



PRUEBA DE EVALUACIÓN 1

KNIME

SALARIOS EN ESTADOS UNIDOS



UAH

14 DE FEBRERO DE 2024

ALFREDO SALVADOR TÉLLEZ

UNIVERSIDAD ALCALÁ DE HENARES

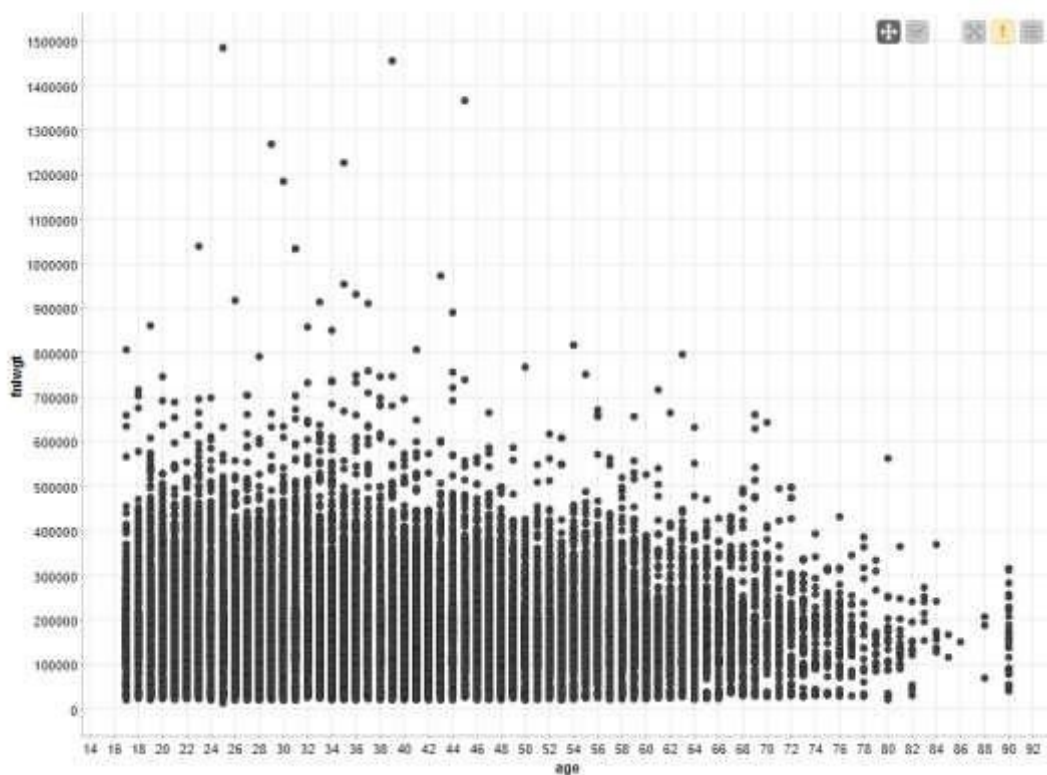
EXPLORACIÓN

Para estudiar los salarios de Estados Unidos nos hemos descargado una base de datos de Kaggle donde hay variables que son muy interesantes para el estudio, como puede ser `target`, donde nos reflejan si el salario está por encima o por debajo de 50.000 dólares anuales. Esta variable va a ser la más usada en nuestro proyecto.

Existen otras variables como el sexo o la raza que también son de interés y podemos correlacionar con la principal, `target`.

Cargando el dataframe podemos observar la distribución de los datos analizando el salario final anual a lo largo del tiempo. Observando que los años de madurez laboral intermedia es donde más se cobran y va descendiendo al pasar los años donde al final existe un estancamiento por la pensión de jubilación.

Existen datos atípicos entre los 25 y 40 años al superar la media como hemos comentado anteriormente.



Utilizamos el nodo `CrossTab` para visualizar las relaciones entre variables. Como la variable `"target"` distribuye los datos en `<` o `>` de 50.000 dólares analizando el cruce de datos con el sexo.

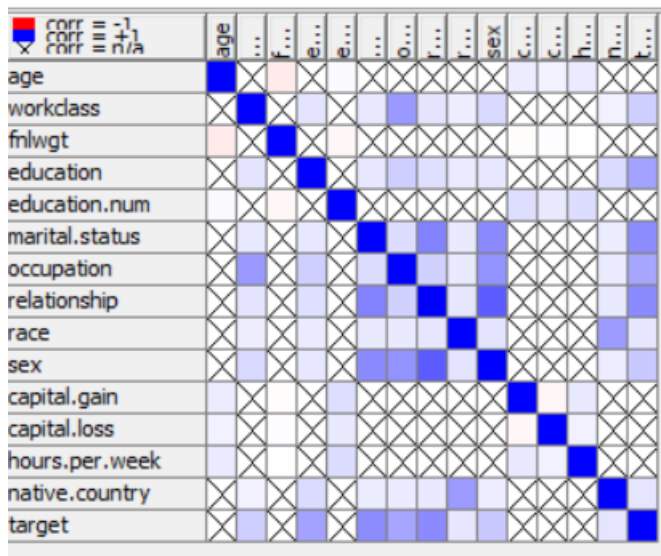
Podemos observar como existen más mujeres que hombres que cobran menos de 50mil euros. También podemos deducir que la mayoría de los salarios en Estados Unidos son más bajos de los 50.000 dólares netos.

Cross Tabulation of sex by target

Frequency Row Percent	<=50K.	>50K.	Total
Female	4.831 89,1164%	590 10,8836%	5.421
Male	7.604 70,0184%	3.256 29,9816%	10.860
Total	12.435	3.846	16.281

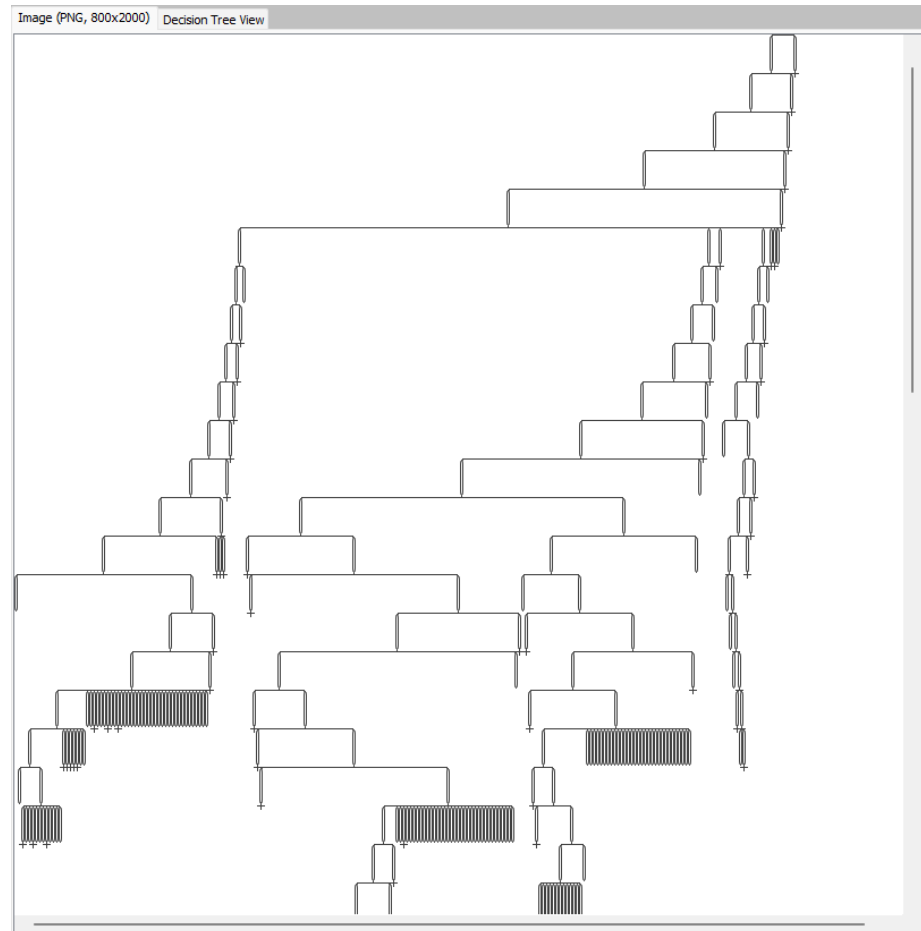
Otro dato para analizar previamente es la correlación de las variables. Decidimos estudiar todas, aunque realmente solo nos interesan dos o tres de ellas.

La variable “target” está notablemente más correlacionada con otras tres variables; ocupación, estado civil y relaciones. Es decir, el salario está vinculado al puesto de trabajo y un factor importante es el estado civil, donde una persona casada puede recibir un salario mayor 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.



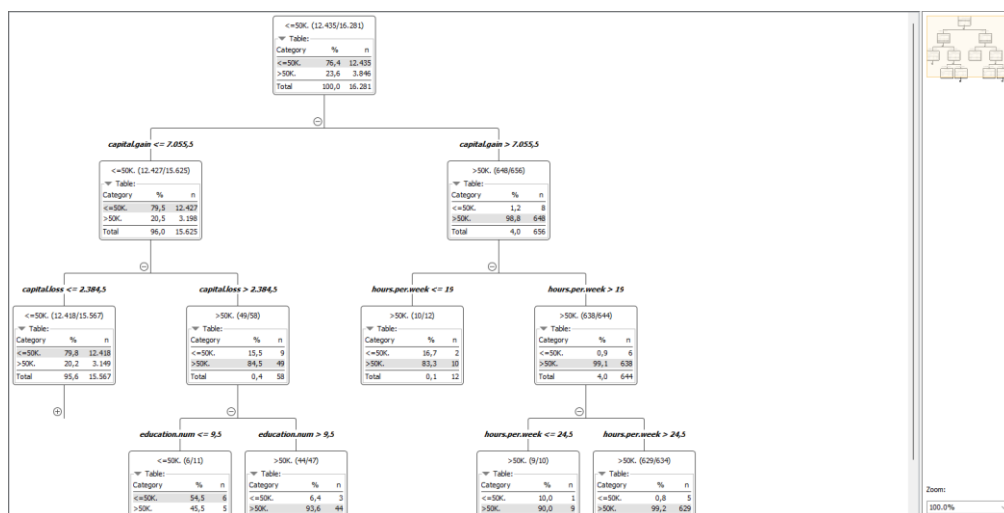
TRAINING

Para datos training introduciremos un árbol previo sin poda para analizar los datos de forma esquematizada y no analizarlos todavía. Nos fijamos en los nodos del árbol de decisión del gráfico de abajo para observar las ramificaciones.



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 tiene unas ganancias superiores a 7.055,5 y trabaja más de 19 horas por semana, tiene una probabilidad del 98,8% de ganar anualmente más de 50k dólares.



EVALUACIÓN

En este apartado añadiremos una segunda rama a nuestra estructura, los árboles de decisión con poda.

Para no caer en sesgos hemos decidido no aumentar la complejidad del árbol, por lo tanto, la simplicidad de árbol es lo que nos permite ser más efectivos y no caer en sobreajustes en los datos.

La idea es obtener un árbol específico que mediante la poda nos aporte nueva información al modelo. Creando un árbol complejo e ir podando los niveles suficientes para que no exista dificultad de predicción. Usamos parte de los datos de entrenamiento para los datos de validación.

SIN PODADO

target \ Pr...	<=50K.	>50K.
<=50K.	11897	538
>50K.	1682	2164

Correct classified: 14.061
Accuracy: 86,364%
Cohen's kappa (κ): 0,579%

PODADO

target \ Pr...	<=50K.	>50K.
<=50K.	12140	295
>50K.	348	3498

Correct classified: 15.638
Accuracy: 96,051%
Cohen's kappa (κ): 0,89%

OVERFITTING

Para evitar el sobreajuste en este apartado intentamos controlar este efecto realizando un participación inicial del 80% test y el 20% para los datos de entrenamiento. Usamos los datos reservados para la evaluación y repetimos el proceso para obtener un modelo más realista que los anteriores.

Dialog - 11:5 - Partitioning (Train/test split)

File

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

Choose size of first partition

☐ Absolute 100

☒ Relative[%] 80

☐ Take from top

☐ Linear sampling

☒ Draw randomly

☐ Stratified sampling S target

☐ Use random seed 1.707.843.577.4

OK Apply Cancel ?

RANDOM 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. Es útil para comparar con los datos de la validación cruzada. Una diferencia muy grande puede suponer que la CV no ha seleccionado correctamente los datos test.

Scorer View

Confusion Matrix



	<=50K. (Predicted)	>50K. (Predicted)	
<=50K. (Actual)	9428	494	95.02%
>50K. (Actual)	1280	1822	58.74%
	88.05%	78.67%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
86.38%	13.62%	0.589	11250	1774

Scorer View

Confusion Matrix



	<=50K. (Predicted)	>50K. (Predicted)	
<=50K. (Actual)	2402	111	95.58%
>50K. (Actual)	310	434	58.33%
	88.57%	79.63%	

Overall Statistics




Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
87.07%	12.93%	0.595	2836	421

CLUSTERING

En esta técnica de análisis no supervisado agrupamos en ítems o núcleos dependiendo de las similitudes de las características de estos mientras que los objetos son distintos.

Una vez que hemos elegido el número de grupos, k como aconsejaba Cipola usaremos 4 y procederemos a elegir los k centroides de manera aleatoria en los datos, donde cada objeto de datos es asignado a un centroide. Finalmente, se recalcula la posición del centroide mediante la determinación de un nuevo centroide basado en el promedio de los objetos pertenecientes al grupo. Este proceso se repite hasta que los centroides dejen de moverse. Para nuestro modelo, hemos seleccionado 4 clústeres y un número máximo de 99 iteraciones.

Clusters


Number of clusters:   

Centroid initialization:

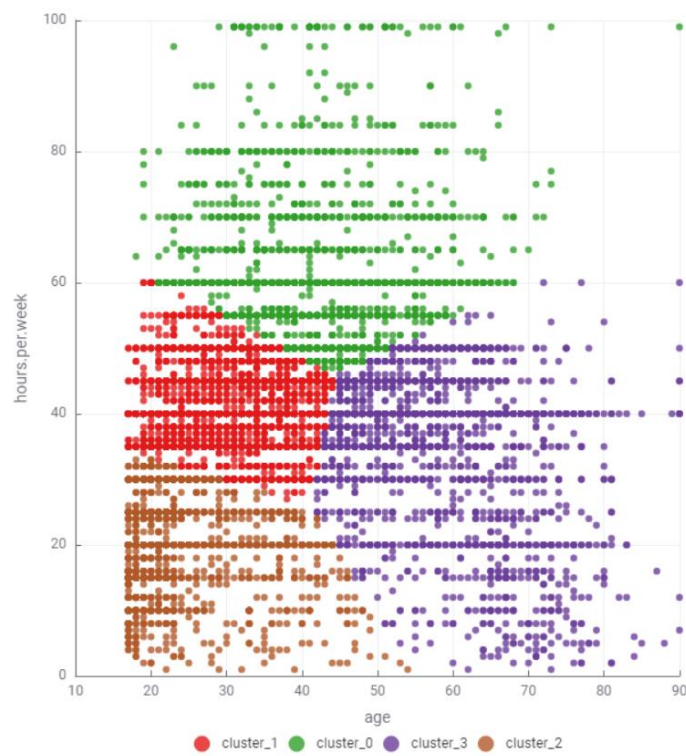
☐ First k rows

☒ Random initialization ☒ Use static random seed

Number of Iterations

Max. number of iterations:  

Scatter Plot Matrix



Data

Dimensions

Manual Wildcard Regex Type

Search Aa

Excludes

workclass
fnlwgt
education
education.n...
marital.stat...
occupation

Any unknown columns

Includes

age
hours.per.week

Color dimension

Cluster

Max rows

250000

Plot

Title

Scatter Plot Matrix

Cancel Ok

Podemos observar por tonos de colores los distintos clúster de las variables “horas por semana” dependiendo de la edad que tenga el individuo.