NP1602 – Introducción al Análisis Multivariado

Proyecto Final:
Árboles de Decisiones

Presentado por:

Yanina Araya P.

Renato Guadamuz F.

Julio 2017

Contenidos

- Caso en estudio
- Objetivo y Justificación
- Análisis descriptivo
- Resultados
  - Classification and Regression Tree (CART)
  - Inferencia condicional
- Conclusiones

Caso en estudio

- Ingreso neto por trabajo
  - ENAHO 2016
  - Personas con trabajo remunerado
  - Mayores de 15 años
- $n = 794$ personas
- Características de las personas:
  - Educación
  - Sociales
  - Demográficas

Objetivo

- Conocer los ingresos y las características de la población que los perciben con el fin de dar el apoyo preciso a los diferentes sectores de la población de acuerdo con sus necesidades.

Justificación

- Analizar la población por sectores para estudiar sus ingresos y por lo tanto intuir sus gastos.
- Estudiar la desigualdad de género a nivel de ingreso en Costa Rica.
- Concientizar a la población y el Estado de las ventajas de una población con educación de calidad.

Variables

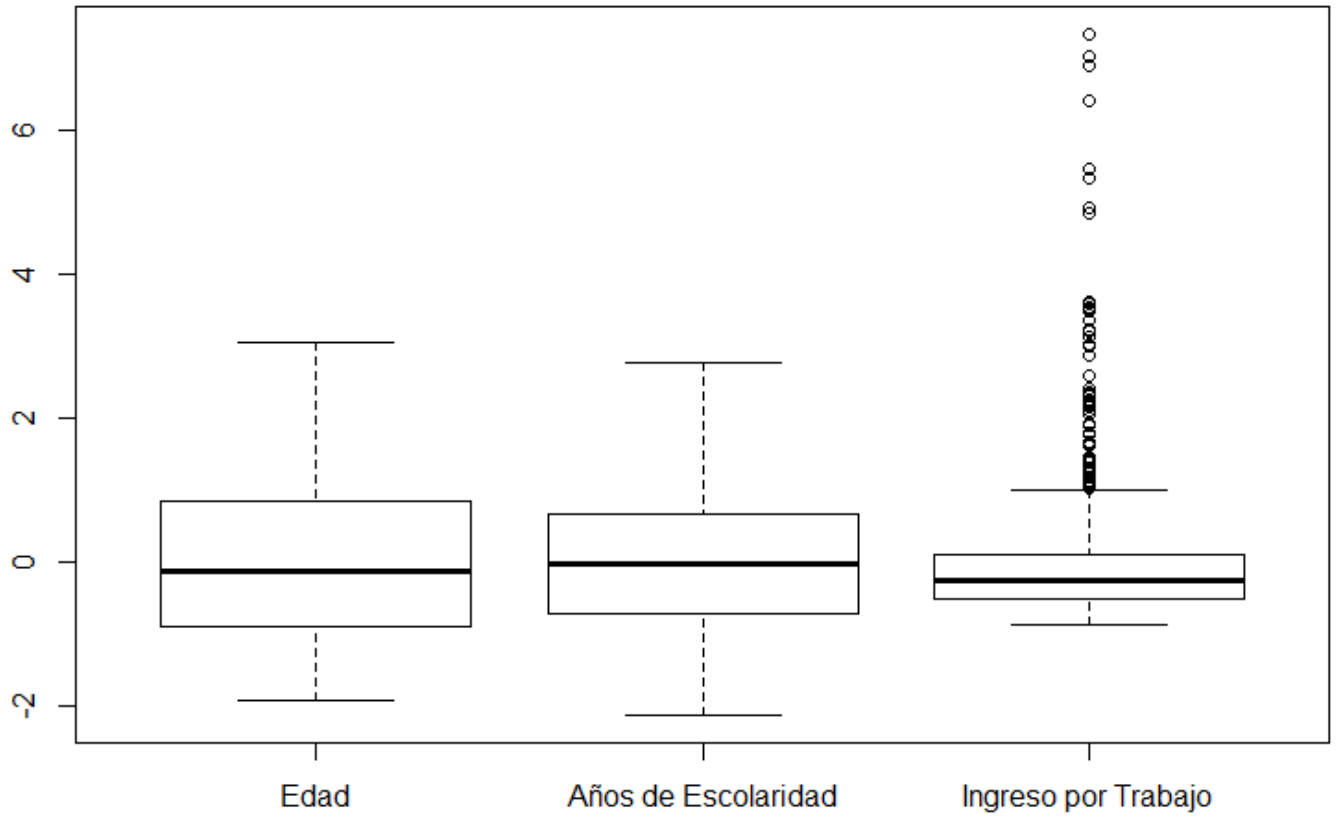
- Continuas:
  - Edad
  - Años de escolaridad
  - **Ingreso neto por trabajo**
- Categóricas:
  - Región
  - Zona
  - Sexo
  - Migrante
  - Nivel de educación
  - Estado Civil

Metodología

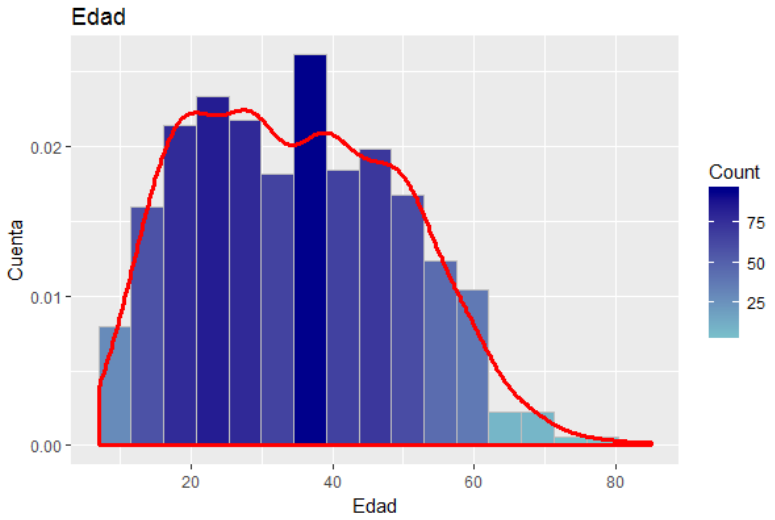
1. Análisis descriptivo
2. Estimación del modelo
   - Predicción (ingreso por trabajo)
3. Poda o restricciones
4. Predicción de variables
5. Interpretación

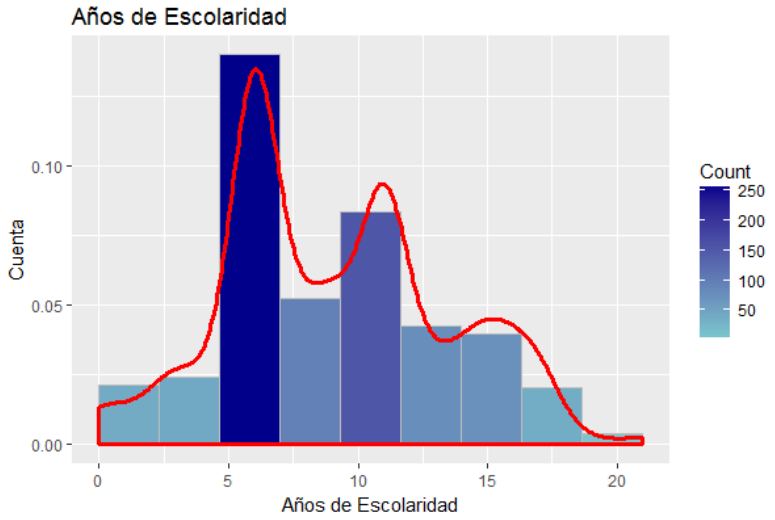
Análisis descriptivos

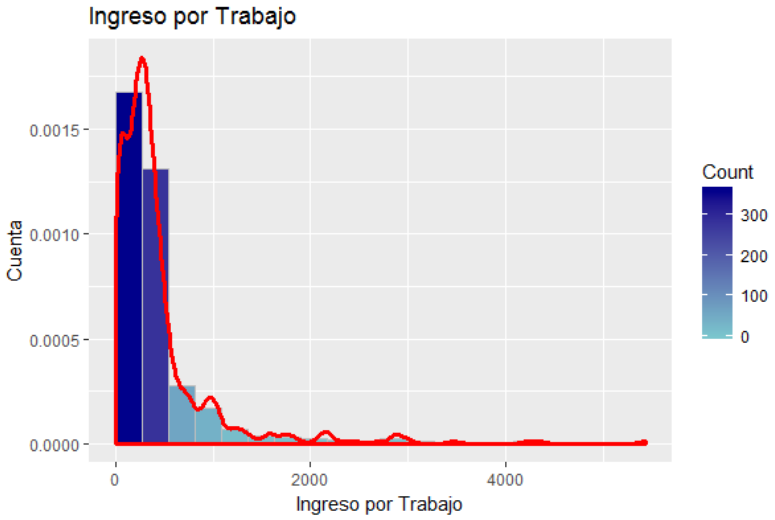
Boxplot

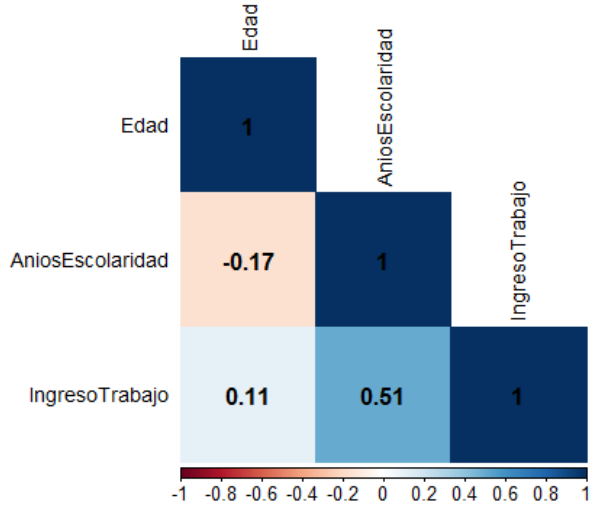


Variables continuas

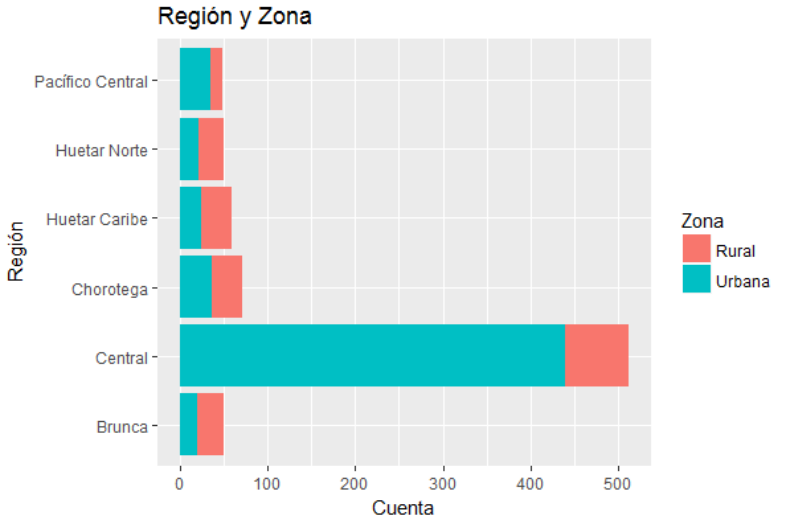


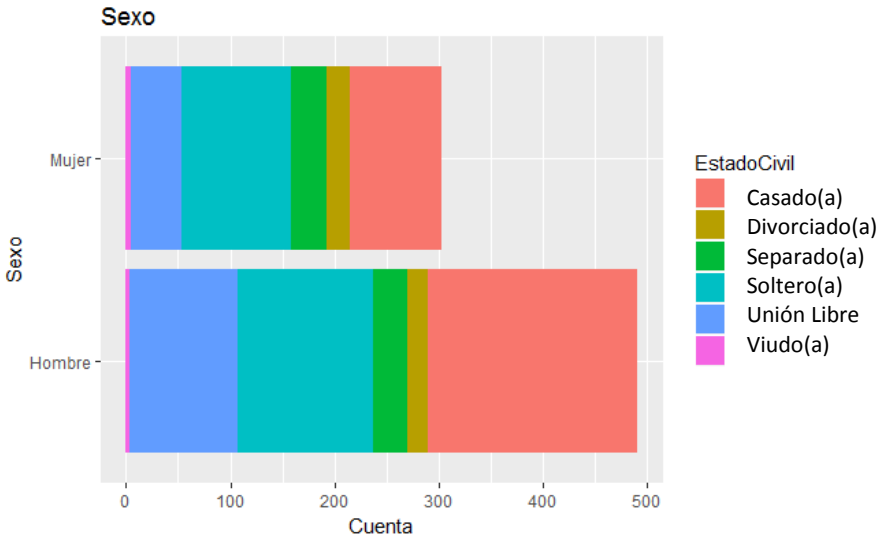






Variables categóricas





Nivel de Educación

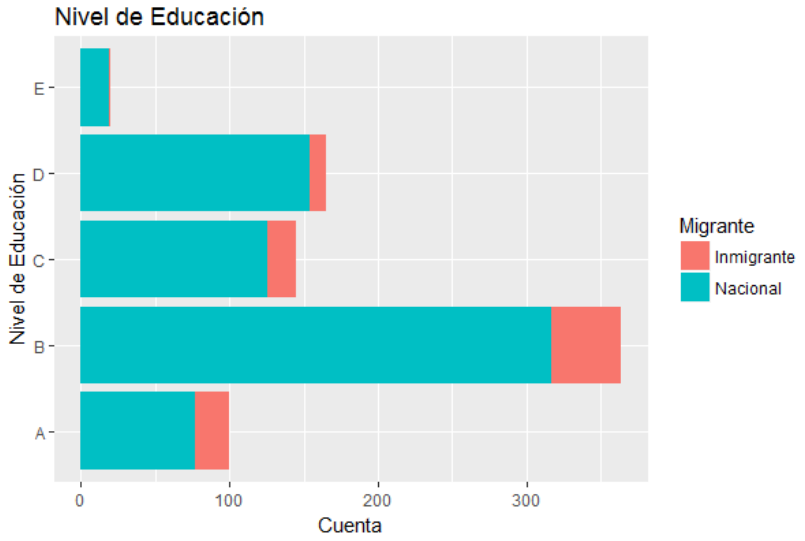
A: Sin educación formal o primaria incompleta

B: Primaria complete o secundaria incompleta

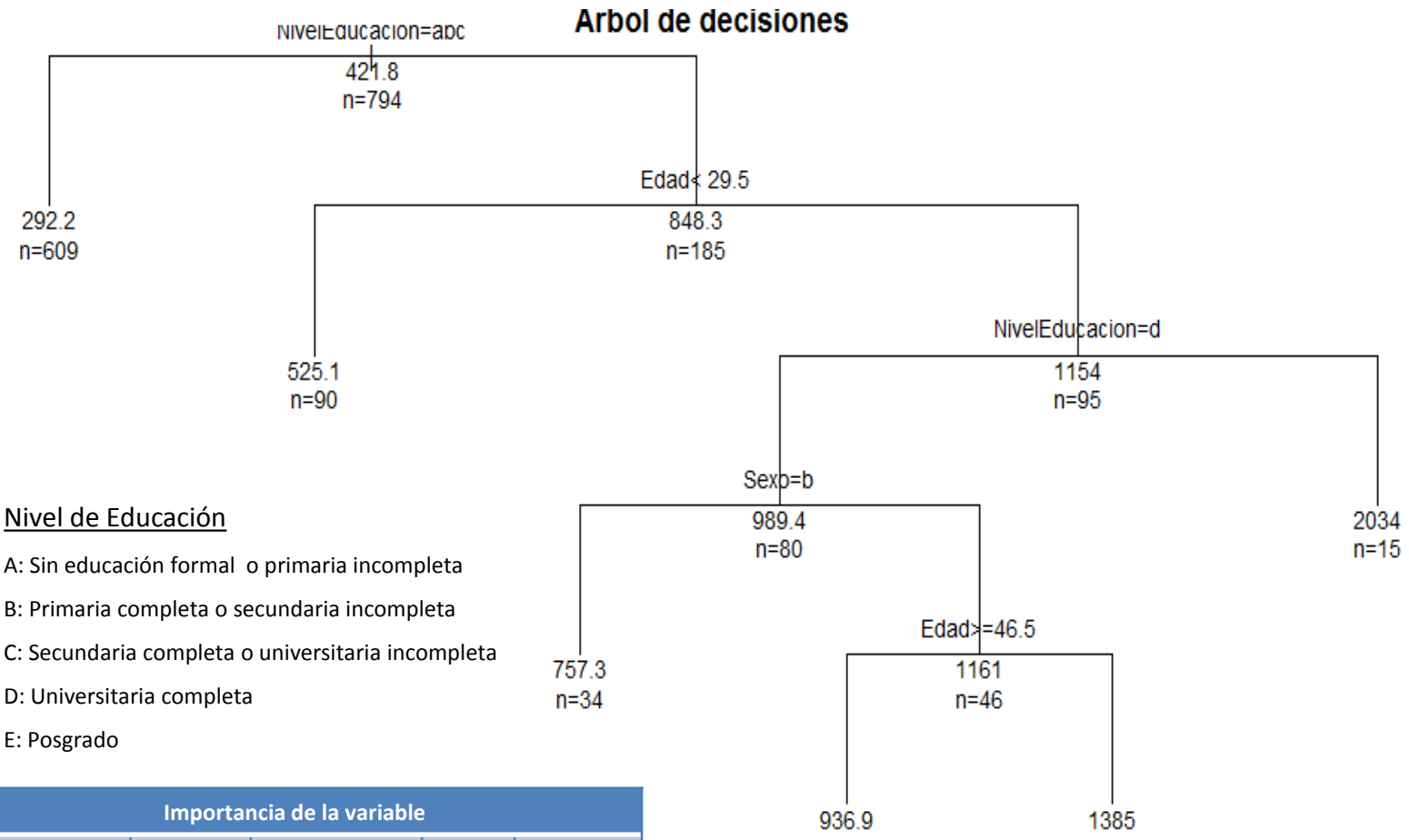
C: Secundaria complete o universitaria incompleta

D: Universitaria completa

E: Posgrado



Árbol de decisiones (CART)



Nivel de Educación

A: Sin educación formal o primaria incompleta

B: Primaria completa o secundaria incompleta

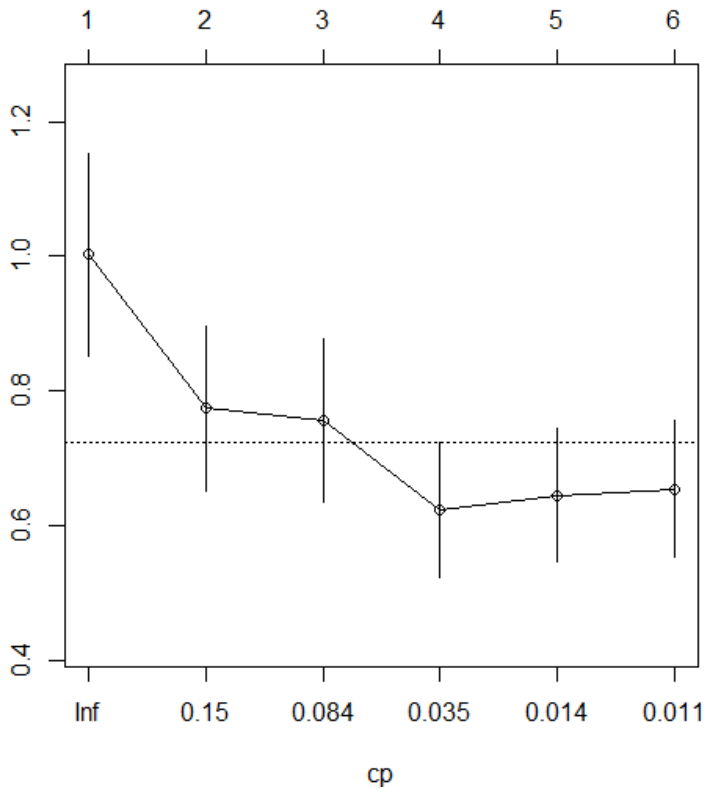
C: Secundaria completa o universitaria incompleta

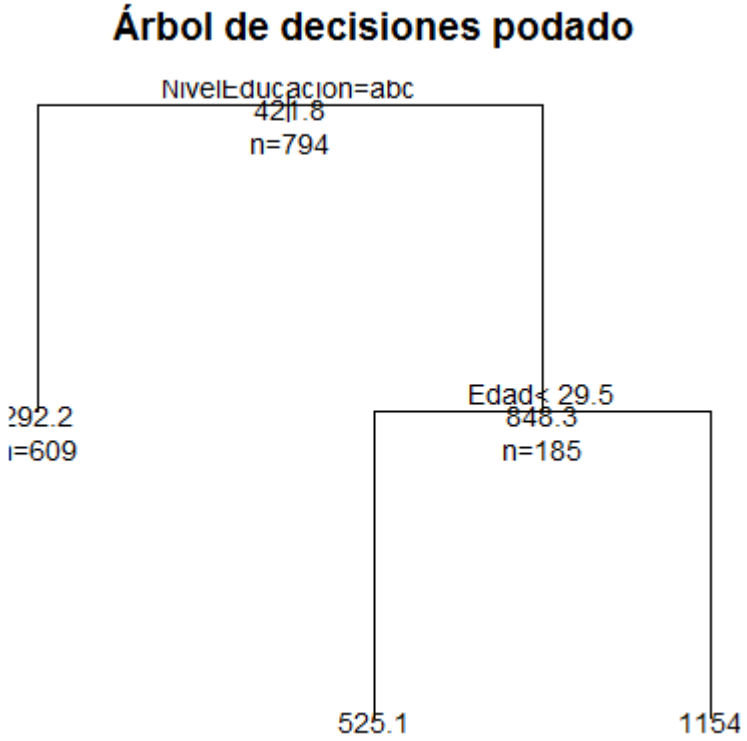
D: Universitaria completa

E: Posgrado

Importancia de la variable				
Nivel Educación	**Edad**	**Estado civil**	**Sexo**	**Región**
60	21	10	6	2

AD-CART Podado





Nivel de Educación

A: Sin educación formal o primaria incompleta

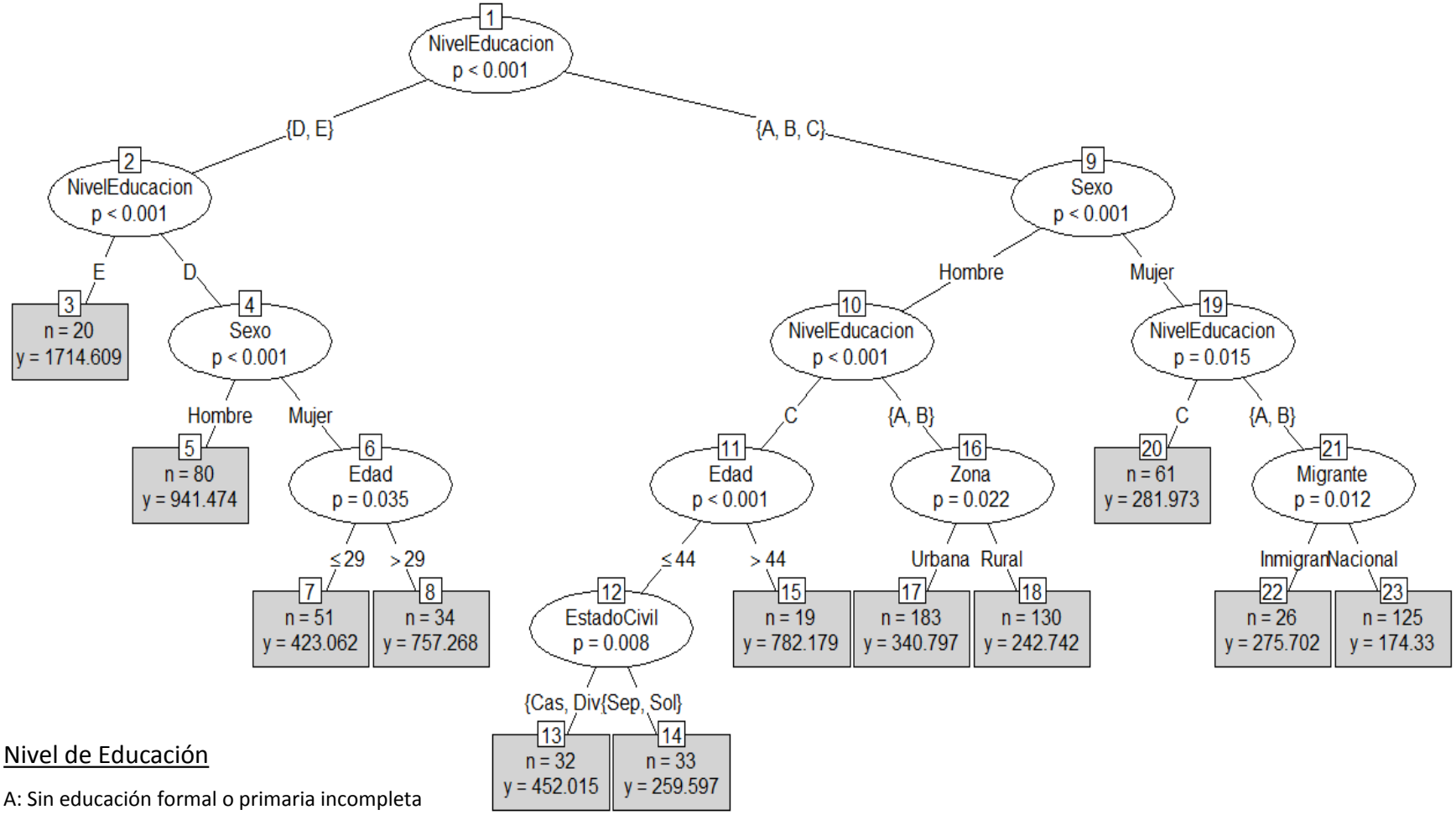
B: Primaria completa o secundaria incompleta

C: Secundaria completa o universitaria incompleta

D: Universitaria completa

E: Posgrado

Árbol de decisiones (inferencia condicional)



Nivel de Educación

A: Sin educación formal o primaria incompleta

B: Primaria completa o secundaria incompleta

C: Secundaria completa o universitaria incompleta

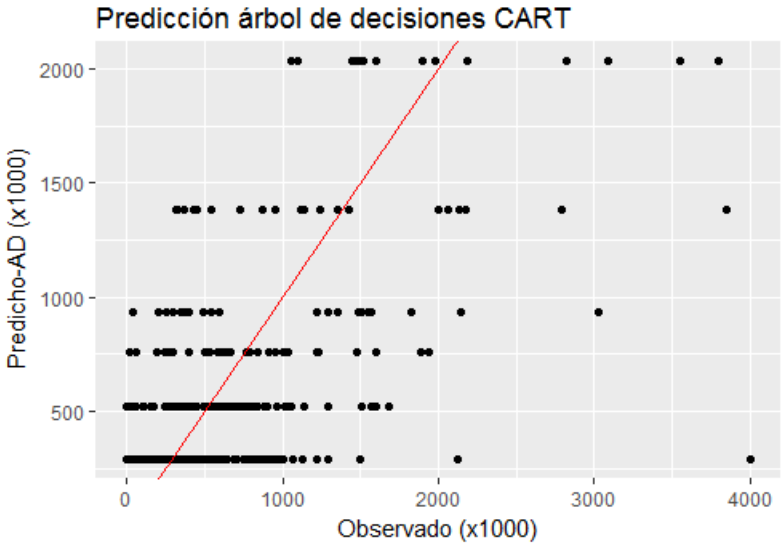
D: Universitaria completa

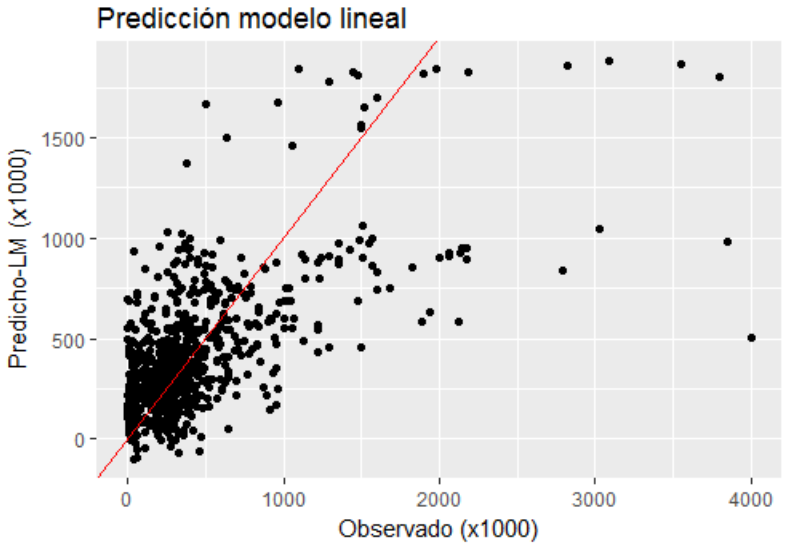
E: Posgrado

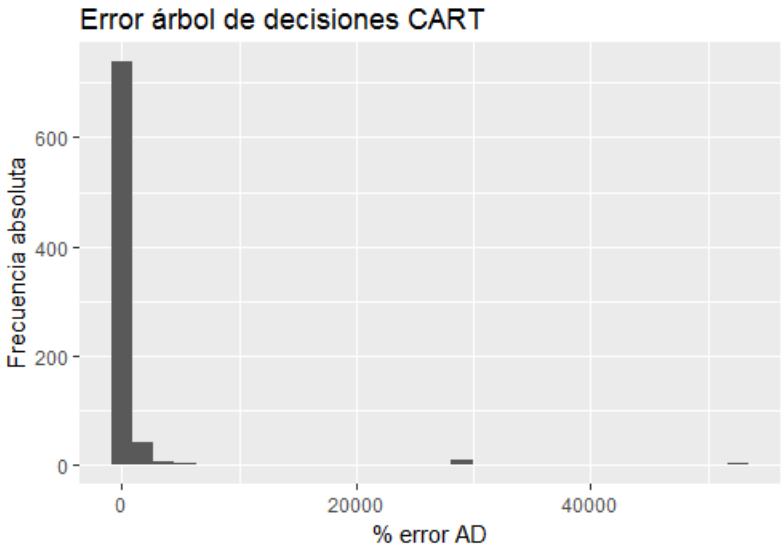
Árbol de decisiones (AD)

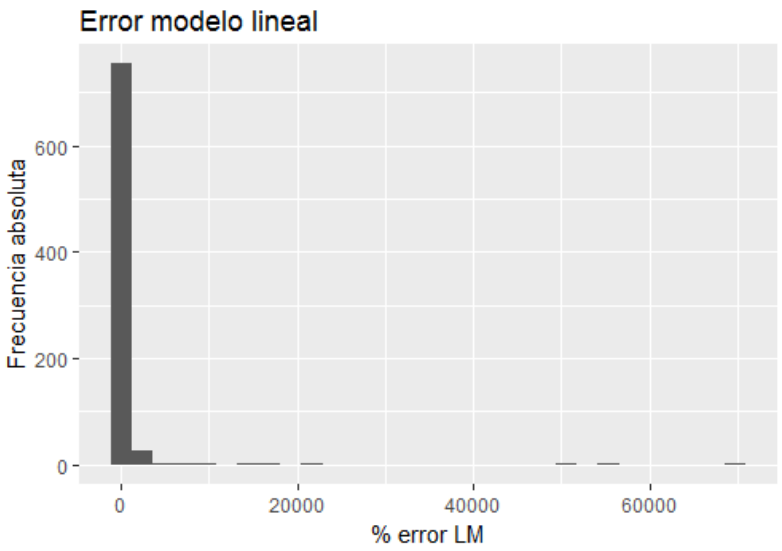
Id	Predicción AD - CART	Predicción modelo lineal	Valor real observado	Diferencia AD - CART	Diferencia ML
1	292.4	292.4	692.9	-58%	-58%
1	292.4	435.3	80.0	262%	438%
3	292.4	313.2	277.1	5%	13%
4	292.4	71.4	4.0	5764%	1349%
5	292.4	64.5	294.6	-1%	-78%
⋮	⋮	⋮	⋮	⋮	⋮
Promedio	**421.8**	**306.8**	**421.8**	**577%**	**490%**
MSE	**213450**	**222560**			

Errores por ambos métodos (%)









Conclusiones

- Las variables de nivel de educación, edad y sexo son las principals variables explicativas.
- Otras variables relevantes son la zona, el estado civil y la condición de inmigrante.
- La variable región, no se encontró significativa con ningún método.
- Los resultados obtenidos tienen sentido, respecto a los ingresos esperados dadas ciertas condiciones.

Conclusiones (cont.)

- Los ingresos aumentan conforme se tienen mayores niveles de educación, mayor edad y para sexo masculino.
- Los ingresos más bajos se registran para mujeres nacionales y con baja escolaridad.
- Los ingresos más altos se dan para personas con estudios de posgrado.
- Resultados de las variables son consistentes con otras investigaciones.
- No es evidente cuál método presenta los mejores resultados, sin embargo, basado en el MSE se prefieren los árboles de decisión.