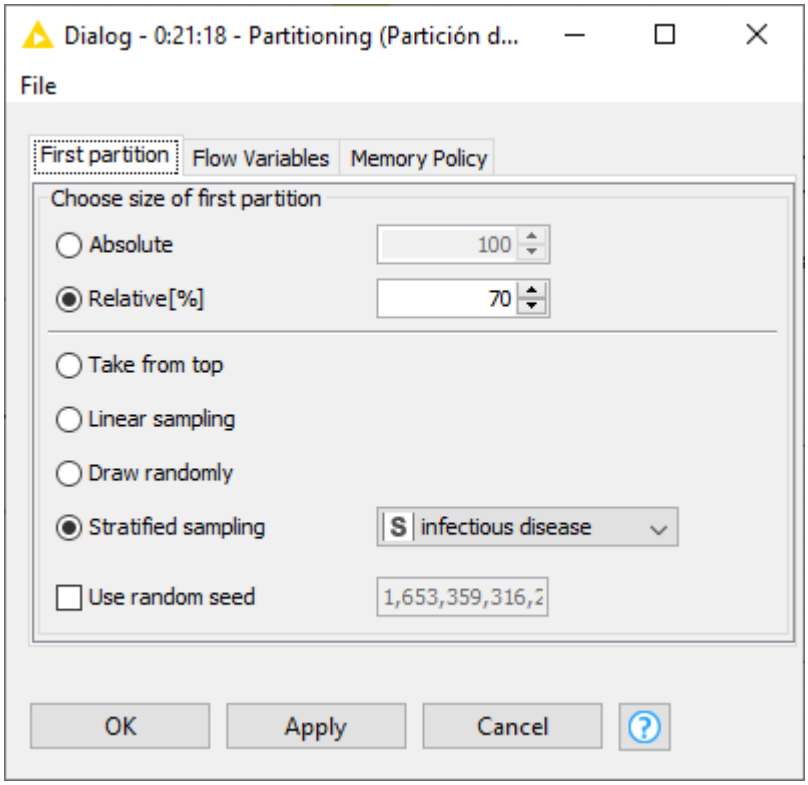
2.2.5 Pantallas de configuración

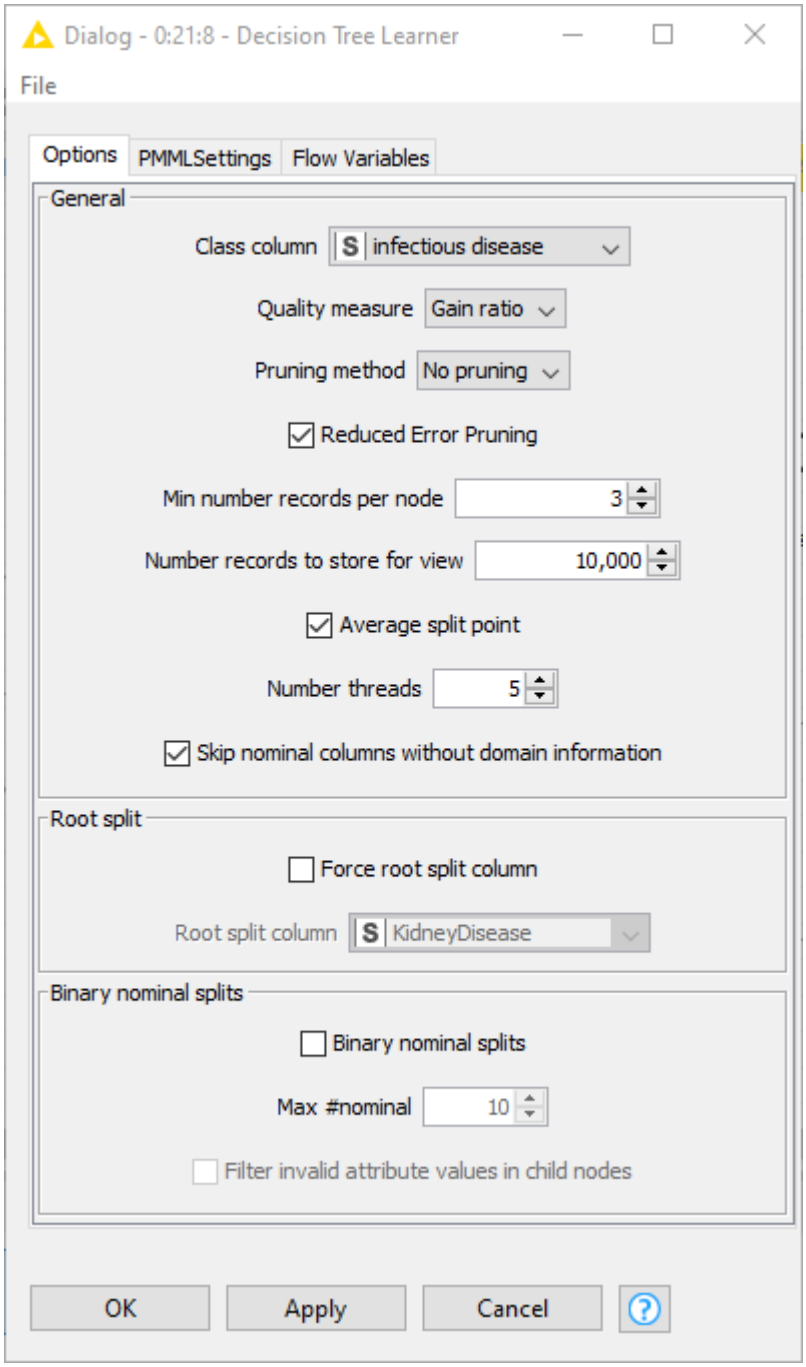

Dialog - 0:21:3 - Partitioning (Partición de ...
File
First partition
Flow Variables
Memory Policy
Choose size of first partition
Absolute
100
Relative[%]
70
Take from top
Linear sampling
Draw randomly
Stratified sampling
S
HeartDisease
Use random seed
1,653,358,419,7
OK
Apply
Cancel


Dialog - 0:21:4 - Decision Tree Learner
File
Options
PMMLSettings
Flow Variables
General
Class column
S HeartDisease
Quality measure
Gain ratio
Pruning method
No pruning
Reduced Error Pruning
Min number records per node
3
Number records to store for view
10,000
Average split point
Number threads
5
Skip nominal columns without domain information
Root split
Force root split column
Root split column
S KidneyDisease
Binary nominal splits
Binary nominal splits
Max #nominal
10
Filter invalid attribute values in child nodes
OK
Apply
Cancel


Dialog - 0:21:19 - Partitioning (Partición d...
File
First partition
Flow Variables
Memory Policy
Choose size of first partition
Absolute
100
Relative[%]
70
Take from top
Linear sampling
Draw randomly
Stratified sampling
S
respiratory disease
Use random seed
1,653,359,373,4
OK
Apply
Cancel


Dialog - 0:21:10 - Decision Tree Learner
File
Options
PMMLSettings
Flow Variables
General
Class column
S
respiratory disease
Quality measure
Gain ratio
Pruning method
No pruning
Reduced Error Pruning
Min number records per node
3
Number records to store for view
10,000
Average split point
Number threads
5
Skip nominal columns without domain information
Root split
Force root split column
Root split column
S
KidneyDisease
Binary nominal splits
Binary nominal splits
Max #nominal
10
Filter invalid attribute values in child nodes
OK
Apply
Cancel


Dialog - 0:21:18 - Partitioning (Partición d...
File
First partition
Flow Variables
Memory Policy
Choose size of first partition
Absolute
100
Relative[%]
70
Take from top
Linear sampling
Draw randomly
Stratified sampling
S
infectious disease
Use random seed
1,653,359,316,2
OK
Apply
Cancel


Dialog - 0:21:8 - Decision Tree Learner
File
Options
PMMLSettings
Flow Variables
General
Class column
S
infectious disease
Quality measure
Gain ratio
Pruning method
No pruning
Reduced Error Pruning
Min number records per node
3
Number records to store for view
10,000
Average split point
Number threads
5
Skip nominal columns without domain information
Root split
Force root split column
Root split column
S
KidneyDisease
Binary nominal splits
Binary nominal splits
Max #nominal
10
Filter invalid attribute values in child nodes
OK
Apply
Cancel