PRUEBA DE EVALUACIÓN 1 KNIME

SALARIOS EN ESTADOS UNIDOS

17 DE FEBRERO DE 2023

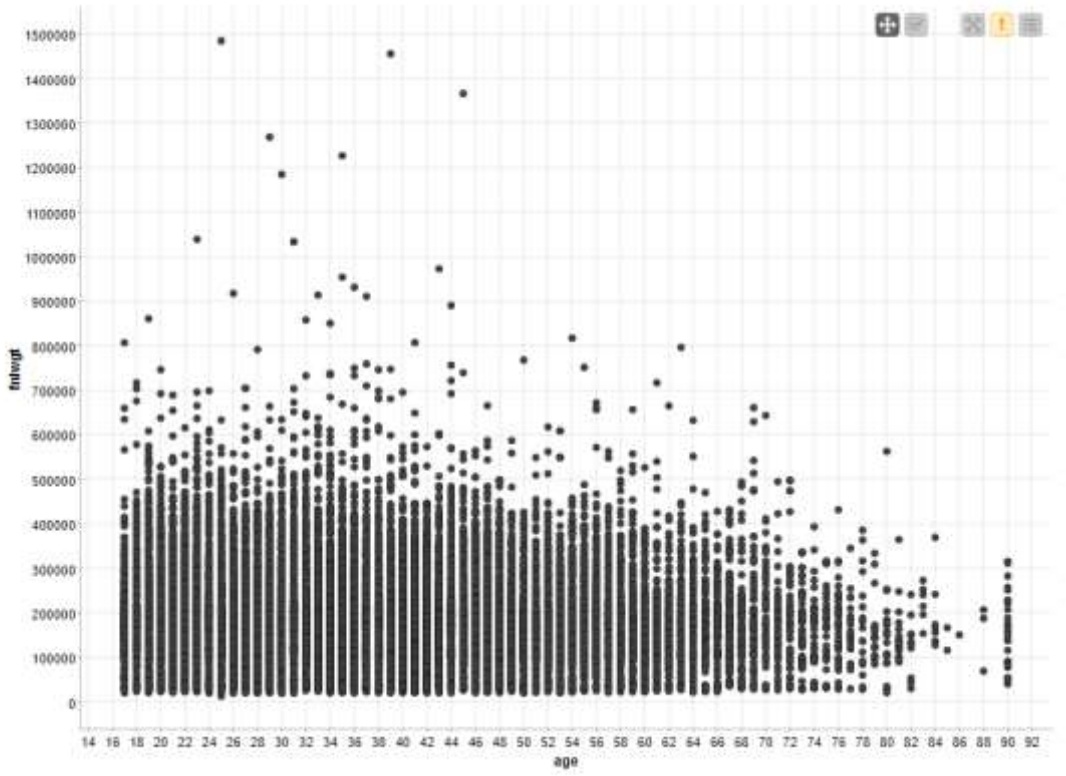
ALFREDO SALVADOR TÉLLEZ

UNIVERSIDAD ALCALÁ DE HENARES

EXPLORACIÓN

Indagamos y observamos los datos con detenimiento, para ellos es imprescindible la preparación previa de ellos para un posterior análisis. La base de datos a tratar se llama “salary” donde intentaremos analizar cómo influye el salario dependiendo de la edad de los trabajadores en E.E.U.U.

Una vez cargado el dataframe podemos observar la distribución de los datos analizando el salario final anual a medida que pasan los años. Donde podemos concretar que en los años intermedios de madurez laboral es donde más se cobra y descienden progresivamente hasta el final donde se puede entender que es la pensión de jubilación. También se pueden observar datos atípicos de personas que cobran muy por encima de la media, donde como hemos comentado también serían de personas de entre 25 y 40 años.



Utilizamos el nodo CrossTab para visualizar las relaciones entre variables. Como la variable “salary” distribuye los datos en < o > de 50.000 dólares podemos estudiar el porcentaje.

Cross Tabulation of salary by age

Frequency Row Percent	17	18	19	20	21	22	23	24	25	26	... (63)	Total
<=50K	73.360.282	1,06E8	1,45E8	1,40E8	1,43E8	1,52E8	1,79E8	1,57E8	1,60E8	1,40E8		4,71E9
	1,5591%	2,2622%	3,088%	3,1773%	3,0389%	3,2306%	3,8003%	3,3368%	3,4039%	2,9689%		
>50K												1,47E9
Total	73.360.282	1,06E8	1,46E8	1,40E8	1,44E8	1,53E8	1,81E8	1,64E8	1,71E8	1,51E8		6,18E9

Frequency
Expected
Deviation
Percent
Row Percent
Column Percent
Cell Chi-Square

Max rows: 10
Max columns: 10

Statistics for Table of salary by age

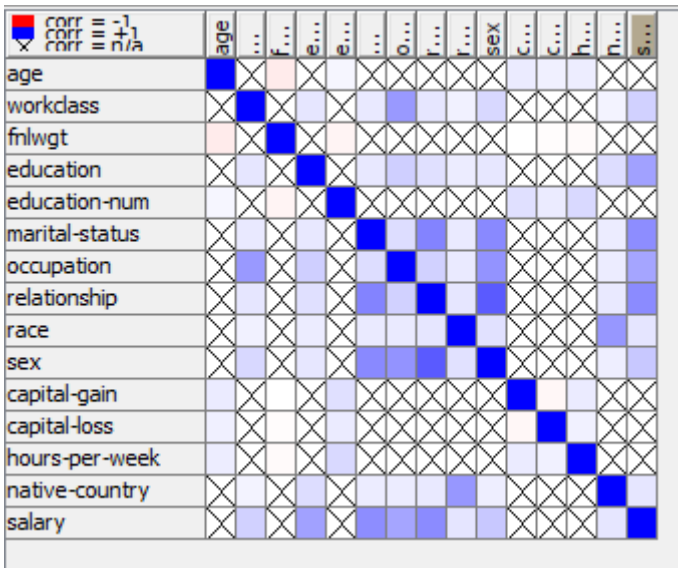
Statistic	DF	Value	Prob
Chi-Square	72	6,20E8	0.0

Total sample size: 6.17937339289

Otro dato para analizar previamente es la correlación de las variables. Decidimos estudiar todas, aunque realmente solo nos interesan dos o tres de ellas.

La variable “salary” está notablemente más correlacionada con otras tres variables; ocupación, estado civil y relaciones. Es decir, el salario está ligado al puesto de trabajo que emplees y un factor importante es si estás casado o no, donde una persona casada puede cobrar más que una persona soltera.

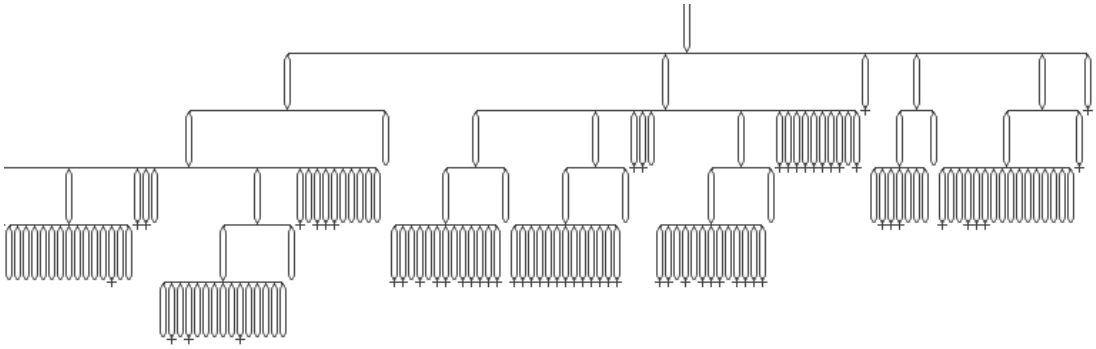
Otro factor para destacar es la variable “sexo” donde el valor de correlación es el más alto de la tabla. Cerca del 70% con la variable relación.



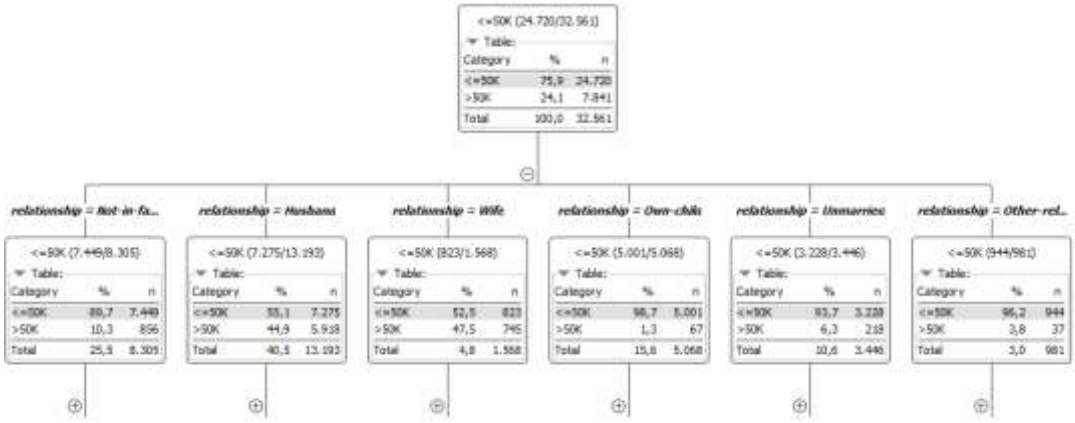
TRAINING

En los datos de entrenamiento podemos introducir un árbol previo sin podado todavía para analizar los datos y poder predecir. Nos fijamos en los nodos de árbol de decisión graficado en una imagen para

plasmar el árbol.



Una vez que el programa ha terminado de construir el árbol, se pueden calcular las puntuaciones predictivas. La puntuación predictiva es un porcentaje del indicador objetivo en el nodo terminal o rama del modelo entrenado.
Las predicciones funcionan de tal forma, si el individuo por ejemplo cae en la rama de los individuos con mujer tiene un 47,5 % de cobrar más de 50 mil dólares.
Esto demuestra que un modelo puede predecir si ciertos átomos tienen una mayor propensión a un indicador objetivo. Es importante recordar que es una predicción, que no es 100% precisa, sin embargo, es mucho más precisa que una suposición aleatoria.



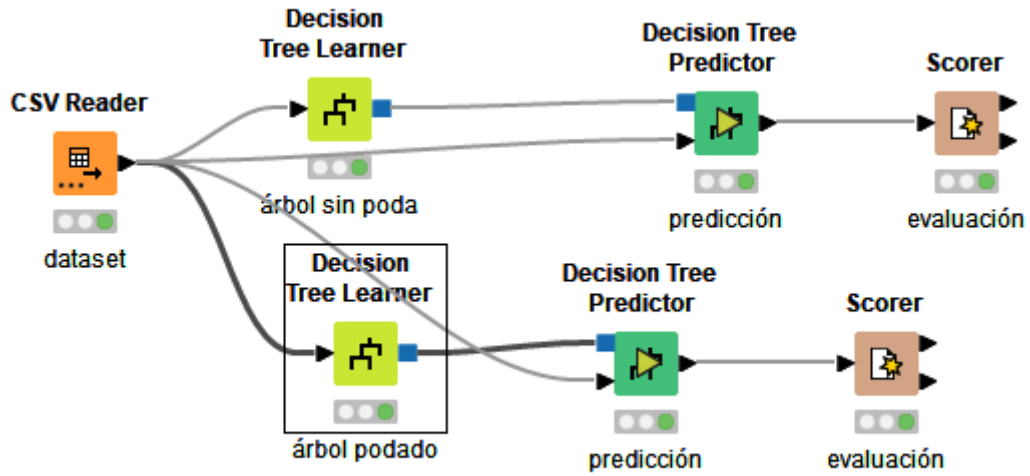
EVALUACIÓN

En esta parte añadimos una segunda rama de nuestra estructura, árboles de decisión podados.

Es fácil crear un árbol de decisión demasiado complejo y no poder interpretarlo bien. La simplicidad es la clave para reducir el esesgo, si es demasiado grande deja de sernos útil, y podemos caer en un sobreajuste de los datos.

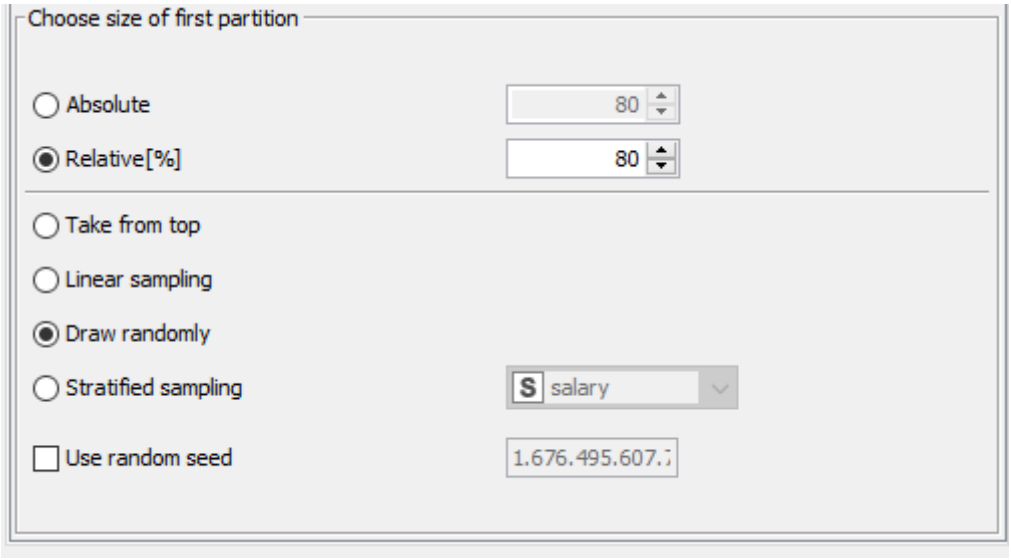
La intención es crear el árbol suficientemente especifico que nos pueda aportar nueva información y esto se consigue mediante la poda. La idea es crear el árbol complejo e ir eliminado los niveles suficientes para

que no exista dificultad de predicción, pero sean precisos en esa predicción. Usamos una parte de los datos de entramiento para usarlos como datos de validación.



OVERFITTING

En este apartado intentaremos controlar el efecto de sobreajuste haciendo una partición inicial del 80% para los datos test y el 20 % para datos de entrenamiento. Usamos los datos reservados para la evaluación y repetimos el proceso para obtener un modelo más realista que los anteriores.



Utilizando la tabla interactiva para analizar mejor los datos, vemos que el coeficiente de determinación es -248. Es decir, que a medida que aumentan los años, el salario disminuye, como hemos comentado anteriormente con el gráfico.

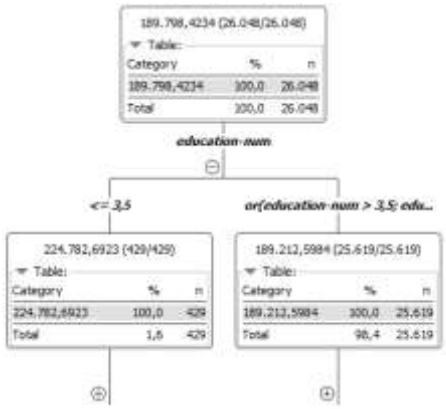
Row ID	D fnlwgt
R^2	-248,832,439....
mean absolut...	189,937.868
mean square...	47,657,918,31...
root mean sq...	218,306.936
mean signed ...	189,937.868
mean absolut...	5,724.849
adjusted R^2	-248,832,439....

File

R²:	-248.832.439,511
Mean absolute error:	189.937,868
Mean squared error:	47.657.918.312,639
Root mean squared error:	218.306,936
Mean signed difference:	189.937,868
Mean absolute percentage error:	5.724,849
Adjusted R²:	-248.832.439,511

RAMDOM FOREST

En este caso la segunda cola de ll árbol de decisión lo hemos destinado a entrenar el bosque aleatoria para una regresión usando solo una rama como columna de destino. Predecimos de igual modo los datos test de la regresión.



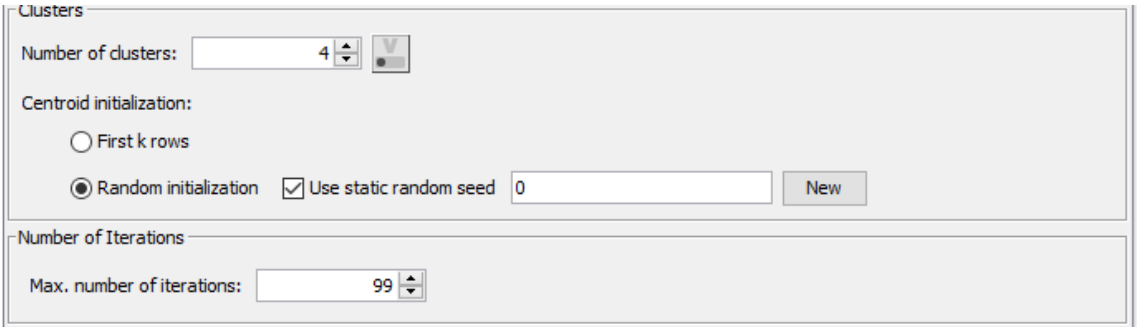
R²:	-3,085
Mean absolute error:	189.572,567
Mean squared error:	47.587.435.782,039
Root mean squared error:	218.145,446
Mean signed difference:	-189.572,567
Mean absolute percentage error:	1
Adjusted R²:	-3,085

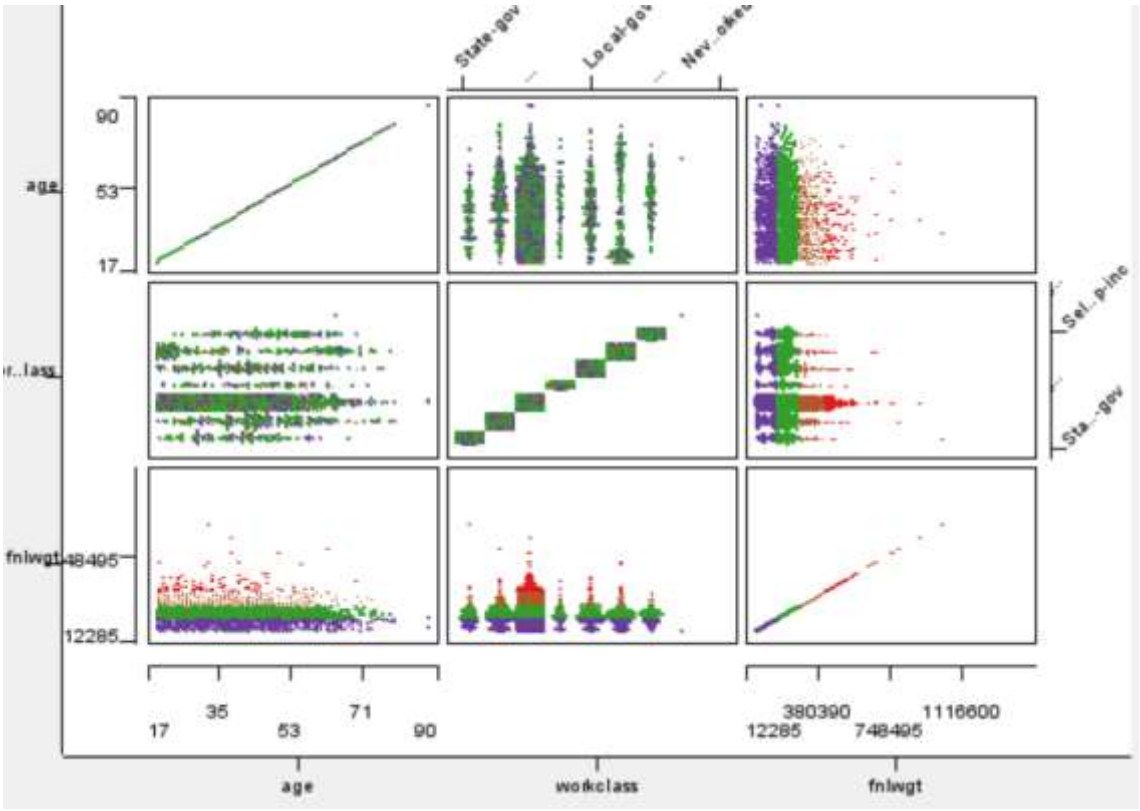
Vemos como el coeficiente de determinación se ha reducido, por lo tanto, el modelo un mejor ajuste, aunque sigue siendo negativo para mostrarnos que las variables elegidas son inversas.

CLUSTERING

En esta técnica de marchine learning hemos pretendido utilizar el análisis estadístico basado en le agrupamiento de ítems en grupos que tengan características similares, conocidos como clúster.

Una vez que hemos elegido el número de grupos, k , procedemos a escoger los k centroides del espacio de datos aleatoriamente. Cada objeto de datos es asignado a un centroide. Y por ultimo se reasigna la posición del centroide tomando un nuevo centroide basado en el promedio de los objetos pertenecientes al grupo. Repitiendo el proceso hasta que estos centroides no se mueven. Hemos elegido 4 clúster para nuestro modelo y un número máximo de 99 interacciones.

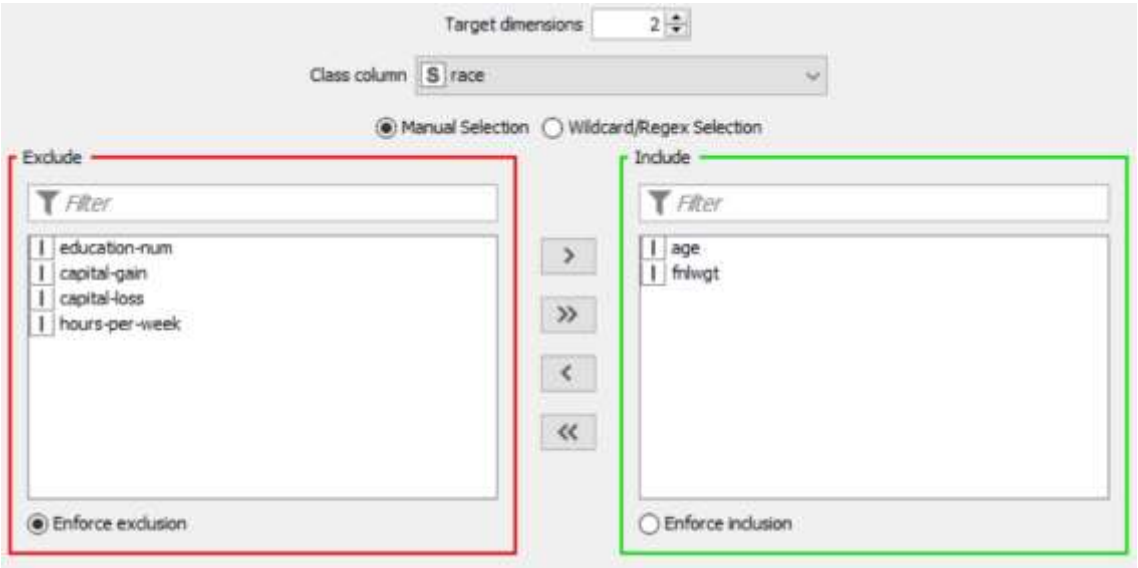




Para la preparación de los gráficos hemos usado nodos para seleccionar gamas de colores de los clústers y la formas de ellos con otro nodo. En los gráficos se pueden ver las distintas iteraciones de los clústers diferenciados por colores y el agrupamiento de ellos por homogeneidad.

ANÁLISIS DISCRIMINANTE

Para el análisis discriminante lineal hemos usado dos salidas dimensionales, incluyendo la edad y el salario final con el objetivo de determinar con criterio a que grupo pertenece un individuo a parit de la información que disponemos.



En el gráfico hemos querido diferenciar los salarias de nuevo en mayor y menores de 50 mil para ver a qé grupo pertenecen según la edad y el salario al final de año. Filtrando los datos discriminantes podemos observar más claramente y con más precisión a qué grupo pueden pertenecer los individuos.