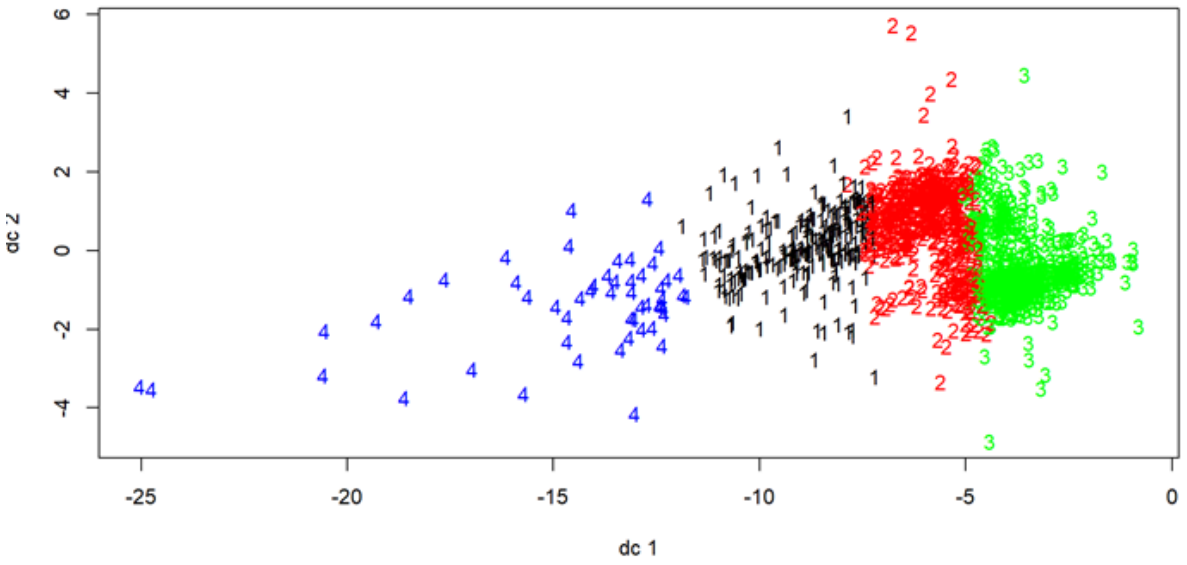
Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
CC3074 – Minería de Datos
Semestre I – 2022
Andrés Paiz 191142
René Ventura
Eduardo Ramirez 19946

Hoja de trabajo 3

1. Descargue los conjuntos de datos de la plataforma kaggle.

2. Incluya un análisis de grupos en el análisis exploratorio. Explique las características de los grupos.

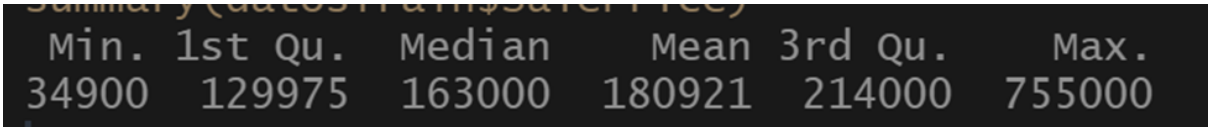


g1	245 obs. of 82 variables
g2	527 obs. of 82 variables
g3	635 obs. of 82 variables
g4	53 obs. of 82 variables

Se utilizó el método de K means debido a que este conjunto de 4 clusters podía ser fácilmente observable desde K means. El grupo 3 contiene la mayor cantidad de

observaciones con 635 y el grupo 4 el menor con solo 53. Siendo el grupo 4 el que contiene las casas más caras y en el grupo 3 las más baratas.

3. Dependiendo del análisis exploratorio elaborado cree una variable respuesta que le permita clasificar las casas en Económicas, Intermedias o Caras. Los límites de estas clases deben
tener un fundamento en la distribución de los datos de precios, y estar bien explicados.



Clase	Rango
Económicas	34,900-152,000
Intermedias	153,000-229,000
Caras	229,426- 755,00

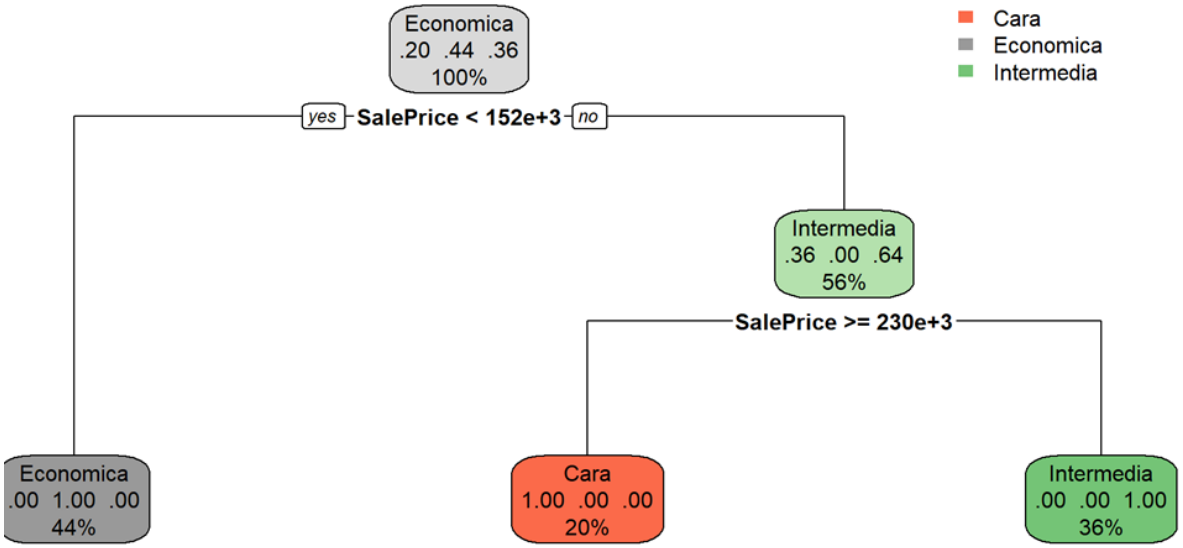
Para la variable respuesta se realizó una clasificación de económicas, intermedias y caras. Para las económicas se tomó únicamente el grupo 3 del cluster, para las intermedias se tomó únicamente el grupo 2 y para las caras se unieron los grupos 1 y 4.

4. Divida el set de datos preprocesados en dos conjuntos: Entrenamiento y prueba. Describa el criterio que usó para crear los conjuntos: número de filas de cada uno, estratificado o no, balanceado o no, etc. Si le proveen un conjunto de datos de prueba y tiene suficientes datos, tómelo como de validación, pero haga sus propios conjuntos de prueba.

Se tomarán como conjuntos hipotéticos el 70% del dataset train para entrenamiento y el otro 30% del mismo dataset como prueba.

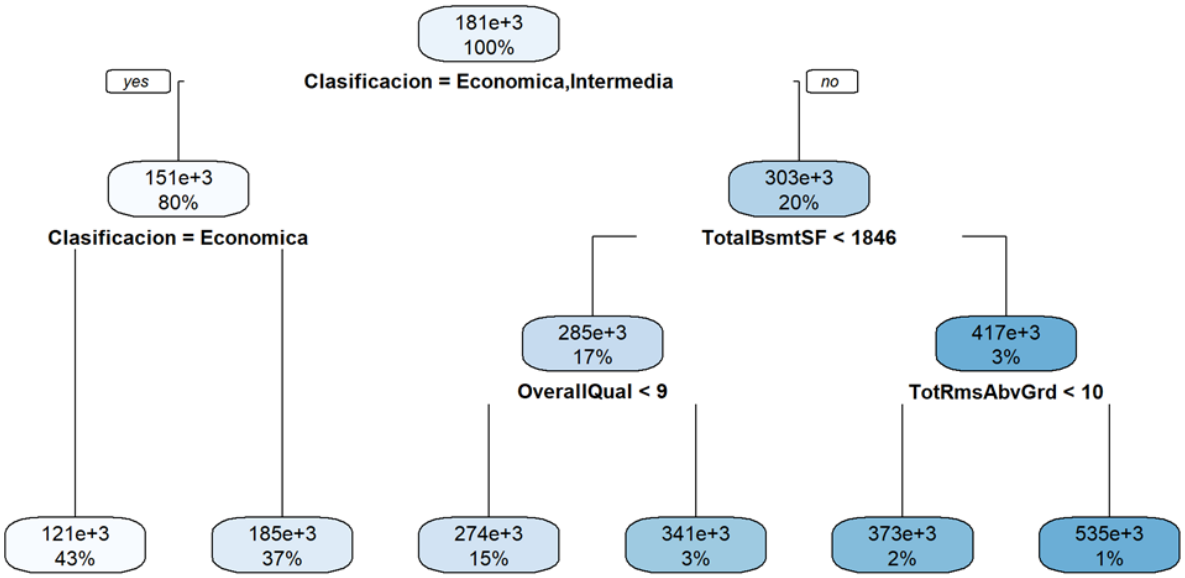
5. Elabore el árbol de clasificación utilizando el conjunto de entrenamiento y la variable respuesta que creó en el punto 4. Explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser reproducible

por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.



El árbol de clasificación realizó el análisis gráfico de partición con la variable respuesta que se indicó hipotéticamente en el ejercicio 4.

6. Elabore el árbol de regresión para predecir el precio de las viviendas utilizando el conjunto de entrenamiento. Explique los resultados a los que llega. Muestre el modelo gráficamente. El experimento debe ser reproducible por lo que debe
fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.



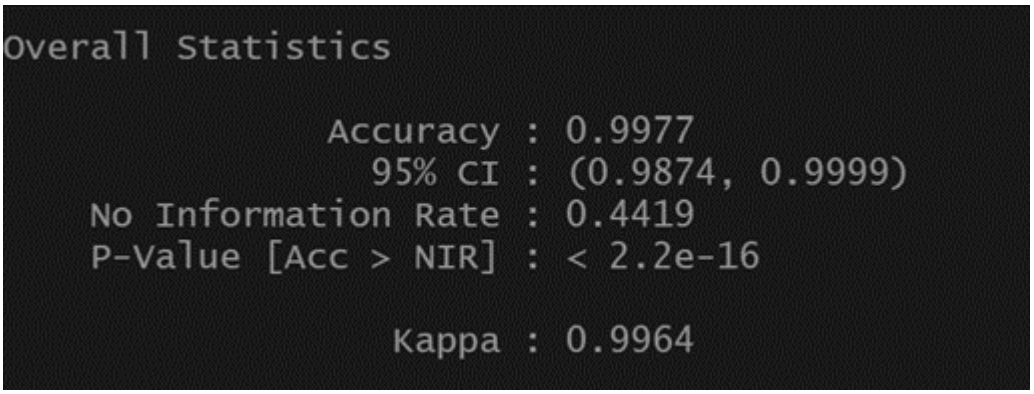
En el lado derecho del gráfico podemos observar los datos con un overall quality mayor.Igualemente un conjunto de solo 1% de la muestra. Al mismo tiempo del lado izquierdo se encuentran las casas con las calidades más bajas, con un precio disminuido.

7. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar y predecir, en dependencia de las características de la
variable respuesta.

	Reference		
Prediction	Cara	Economica	Intermedia
Cara	88	0	0
Economica	0	194	0
Intermedia	1	0	156

Se observó que el algoritmo fue efectivo a la hora de predecir o clasificar la variable de respuesta en nuestro conjunto de prueba. Se puede observar que solamente ocurrió un error en la predicción de una casa que era intermedia.

8. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión para el árbol de clasificación. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores.

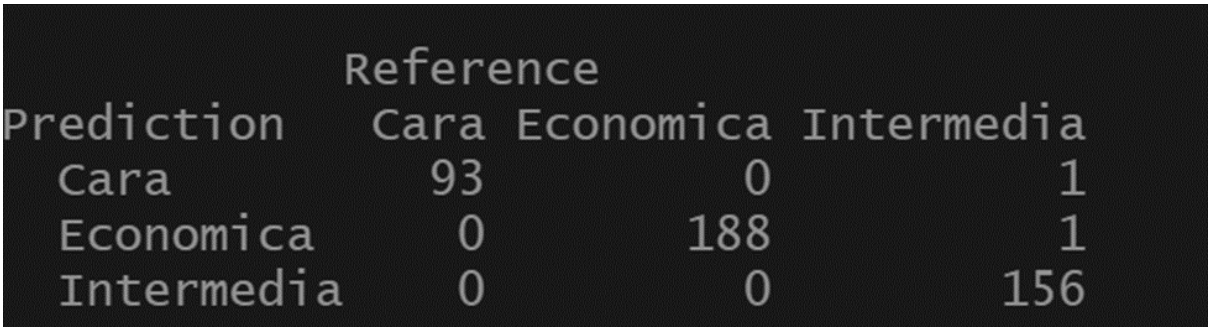


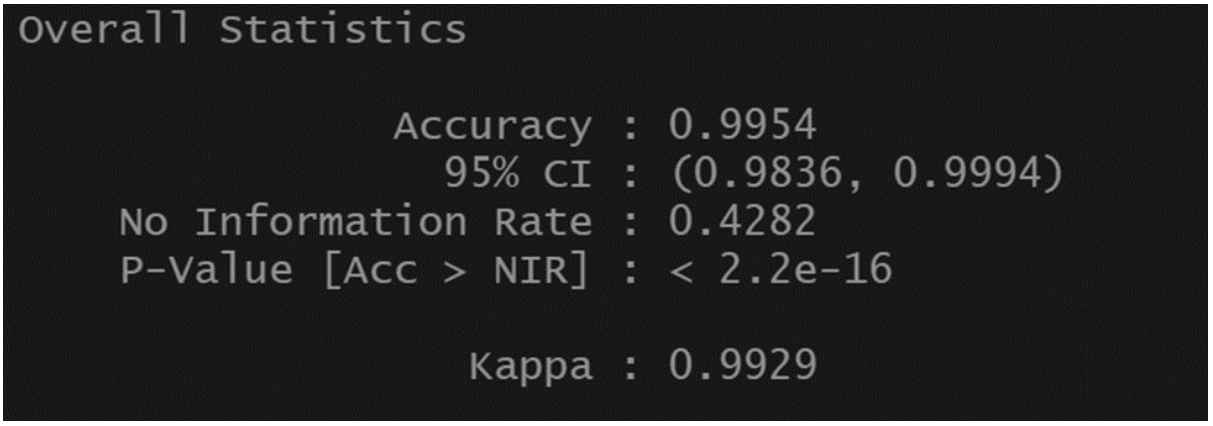
El algoritmo cometió un error en la clasificación de una casa que realmente era intermedia y no cara. El error observado nos permite observar como una casa puede compartir características con otras casas que pueden no compartir el mismo precio. El error nos puede hacer asumir que para crear nuestra variable de respuesta de clasificación, se podría basar en otras variables del dataset que no sean sale price, lo cual permitiría una mejor clasificación.

9. Analice el desempeño del árbol de regresión.

Se observó que el árbol de represión tuvo múltiples problemas al intentar predecir resultados de sale price. Igualmente se observó que resultaba aún menos efectivo a la hora de utilizar valores numéricos únicos. Comparándolo con el árbol de clasificación, se considera menos efectivo.

10. Repita los análisis usando random forest como algoritmo de predicción, explique sus resultados comparando ambos algoritmos.





Se puede observar que en el conjunto de prueba la mayor parte fueron clasificados correctamente. Se cometió un error al clasificar una económica cuando en realidad era una intermedia, lo mismo sucede con una intermedia que era cara y una intermedia que era económica.