PARTE 1: CLASIFICACIÓN DE ASIGNATURAS DE PRIMERO EN LA TAXONOMÍA OBTENIDA UTILIZANDO LA REJILLA DE REPERTORIO SOBRE LAS ASIGNATURAS DE SEGUNDO

Del ejercicio de clase de la semana pasada habréis obtenido clases de asignaturas de acuerdo a vuestro propio criterio utilizando la técnica de la rejilla de repertorio sobre las asignaturas de segundo curso, jerarquizadas en forma de taxonomía simple. Además para cada grupo habréis obtenido los criterios que definen esos grupos

Se trata de aplicar esos criterios para decidir para cada asignatura de primero a que grupos de los obtenidos pertenece (su clasificación en la taxonomía). Si no os convence la clasificación, replantearos los criterios, e incluso en su caso la coherencia de los grupos obtenidos, indicando las asignaturas de primero para las que no os convence la clasificación y por qué.

Se entrega la taxonomía, los criterios de cada grupo, y la clasificación para cada una de estas asignatura:
Algebra Lineal y Estructuras Matemáticas; Cálculo; Fundamentos de Programación; Fundamentos de Software; Fundamentos Físicos y Tecnológicos; Estadística; Ingeniería, Empresa y Sociedad; Lógica y Métodos Discretos; Metodología de la Programación; y Tecnología y Organización de Computadores.

Añadir comentarios sobre si os parece lógica la clasificación.

Árbol del ejercicio anterior

Partiendo del árbol obtenido en la aplicación, tenemos nueve categorías diferentes en las que clasificar nuestras asignaturas de primero:

1. Todas las asignaturas.
2. Asignaturas orientadas a la programación.
3. Concurrencia y cálculos.
4. Software.
5. Procesamiento de datos.
6. Inteligencia artificial.
7. Programación + BD.
8. Bajo nivel.
9. Orientados a hardware.

Con más constructores, podríamos haber obtenido grupos más refinados. De partida, hay muchas asignaturas de primero relacionadas con la física , las matemáticas y las empresariales, como ALEM, Cálculo, FFT, Estadística, IES y LMD. Estas asignaturas no encuadran objetivamente en ninguna de nuestras categorías, ya que en ningún constructor hemos tenido en cuenta la carga matemática, física o de empresariales que tiene una asignatura, ni hemos hecho un constructor relacionado. Lo más lógico sería añadir constructores y realizar de nuevo todo el proceso.

Otras asignaturas sin embargo podrían estar en alguno de los grupos formados:

- ➔ FP: la incluiría en el 2. Si refinamos más, debe formar un grupo con ED.
- ➔ FS: en el 9. Preparatorio de SO.
- ➔ MP: ídem a FP.
- ➔ TOC: en el grupo 8. Se trabaja con puertas lógicas y álgebra de Boole, que es con lo que trabaja el hardware a bajo nivel.

PARTE 2: CONOCIMIENTO PARA DESARROLLAR SBC PARA ACONSEJAR QUE RAMA ELEGIR UTILIZANDO APRENDIZAJE DE ÁRBOLES DE DECISIÓN

Seguramente ya habréis comprobado que asesorar en qué asignaturas matricularse es un problema complejo con muchas opciones distintas. Por ello nos vamos a centrar en un subproblema concreto, aconsejar que rama o mención elegir.

Supongamos que inicialmente vamos a intentar aconsejar que rama elegir a partir de las siguientes características del alumno que pide consejo:

- Si le gusta las matemáticas
- En qué quiere trabajar
- La nota media que ha obtenido en las asignaturas que ha cursado hasta ahora
- Cómo de trabajador se considera
- Si le gusta programar
- Si prefiere asignaturas teóricas o prácticas

Para ello vamos a recoger una serie de alumnos y preparamos una tabla para que el experto rellene cual sería la respuesta o respuestas posibles para ese caso:

Caso	Gusta Matemáticas	Quiere trabajar	Nota media	Gusta hardware	Es trabajador	Gusta Programar	Prefiere teóricas o prácticas	Rama(s) aconsejada
Alumno 1	Si	Docencia	Alta	No	Mucho	Si	teóricas	**CSI**
Alumno 2	No	Empresa Pública	Media	No	Normal	No	prácticas	**SI**
Alumno 3	Si	Empresa Privada	Media	Si	Normal	Si	ambas	**CSI**
	Si	Empresa Privada	Media	Si	Normal	Si	ambas	**IS**
Alumno 4	No	Empresa Privada	Baja	No	Poco	Si	ambas	**IS**
	No	Empresa Privada	Baja	No	Poco	Si	ambas	**TI**
Alumno 5	No	le da igual	Alta	Si	Mucho	Si	prácticas	**IC**

Alumno 6	Si	Docencia	Media	No	Poco	Si	teóricas	**TI**
	Si	Docencia	Media	No	Poco	Si	teóricas	**IS**
Alumno 7	No	Docencia	Alta	No	Normal	Si	prácticas	**IS**
Alumno 8	No	Empresa Pública	Baja	Si	Normal	No	ambas	**IC**
Alumno 9	Si	Empresa Privada	Alta	No	Normal	Si	ambas	**CSI**
	Si	Empresa Privada	Alta	No	Normal	Si	ambas	**SI**
	Si	Empresa Privada	Alta	No	Normal	Si	ambas	**IS**
Alumno 10	No	Empresa Pública	Baja	No	Poco	Si	prácticas	**IS**
Alumno 11	Si	le da igual	Baja	No	Poco	Si	ambas	**TI**
Alumno 12	Si	Empresa Privada	Alta	Si	Normal	No	prácticas	**SI**
Alumno 13	No	Empresa Privada	Baja	Si	Poco	Si	prácticas	**IC**
	No	Empresa Privada	Baja	Si	Poco	Si	prácticas	**IS**
Alumno 14	No	Empresa Privada	Alta	Si	Normal	Si	ambas	**SI**
	No	Empresa Privada	Alta	Si	Normal	Si	ambas	**IC**
Alumno 15	Si	Docencia	Media	No	Mucho	Si	teóricas	**CSI**

Alumno 16	No	Docencia	Media	No	Normal	No	teóricas	**SI**
Alumno 17	No	Empresa Pública	Media	Si	Poco	No	prácticas	**IC**
Alumno 18	Si	le da igual	Media	No	Normal	Si	ambas	**CSI**
	Si	le da igual	Media	No	Normal	Si	ambas	**IS**
Alumno 19	No	Empresa Pública	Alta	Si	Mucho	Si	prácticas	**IS**
Alumno 20	Si	le da igual	baja	No	Poco	No	teóricas	**CSI**

Ejercicio propuesto:

0) Opcional, antes de empezar a aplicar la técnica del aprendizaje de árboles de decisión, añade o modifica características, modifica los valores posibles (adaptando la tabla), y añade ejemplos de alumnos que consideres que serían interesantes, obteniendo una nueva tabla adaptada a tu criterio sobre la que trabajar.

1) Haz tú de experto y rellena el campo de las Ramas Aconsejadas indicando la rama o las ramas que aconsejarias en cada caso (en muchos casos pensarás le puedes aconsejar más de una)
2) Aplicar un algoritmo de aprendizaje para obtener un árbol de clasificación a partir de esa tabla rellena (Ojo cuando tengas dos o más ramas aconsejadas para un alumno, el árbol deberás obtenerlo creando un ejemplo para cada rama, además tendrá que decidir cómo manejar los atributos donde algún valor no es incompatible con otros)
3) Obtener las reglas de decisión asociadas al árbol obtenido
4) Analizar las reglas obtenidas:
    a) En el caso de que no se consideren válidas, proponer modificaciones a las mismas o añadir ejemplos que complementen las posibles situaciones y que son las que harían no válidas esa regla. Si se añaden nuevos casos revisar de nuevo las reglas a obtener.
    b) En el caso donde el algoritmo de aprendizaje no sea capaz de decidir, pero si hay muchas más posibilidades de que ocurra una de las posibles opciones, proponer esa como regla por defecto y analizar bajo qué condiciones habría que descartar la opción por defecto.
    c) En el caso de varias opciones sin que una sea claramente preponderante, analizar cómo podría tratarse esa situación.

Se entrega una descripción del proceso seguido: la tabla propia inicial, las reglas iniciales obtenidas por el algoritmo, y para cada regla inicial si se considera adecuado o si se debe modificar, en cuyo se indica las modificaciones y el por qué.

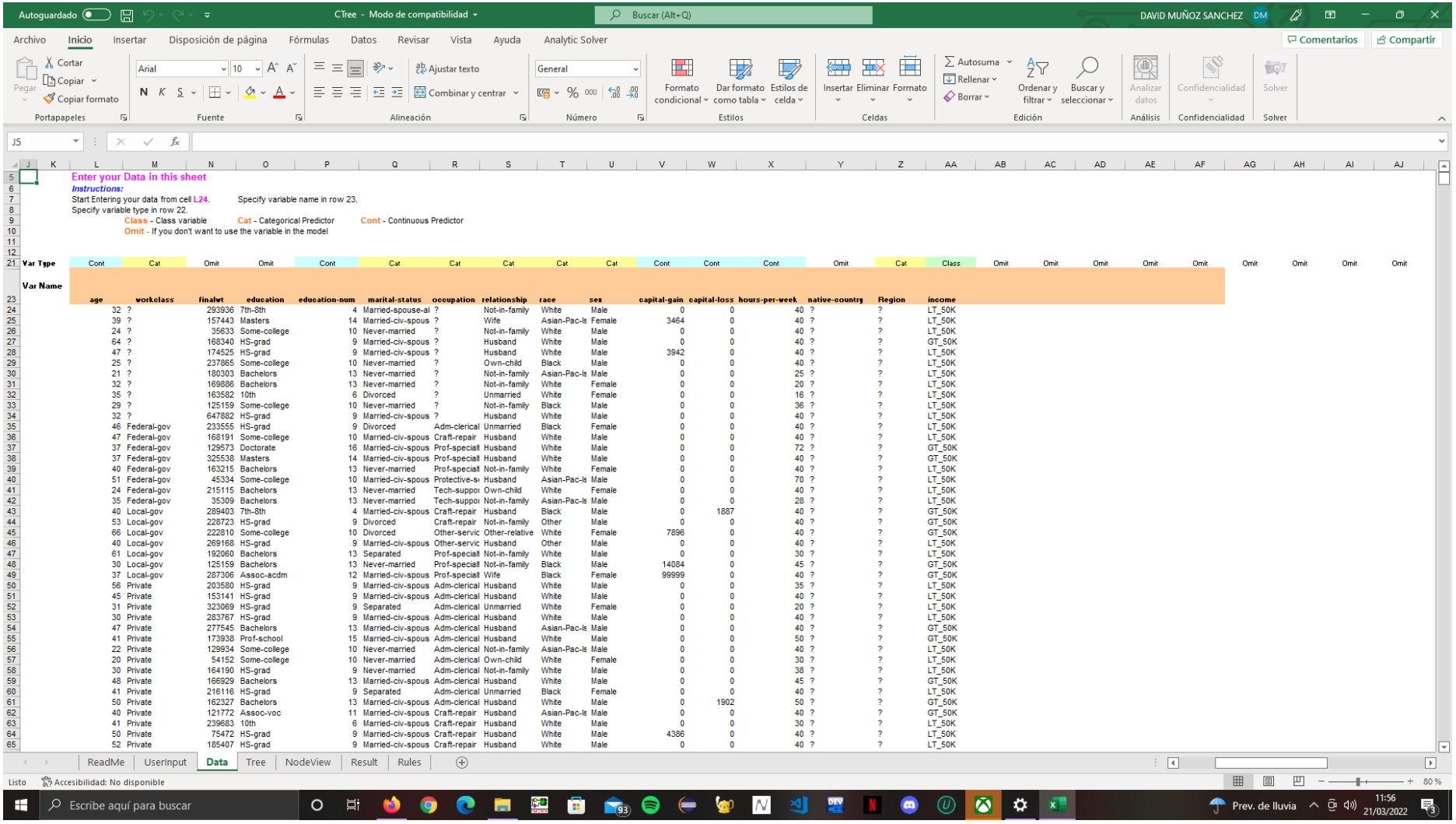
La tabla en el enunciado está rellena. En muchos casos se puede recomendar más de una rama. Las ramas son:

- CSI (Computación y Sistemas Inteligentes).
- IC (Ingeniería de Computadores).
- IS (Ingeniería del Software).
- SI (Sistemas de Información).
- TI (Tecnologías de la Información).

El software de árbol de decisión que voy a utilizar será CTree. Es una hoja de cálculo, adjunto captura de pantalla del programa en la hoja siguiente (número 7).

En la zona de datos que es la que se adjunta en la pantalla, debemos introducir nuestra tabla, todo son categorías menos la rama elegida, que es una clase.

En la página 8 se adjunta captura de pantalla de los datos introducidos y continúa el documento (se adjuntan capturas de la configuración para generar el árbol y del árbol resultante, así como otra información relativa al árbol de decisión).



Enter your Data in this sheet

Instructions:

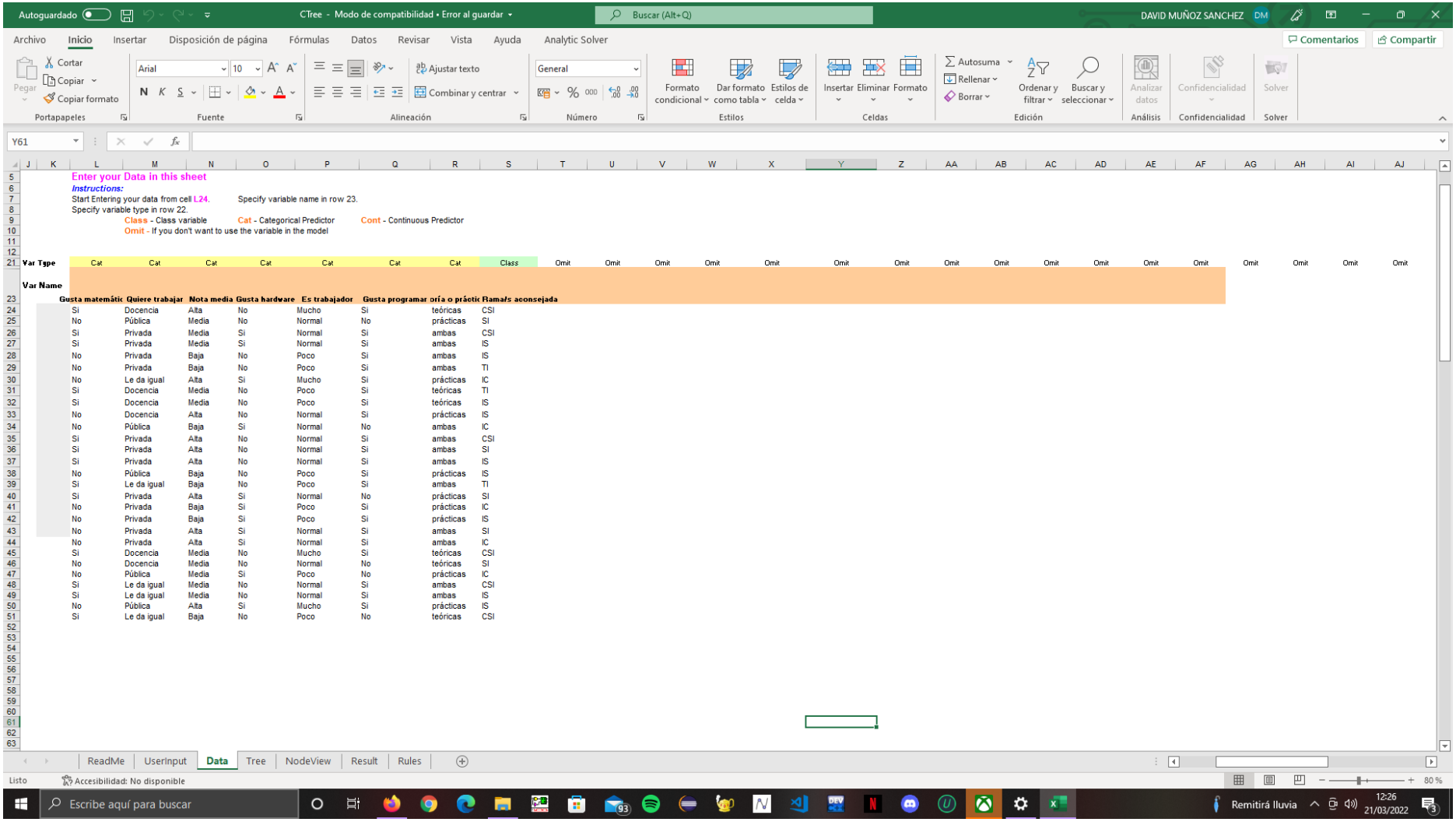
Start Entering your data from cell L24. Specify variable name in row 23.

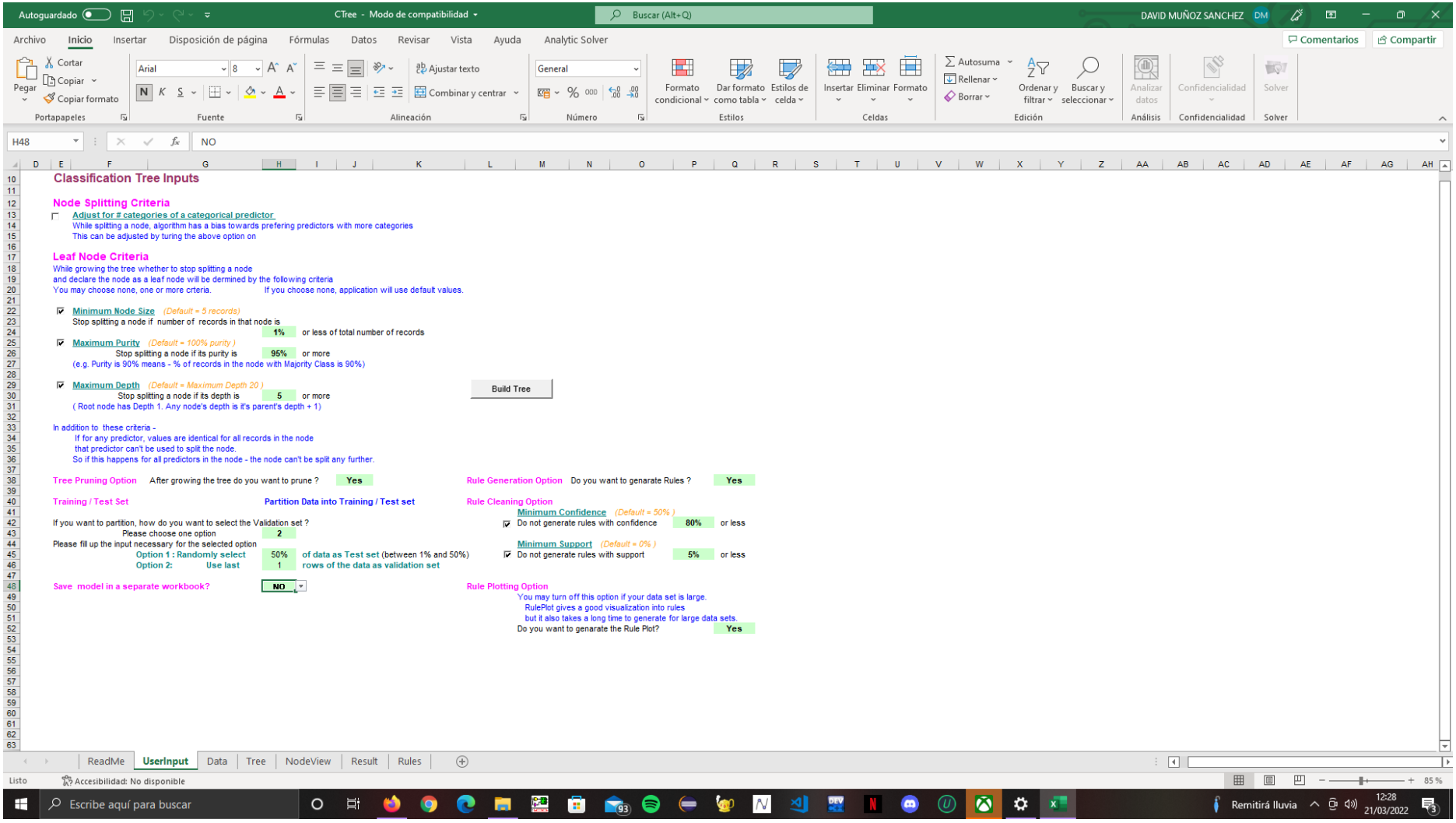
Specify variable type in row 22.

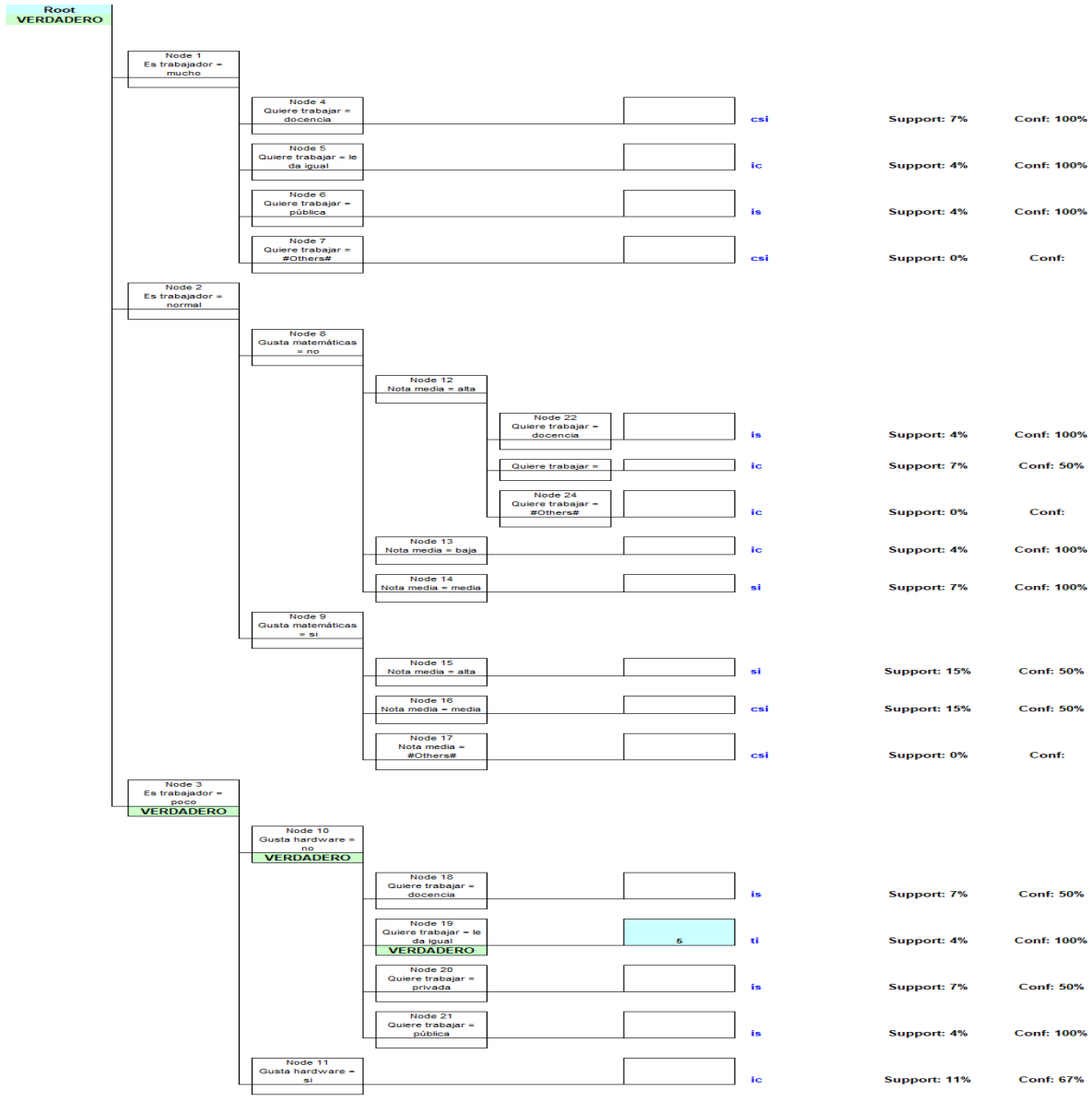
Class - Class variable Cat - Categorical Predictor Cont - Continuous Predictor

Omit - If you don't want to use the variable in the model

Var Type	Cont	Cat	Omit	Omit	Cont	Cat	Cat	Cat	Cat	Cat	Cont	Cont	Cont	Omit	Cat	Class	Omit	Omit	Omit	Omit	Omit	Omit	Omit	Omit	Omit
Var Name	age	workclass	finalwt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	Region	income									
23																									
24	32	?	293936	7th-8th	4	Married-spouse-al	?	Not-in-family	White	Male	0	0	40	?	?	LT_50K									
25	39	?	157443	Masters	14	Married-civ-spous	?	Wife	Asian-Pac-Is	Female	3464	0	40	?	?	LT_50K									
26	24	?	35633	Some-college	10	Never-married	?	Not-in-family	White	Male	0	0	40	?	?	LT_50K									
27	64	?	168340	HS-grad	9	Married-civ-spous	?	Husband	White	Male	0	0	40	?	?	GT_50K									
28	47	?	174525	HS-grad	9	Married-civ-spous	?	Husband	White	Male	3942	0	40	?	?	LT_50K									
29	25	?	237865	Some-college	10	Never-married	?	Own-child	Black	Male	0	0	40	?	?	LT_50K									
30	21	?	180303	Bachelors	13	Never-married	?	Not-in-family	Asian-Pac-Is	Male	0	0	25	?	?	LT_50K									
31	32	?	169886	Bachelors	13	Never-married	?	Not-in-family	White	Female	0	0	20	?	?	LT_50K									
32	35	?	163582	10th	6	Divorced	?	Unmarried	White	Female	0	0	16	?	?	LT_50K									
33	29	?	125159	Some-college	10	Never-married	?	Not-in-family	Black	Male	0	0	36	?	?	LT_50K									
34	32	?	647882	HS-grad	9	Married-civ-spous	?	Husband	White	Male	0	0	40	?	?	LT_50K									
35	46	Federal-gov	233555	HS-grad	9	Divorced	Adm-clerical	Unmarried	Black	Female	0	0	40	?	?	LT_50K									
36	47	Federal-gov	168191	Some-college	10	Married-civ-spous	Craft-repair	Husband	White	Male	0	0	40	?	?	LT_50K									
37	37	Federal-gov	129573	Doctorate	16	Married-civ-spous	Prof-specialt	Husband	White	Male	0	0	72	?	?	GT_50K									
38	37	Federal-gov	325538	Masters	14	Married-civ-spous	Prof-specialt	Husband	White	Male	0	0	40	?	?	GT_50K									
39	40	Federal-gov	163215	Bachelors	13	Never-married	Prof-specialt	Not-in-family	White	Female	0	0	40	?	?	LT_50K									
40	51	Federal-gov	45334	Some-college	10	Married-civ-spous	Protective-s	Husband	Asian-Pac-Is	Male	0	0	70	?	?	LT_50K									
41	24	Federal-gov	215115	Bachelors	13	Never-married	Tech-suppo	Own-child	White	Female	0	0	40	?	?	LT_50K									
42	35	Federal-gov	35309	Bachelors	13	Never-married	Tech-suppo	Not-in-family	Asian-Pac-Is	Male	0	0	28	?	?	LT_50K									
43	40	Local-gov	289403	7th-8th	4	Married-civ-spous	Craft-repair	Husband	Black	Male	0	1887	40	?	?	GT_50K									
44	53	Local-gov	228723	HS-grad	9	Divorced	Craft-repair	Not-in-family	Other	Male	0	0	40	?	?	GT_50K									
45	66	Local-gov	222810	Some-college	10	Divorced	Other-servic	Other-relative	White	Female	7896	0	40	?	?	GT_50K									
46	40	Local-gov	269168	HS-grad	9	Married-civ-spous	Other-servic	Husband	Other	Male	0	0	40	?	?	LT_50K									
47	61	Local-gov	192060	Bachelors	13	Separated	Prof-specialt	Not-in-family	White	Male	0	0	30	?	?	LT_50K									
48	30	Local-gov	125159	Bachelors	13	Never-married	Prof-specialt	Not-in-family	Black	Male	14084	0	45	?	?	GT_50K									
49	37	Local-gov	287306	Assoc-acdm	12	Married-civ-spous	Prof-specialt	Wife	Black	Female	99999	0	40	?	?	GT_50K									
50	56	Private	203580	HS-grad	9	Married-civ-spous	Adm-clerical	Husband	White	Male	0	0	35	?	?	LT_50K									
51	45	Private	153141	HS-grad	9	Married-civ-spous	Adm-clerical	Husband	White	Male	0	0	40	?	?	LT_50K									
52	31	Private	323069	HS-grad	9	Separated	Adm-clerical	Unmarried	White	Female	0	0	20	?	?	LT_50K									
53	30	Private	283767	HS-grad	9	Married-civ-spous	Adm-clerical	Husband	White	Male	0	0	40	?	?	LT_50K									
54	47	Private	277545	Bachelors	13	Married-civ-spous	Adm-clerical	Husband	Asian-Pac-Is	Male	0	0	40	?	?	GT_50K									
55	41	Private	173938	Prof-school	15	Married-civ-spous	Adm-clerical	Husband	White	Male	0	0	50	?	?	GT_50K									
56	22	Private	129934	Some-college	10	Never-married	Adm-clerical	Not-in-family	Asian-Pac-Is	Male	0	0	40	?	?	LT_50K									
57	20	Private	54152	Some-college	10	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	?	?	LT_50K									
58	30	Private	164190	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	38	?	?	LT_50K									
59	48	Private	166929	Bachelors	13	Married-civ-spous	Adm-clerical	Husband	White	Male	0	0	45	?	?	GT_50K									
60	41	Private	216116	HS-grad	9	Separated	Adm-clerical	Unmarried	Black	Female	0	0	40	?	?	LT_50K									
61	50	Private	162327	Bachelors	13	Married-civ-spous	Adm-clerical	Husband	White	Male	0	1902	50	?	?	GT_50K									
62	40	Private	121772	Assoc-voc	11	Married-civ-spous	Craft-repair	Husband	Asian-Pac-Is	Male	0	0	40	?	?	GT_50K									
63	41	Private	239683	10th	6	Married-civ-spous	Craft-repair	Husband	White	Male	0	0	30	?	?	LT_50K									
64	50	Private	75472	HS-grad	9	Married-civ-spous	Craft-repair	Husband	White	Male	4386	0	40	?	?	LT_50K									
65	52	Private	185407	HS-grad	9	Married-civ-spous	Craft-repair	Husband	White	Male	0	0	40	?	?	LT_50K									


Autoguardado
CTree - Modo de compatibilidad • Error al guardar
Buscar (Alt+Q)
DAVID MUÑOZ SANCHEZ
Archivo Inicio Insertar Disposición de página Fórmulas Datos Revisar Vista Ayuda Analytic Solver
Comentarios Compartir
Y61
Enter your Data in this sheet
Instructions:
Start Entering your data from cell L24.
Specify variable name in row 23.
Specify variable type in row 22.
Class - Class variable
Cat - Categorical Predictor
Cont - Continuous Predictor
Omit - If you don't want to use the variable in the model
Var Type
Var Name
Gusta matemátic Quiere trabajar Nota media Gusta hardware Es trabajador Gusta programar oría o práctic Rama/s aconsejada
Si Docencia Alta No Mucho Si teóricas CSI
No Pública Media No Normal No prácticas SI
Si Privada Media Si Normal Si ambas CSI
Si Privada Media Si Normal Si ambas IS
No Privada Baja No Poco Si ambas IS
No Privada Baja No Poco Si ambas TI
No Le da igual Alta Si Mucho Si prácticas IC
Si Docencia Media No Poco Si teóricas TI
Si Docencia Media No Poco Si teóricas IS
No Docencia Alta No Normal Si prácticas IS
No Pública Baja Si Normal No ambas IC
Si Privada Alta No Normal Si ambas CSI
Si Privada Alta No Normal Si ambas SI
Si Privada Alta No Normal Si ambas IS
No Pública Baja No Poco Si prácticas IS
Si Le da igual Baja No Poco Si ambas TI
Si Privada Alta Si Normal No prácticas SI
No Privada Baja Si Poco Si prácticas IC
No Privada Baja Si Poco Si prácticas IS
No Privada Alta Si Normal Si ambas SI
No Privada Alta Si Normal Si ambas IC
Si Docencia Media No Mucho Si teóricas CSI
No Docencia Media No Normal No teóricas SI
No Pública Media Si Poco No prácticas IC
Si Le da igual Media No Normal Si ambas CSI
Si Le da igual Media No Normal Si ambas IS
No Pública Alta Si Mucho Si prácticas IS
Si Le da igual Baja No Poco No teóricas CSI
ReadMe UserInput Data Tree NodeView Result Rules
Listo
Accesibilidad: No disponible
80 %
Escribe aquí para buscar
Remitirá lluvia
12:26
21/03/2022


Autoguardado
CTree - Modo de compatibilidad
Buscar (Alt+Q)
DAVID MUÑOZ SANCHEZ
Archivo Inicio Insertar Disposición de página Fórmulas Datos Revisar Vista Ayuda Analytic Solver
Comentarios Compartir
H48
NO
Classification Tree Inputs
Node Splitting Criteria
Adjust for # categories of a categorical predictor
While splitting a node, algorithm has a bias towards prefering predictors with more categories
This can be adjusted by turing the above option on
Leaf Node Criteria
While growing the tree whether to stop splitting a node
and declare the node as a leaf node will be dermined by the following criteria
You may choose none, one or more crteria.
If you choose none, application will use default values.
Minimum Node Size (Default = 5 records)
Stop splitting a node if number of records in that node is
1% or less of total number of records
Maximum Purity (Default = 100% purity)
Stop splitting a node if its purity is
95% or more
(e.g. Purity is 90% means - % of records in the node with Majority Class is 90%)
Maximum Depth (Default = Maximum Depth 20)
Stop splitting a node if its depth is
5 or more
(Root node has Depth 1. Any node's depth is it's parent's depth + 1)
Build Tree
In addition to these criteria -
If for any predictor, values are identical for all records in the node
that predictor can't be used to split the node.
So if this happens for all predictors in the node - the node can't be split any further.
Tree Pruning Option
After growing the tree do you want to prune ?
Yes
Rule Generation Option
Do you want to genarate Rules ?
Yes
Training / Test Set
Partition Data into Training / Test set
If you want to partition, how do you want to select the Validation set ?
Please choose one option
2
Please fill up the input necessary for the selected option
Option 1 : Randomly select
50%
of data as Test set (between 1% and 50%)
Option 2: Use last
1
rows of the data as validation set
Save model in a separate workbook?
NO
Rule Cleaning Option
Minimum Confidence (Default = 50%)
Do not generate rules with confidence
80%
or less
Minimum Support (Default = 0%)
Do not generate rules with support
5%
or less
Rule Plotting Option
You may turn off this option if your data set is large.
RulePlot gives a good visualization into rules
but it also takes a long time to generate for large data sets.
Do you want to genarate the Rule Plot?
Yes
ReadMe UserInput Data Tree NodeView Result Rules
Listo
Accesibilidad: No disponible
85 %
Escribe aquí para buscar
Remitirá lluvia
12:28
21/03/2022


Root
VERDADERO
Node 1
Es trabajador = mucho
Node 4
Quiere trabajar = docencia
csi
Support: 7%
Conf: 100%
Node 5
Quiere trabajar = le da igual
ic
Support: 4%
Conf: 100%
Node 6
Quiere trabajar = pública
is
Support: 4%
Conf: 100%
Node 7
Quiere trabajar = #Others#
csi
Support: 0%
Conf:
Node 2
Es trabajador = normal
Node 8
Gusta matemáticas = no
Node 12
Nota media = alta
Node 22
Quiere trabajar = docencia
is
Support: 4%
Conf: 100%
Quiere trabajar =
ic
Support: 7%
Conf: 50%
Node 24
Quiere trabajar = #Others#
ic
Support: 0%
Conf:
Node 13
Nota media = baja
ic
Support: 4%
Conf: 100%
Node 14
Nota media = media
si
Support: 7%
Conf: 100%
Node 9
Gusta matemáticas = si
Node 15
Nota media = alta
si
Support: 15%
Conf: 50%
Node 16
Nota media = media
csi
Support: 15%
Conf: 50%
Node 17
Nota media = #Others#
csi
Support: 0%
Conf:
Node 3
Es trabajador = poco
VERDADERO
Node 10
Gusta hardware = no
VERDADERO
Node 18
Quiere trabajar = docencia
is
Support: 7%
Conf: 50%
Node 19
Quiere trabajar = le da igual
VERDADERO
5
ti
Support: 4%
Conf: 100%
Node 20
Quiere trabajar = privada
is
Support: 7%
Conf: 50%
Node 21
Quiere trabajar = pública
is
Support: 4%
Conf: 100%
Node 11
Gusta hardware = si
ic
Support: 11%
Conf: 67%

La foto no se aprecia bien a simple vista, pero adjunto la foto para que se pueda observar mejor.

A partir del árbol generado obtenemos las siguientes reglas:

1. Si es muy trabajador y quiere trabajar en la empresa privada o en la docencia recomienda CSI.
2. Si es muy trabajador y le da igual donde trabajar recomienda IC.
3. Si es muy trabajador y quiere trabajar en la pública, recomienda IS.
4. Si es normal trabajando, no le gustan las matemáticas, su nota media es alta y quiere trabajar en cualquier sitio menos la docencia, recomienda IC.
5. Ídem a la anterior pero quiere trabajar en la docencia, entonces recomienda IS.
6. Igual que la cuatro pero, su nota media es media y no tenemos en cuenta donde quiere trabajar, recomienda SI.
7. Igual a la anterior pero la nota media es baja, recomienda IC.
8. Si es normal trabajando, le gustan las matemáticas y tiene nota media alta, SI.
9. Igual que antes pero con nota media media, recomienda CSI.
10. Igual pero la nota media es baja, recomienda CSI.
11. Si es poco trabajador, no le gusta el hardware y quiere trabajar en la privada, en la pública o en docencia, IS.
12. Igual pero si no especifica dónde trabajar, TI.
13. Igual que la 11 pero si le gusta el hardware recomienda IC.

El árbol de decisión presenta varios problemas. La opción primera de si trabaja mucho, poco o normal debería cambiarse por hardware y software, ya que después tendríamos menos reglas incoherentes, como en la 11 y la 12, que recomienda TI e IS indistintamente, o en la rama de muy trabajador, que no se ha tenido en cuenta la carga matemática, siendo determinante para recomendar CSI.

El árbol se debería hacer con más casos y más columnas de información en la tabla. En mi opinión, el resultado que he obtenido, está demasiado condicionado por el lugar dónde quiere trabajar el alumno y lo que él piensa que trabaja. La regla por defecto podría ser IS, que es mayoritaria (un 33,33% de proporción con respecto a las demás).

Fuentes

Saha, Angshuman. "Data Mining in Excel - SayHello2Angshu." *Google Sites*, https://www.sites.google.com/site/sayhello2angshu/dminexcel. Accessed 21 March 2022.