

UNIVERSIDAD DE COSTA RICA
FACULTAD DE CIENCIAS ECONÓMICAS
ESCUELA DE ESTADÍSTICA
PROYECTO FINAL:
ANÁLISIS DE INGRESOS EN COSTA RICA
ESTUDIANTES:
YANINA ARAYA PORRAS
RENATO GUADAMUZ FLORES
SEDE RODRIGO FACIO

JULIO 2017

RESUMEN

El presente artículo estima el ingreso promedio de los trabajadores en Costa Rica de acuerdo con características demográficas, sociales, académicas, entre otras, tomadas de la Encuesta Nacional de Hogares (ENAHOG) 2016 generada por el Instituto Nacional de Estadística y Censos (INEC) por medio de la metodología multivariada de árboles de decisión. Los árboles de decisión son un modelo de clasificación o regresión utilizado en la inteligencia artificial, cuya principal característica es su fácil interpretación visual. Se busca determinar la eficacia de los árboles de decisión, es decir, si pueden ser una herramienta de apoyo para la estimación promedio del ingreso de los costarricenses.

Palabras clave: árboles de decisión, ingreso, estimación, Costa Rica.

ÍNDICE

| | |
|---|----|
| 1. INTRODUCCIÓN..... | 1 |
| 2. MARCO CONCEPTUAL..... | 2 |
| 3. MARCO ACTUAL DE LA CUESTIÓN..... | 4 |
| 4. METODOLOGÍA..... | 5 |
| 4.1 Población de estudio..... | 5 |
| 4.2 Marco muestral de viviendas..... | 6 |
| 4.3 Dominios de estudio..... | 6 |
| 4.4 Estratificación del marco muestral para selección..... | 6 |
| 4.4 Diseño de muestreo..... | 6 |
| 4.5 Tamaño distribución y selección de la muestra..... | 7 |
| 5. RESULTADOS..... | 7 |
| 5.1 Análisis descriptivo de los datos..... | 7 |
| 5.2 Árbol de decisiones (CART)..... | 11 |
| 5.3 Árbol de decisión (inferencia condicional)..... | 12 |
| 5.4 Comparación entre árboles de decisión y regresión lineal..... | 13 |
| 6. CONCLUSIÓN Y DISCUSIONES..... | 15 |
| 7. BIBLIOGRAFÍA..... | 16 |

TABLA DE CUADROS

| | |
|--|----|
| Cuadro 1. comparación de resultados según metodología..... | 14 |
|--|----|

TABLA DE GRÁFICOS

| | |
|--|----|
| Gráfico 1. Gráficos de cajas para las variables continuas..... | 8 |
| Gráfico 2. Histogramas (a) edad, (b) años de escolaridad, (c) Ingreso..... | 8 |
| Gráfico 3. Análisis de correlaciones entre variables continuas | 9 |
| Gráfico 4. Frecuencia por región y zona..... | 9 |
| Gráfico 5. Frecuencia por sexo y estado civil..... | 10 |
| Gráfico 6. Frecuencia en migrantes y nivel de educación | 10 |
| Gráfico 7. Árbol de decisiones (CART) | 11 |
| Gráfico 8. Árbol de decisión (inferencia condicional)..... | 12 |
| Gráfico 9. Predicción y error según modelo CART | 14 |
| Gráfico 10. Predicción y error según modelo lineal | 15 |

TABLA DE ABREVIATURAS

| | |
|-------|---|
| ENAH | Encuesta Nacional de Hogares |
| ENIGH | Encuesta Nacional de Ingresos y Gastos de Hogares |
| INEC | Instituto Nacional de Estadísticas y Censos |
| OIT | Organización Internacional de Trabajo |
| UPM | Unidades Primarias de Muestreo |
| USM | Unidades Secundarias de Muestreo |

1. INTRODUCCIÓN

En Costa Rica, el Instituto Nacional de Estadísticas y Censos (INEC) es el ente encargado de gestionar las estadísticas nacionales para orientar las decisiones que promuevan el desarrollo del país, bajo esta función, es el encargado de realizar la Encuesta Nacional de Hogares cada año con el fin de recolectar datos cuyo enfoque está asociado a al nivel de bienestar de la población especialmente centrados en la conformación del ingreso de los hogares, su distribución y características de los hogares y la población en situación de pobreza (Instituto Nacional de Estadística y Censos (INEC), 2015).

La Encuesta se realiza en julio de cada año, y permite obtener información a nivel nacional, regional y para la zona urbana y rural.

Las Encuestas de Hogares forman parte de un programa permanente que inició en 1976. Desde ese año se han realizado dos actualizaciones metodológicas y conceptuales.

Cada nación es responsable de brindar calidad de vida a sus ciudadanos, puesto que el Estado es la organización social, política, y económica, conformada por un conjunto de instituciones, que tienen el poder de regular la vida en sociedad, por lo tanto debe conocer las necesidades de sus habitantes y trabajar constantemente por brindar las condiciones más adecuadas.

Es necesario conocer la distribución de la riqueza y de qué depende dicha distribución para poder brindar oportunidades de crecimiento y orientar la formulación de políticas y planes de desarrollo para el bienestar de la sociedad.

La pobreza es un problema social que afecta a toda la población, puesto que la pobreza trae consigo no sólo la falta de abastecimiento de las necesidades básicas, sino que acarrea problemas de salud, delincuencia, violencia, entre otros males, por lo que es necesario 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 y que además, podrían estar en riesgo social.

Es por esto que, haciendo uso de la Encuesta Nacional de Hogares (ENAH) del año 2016, el presente estudio se centra en analizar las características de los individuos que determinan el ingreso para los costarricenses y poder determinar en qué basan dichas diferencias y qué tan grandes son, tomando en consideración inicial la hipótesis de que las mujeres, personas de baja escolaridad y edades menores son las que presentan los menores ingresos.

2. MARCO CONCEPTUAL

Johnson, Riggs y Downey (1987) en su artículo presentan modelos de regresión lineal múltiple con el fin de comparar modelos para determinar la equidad en salarios en una universidad del oeste de Estados Unidos donde se muestra como variable dependiente el salario y como variables independientes se toman 17 variables que incluyen: años de educación, nivel de educación, años desde que se egresó, salario inicial, años desde el último nombramiento, tipo de nombramiento (9 o 12 meses), pago dividido (de dos o más departamentos), cargo que desempeña (profesor, instructor, asistente, profesor asociado), años en el cargo, titularidad de universidad (académico), años como académico, puntaje de mérito promedio, años en el sistema de mérito, departamento, nivel de grado requerido por el departamento, responsabilidad dentro del departamento, y experiencia previa. Se formaron ocho modelos en donde la variable dependiente es el salario, lo que varía de un modelo a otro son las variables predictoras y si se considera tanto a mujeres como hombres o sólo hombres. Para los modelos analizados, los coeficientes de determinación varían entre 0.44 y 0.82.

Lazar (2004) realiza un modelo donde se toman en cuenta catorce variables de las cuales ocho son categóricas y seis continuas. Se toman características como son la edad, estado civil, educación, ocupación, sexo, país de origen, horas laboradas por semana, entre otras, las cuales definen si la persona ganas más o menos de \$50 000. Se realizó el análisis tomando en consideración seis bases de datos distintas las cuales toman diferente número de variables de acuerdo a criterios de experto, así como a análisis por componentes, a estas bases de datos se les aplicó la metodología de *Support VectorMachine* (SVM) para clasificar el ingreso de las personas de acuerdo a diferentes características y logra explicar el 85% de la varianza, lo mismo obtenido por Dumas (2016) realizando el análisis haciendo uso de *Random Forest*.

Liu et al. (2006) realizan un modelo para predecir el ingreso por medio de una regresión bayesiana, donde lo que en principio se desea probar es si los estudiantes calificados como talentosos o brillantes tienden a obtener ingresos mayores y si el género, nacionalidad, nivel de educación y horas de trabajo también tienen influencia en el ingreso percibido por las personas.

Rodríguez (2012) en su trabajo titulado “Análisis de la estructura salarial de una empresa” presenta criterios que determinan el salario de los trabajadores para una empresa específica y proporciona un modelo que establece que trabajadores reciben una remuneración por encima o por debajo de lo justo, con el fin de establecer criterios unificados para la remuneración. Para

alcanzar dicho objetivo, hace uso de un análisis de regresión múltiple, donde analiza la relación existente entre el salario actual (como variable dependiente) y el salario inicial, los meses de experiencia previa a la contratación, y los meses desde el comienzo del contrato (como variables independientes), donde una vez obtenido el modelo con las variables en cuestión se obtiene explicar un 80.4% de la varianza.

En Costa Rica se tiene que según el informe de la Encuesta Nacional de Hogares de Julio 2016, como resultados generales se presentan los ingresos con los que disponen los hogares, además pone en manifiesto la desigualdad que se vive en el país. Las comparaciones se realizan de acuerdo al análisis de años anteriores, así como por zonas (urbana y rural), región de planificación, y quintil de ingreso. Así mismo, la desigualdad se mide por medio del Coeficiente de Gini. Según este estudio se estima que el ingreso neto promedio por hogar es de ₡1 027 291 mensuales. En cuanto a la distribución del ingreso, el 20% de los hogares con mayores ingresos acumulan el 50.7% del total de los ingresos de los hogares a nivel nacional y el coeficiente de Gini se estima en 0.521.

Según la Encuesta Continua de Empleo al primer trimestre del 2017, la población ocupada nacional se estimó en 2,06 millones, La tasa de ocupación es producto de relacionar la población ocupada con la población en edad de trabajar (de 15 años o más) para este trimestre fue de 54,0 %. Además, la tasa de ocupación se puede dividir por sexo, presentando una tasa para hombres de 68,0 % y para el sexo femenino de 39,9 %. Según la zona de residencia, 1,53 millones de personas ocupadas residen en la zona urbana, mientras que 528 mil personas ocupadas residen en la zona rural. Esto significó una tasa de ocupación urbana de 54,6 %, mientras que la tasa de ocupación rural se estimó en 52,5 % en este periodo. La tasa de desempleo es de 9,1 %, significando por sexo, una tasa de desempleo de 7,8 % para la población masculina y para las mujeres de 11,3 %. En cuanto a la posición en el empleo, se estimó 1,57 millones personas asalariadas, que representaron cerca del 76 % de la población ocupada total mientras que la población independiente se estima en 449 mil personas.

La Encuesta Nacional de Ingresos y Gastos de los Hogares 2013 (ENIGH) presenta información sobre la composición del presupuesto de los hogares residentes en el país, mediante el conocimiento de sus ingresos y su destino en la adquisición de bienes y servicios. Con respecto a la distribución de los ingresos mensuales, el 10% de los hogares con menor ingreso perciben un promedio por hogar de 179 502 colones mensuales, mientras que, en el otro extremo, del 10% de

los hogares con mayores ingresos, este porcentaje suma el 34%, con un promedio por hogar de 2 888 911 colones mensuales, en los primeros ocho deciles, se concentra el 50% de los ingresos corrientes totales, y la otra mitad del ingreso total de los hogares se reparte entre los que pertenecen a los últimos dos deciles. Esta desigualdad en la distribución del ingreso se resume por medio del Coeficiente de Gini. Este índice toma valores entre cero y uno, donde cero corresponde a la perfecta igualdad y uno corresponde a la perfecta desigualdad. Para Costa Rica el coeficiente de Gini para hogares e individuos se encuentra entre valores de 0.49 y 0.5.

La Encuesta Nacional de Empresas 2016 presenta la remuneración promedio mensual de 7587 empresas nacionales de acuerdo a características como remuneración por sexo, por grupo ocupacional y horas laboradas dependiendo del sector en el que se labore. Los sectores que se toman dentro del análisis son el de manufactura, comercio, transporte y almacenamiento, alojamiento y servicios de comida, enseñanza, otros servicios y otros sectores.

En el año 2015, el periódico El Financiero lanzó una aplicación que hace uso de un modelo estadístico de regresión lineal Gaussiana con función de enlace identidad para conocer los factores que explican las diferencias en el salario bruto de los trabajadores costarricenses entre los 15 y 65 años, tomando en cuenta variables como la edad (experiencia), género, zona de residencia, sector donde labora, rama en la que labora, nivel de escolaridad y horas trabajadas semanalmente. Los datos fueron tomados de la ENIGH del 2013. El modelo estimado no es predictivo si no lo que pretende es mostrar el comportamiento del ingreso monetario del mercado laboral en el año 2013 y presentar resultados de sueldos promedios en el mercado. Dicho modelo explica el 58% de las diferencias salariales en el país.

3. MARCO ACTUAL DE LA CUESTIÓN

Hay diferentes factores que van a determinar el ingreso que percibe una persona, es sabido que, por lo general, factores como el nivel de educación generan una gran influencia en el valor percibido de este monto, sin embargo, hay otros factores que pueden analizarse. En el año 2015, Fernández y Fernández (2015) realizaron una Calculadora Salarial donde estudian con base en datos de la ENIGH 2013 los ingresos promedio para ciertos perfiles de trabajadores. En este estudio se encuentra que las variables más importantes son el tipo de jornada laboral y los años de educación formal, sin embargo variables como el sexo, el sector laboral, la zona de residencia,

la rama de especialización y la edad son factores que también influyen en cuanto se perciba mensualmente.

Calderón y Bonilla (2006) exponen la problemática que sufren las mujeres en el sector laboral nacional debido a la ideología patriarcal, donde la brecha salarial entre hombres y mujeres es uno de los indicadores más claros de la inequidad de género, puesto que dicha brecha se presenta en zonas tanto urbanas como rurales, sector institucional (público o privado) y nivel educativo, a pesar de que desde el año 1952 el convenio número 100 de la Organización Internacional de Trabajo (OIT) estableciera “igual remuneración por igual trabajo”

El estudio realizado con datos del 2013 por Fernández y Fernández es el más reciente a nivel nacional. Además, los estudios de percepción de ingresos en su mayoría se basan en modelos de regresión lineal. Es por esto que contando con datos del INEC del año 2016 y haciendo uso del análisis multivariante de árboles de decisión se puede analizar y comparar los resultados por medio de dos técnicas de análisis multivariantes como son el análisis de regresión lineal y árboles de decisión. De esta manera, se pueden analizar las diferencias que se han presentado en los últimos tres años. Dentro de las limitaciones en dicha comparación se tiene que las encuestas utilizadas, aunque son ambas del INEC persiguen objetivos distintos por lo que los datos recolectados no son los mismos por lo que las variables van a variar, por lo tanto, no se pueden hacer comparaciones directas entre resultados.

Además, el tema de las desigualdades de género es un tema que ha generado controversia a lo largo de los años. Aún en el 2013 se presentan diferencias importantes, sin embargo, no se presenta como el factor de mayor peso, por lo que es interesante analizar qué tantas diferencias arroja esta característica y pensar en acciones que vayan allá más que el solo conocimiento a voces de la desigualdad de género en el país.

4. METODOLOGÍA

4.1 Población de estudio

Conformada por el conjunto de todas las viviendas individuales ocupadas y sus ocupantes residentes habituales, tanto del área urbana como del área rural de Costa Rica. Se excluye la población residente en las viviendas colectivas.

4.2 Marco muestral de viviendas

Estos son los instrumentos utilizados para la selección de muestras de encuestas a viviendas. Las Unidades Primarias de Muestreo (UPM), se toman como áreas geográficas en las que se divide cada distrito del país y contienen en promedio 150 viviendas en las zonas urbanas y 100 viviendas en promedio en las zonas rurales.

Además del conjunto de UPM, se complementa con información demográfica y socioeconómica tomada de los censos nacionales. También se hace uso de mapas cartográficos que permiten la ubicación, ordenamiento y estratificación del marco.

4.3 Dominios de estudio

El diseño muestral de la ENAHO considera como dominios de estudio el nivel nacional, zona urbana y rural, y las seis regiones de planificación. Para estos dominios de estudios se garantizan errores relativos menores al 15% para el porcentaje de hogares en nivel de pobreza, y menores al 10% para otras variables importantes de la encuesta, como es el ingreso per cápita del hogar.

4.4 Estratificación del marco muestral para selección

El diseño muestral de la ENAHO toma como variables de estratificación las regiones de planificación, la clasificación por zona urbana o rural, y por último en la región Central se realiza una clasificación socioeconómica de la zona urbana.

4.4 Diseño de muestreo

El diseño muestral corresponde a un diseño probabilístico de áreas, estratificado y bietápico. De áreas, porque las unidades de muestreo son áreas geográficas o UPM y las probabilidades de selección de las viviendas están asociadas a estas UPM, estratificado ya que para la distribución y selección de la muestra se definieron los estratos de interés antes señalados, y bietápico porque en una primera etapa se seleccionaron las UPM, y en una segunda etapa se seleccionan viviendas o Unidades Secundarias de Muestreo (USM) dentro de las UPM seleccionadas en la primera etapa.

4.5 Tamaño distribución y selección de la muestra

Inicialmente se considera la capacidad técnica y presupuestaria del INEC, y los niveles de precisión obtenidos en las anteriores encuestas en las principales variables de investigación: el ingreso per cápita del hogar y el porcentaje de hogares pobres. Se decide mantener el mismo tamaño de muestra de aproximadamente 13 400 viviendas a nivel nacional, y se trata de homogenizar la precisión relativa de las variables de interés en estos estratos.

Para disminuir el cansancio de la población entrevistada y los posibles rechazos de entrevista, se considera un esquema de rotación del 25% de viviendas en la muestra, esto implica que cada cinco años se renueva completamente la muestra. Para establecer este esquema de rotación, se utiliza un muestreo replicado que consiste en la selección de cuatro réplicas independientes y semejantes en diseño. Este diseño permite además hacer ajustes en la muestra según lo requieran las necesidades de estimación y precisión en los estratos planteados. La selección de la muestra se efectuó en dos etapas: en la primera etapa se seleccionaron UPM con probabilidad proporcional al tamaño, y en una segunda etapa dentro de cada UPM se seleccionaron viviendas o USM con probabilidades iguales de selección, mediante un muestreo sistemático con arranque aleatorio.

5. RESULTADOS

Por medio de árboles de decisión se realiza el análisis para determinar el ingreso neto en colones. Como variables continuas se toman además de la variable objetivo (Ingreso) las variables de edad y años de escolaridad, mientras que como variables categóricas se toman la región, zona, sexo, si es migrante o no, nivel de educación y estado civil.

5.1 Análisis descriptivo de los datos

Se presenta a continuación el análisis descriptivo para las variables continuas.

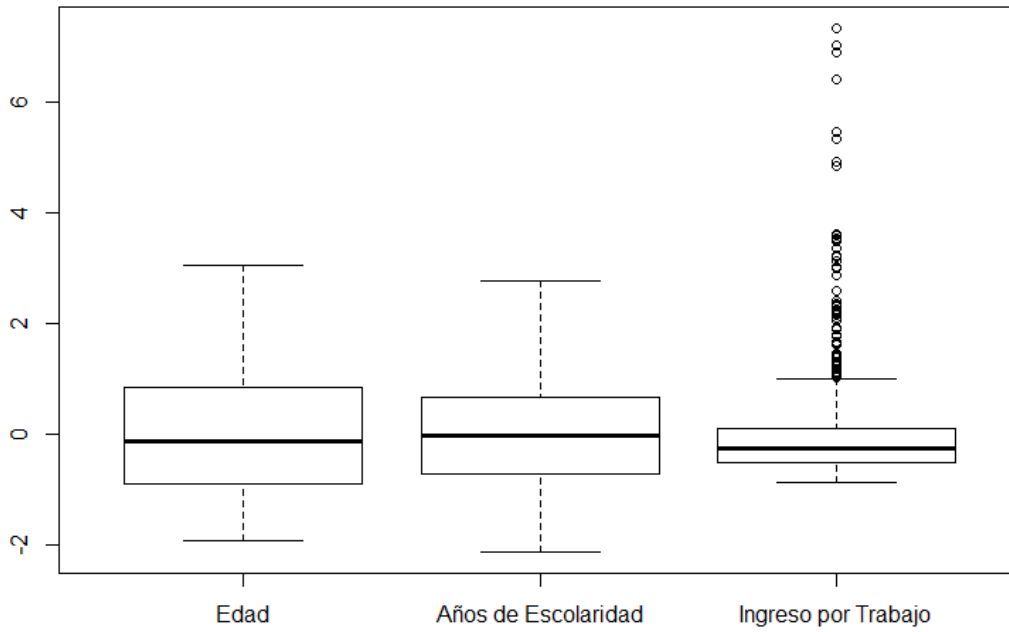


Gráfico 1. Gráficos de cajas para las variables continuas

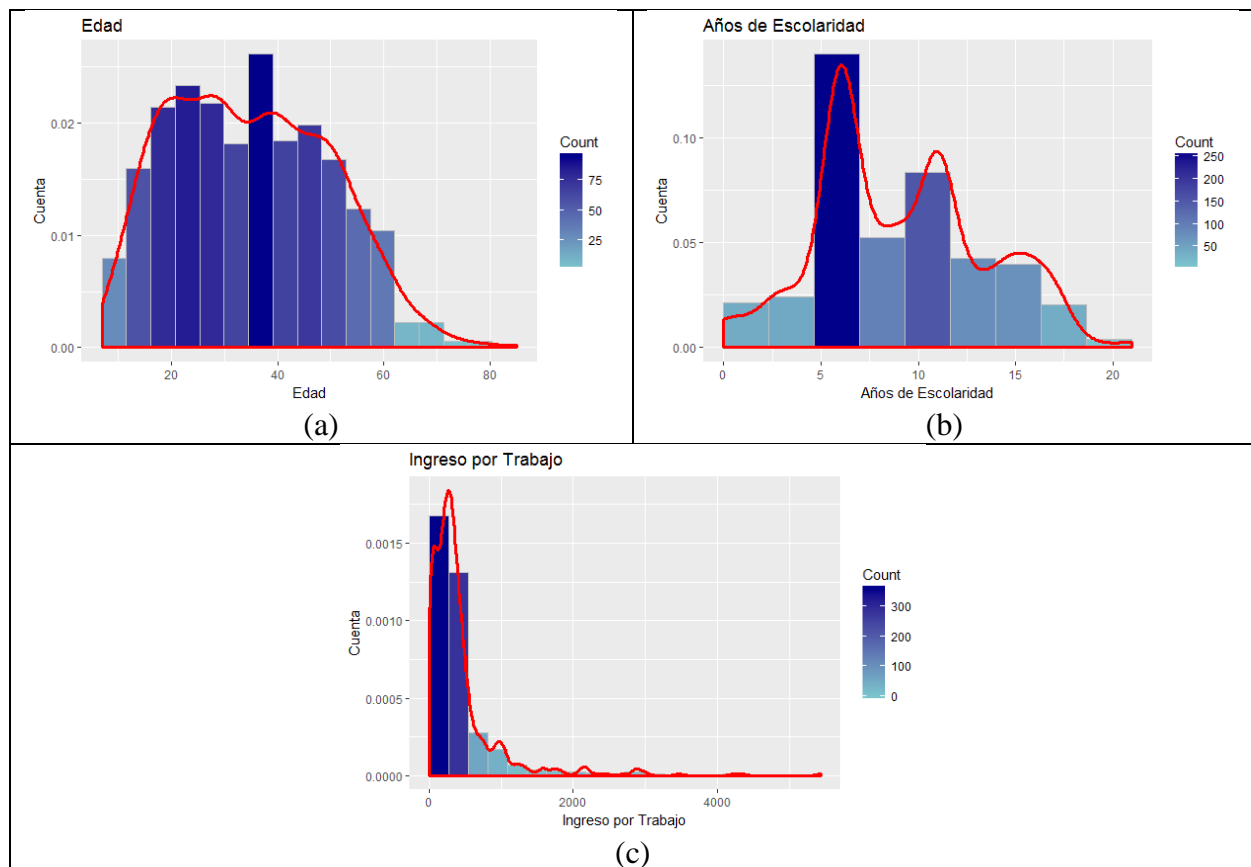


Gráfico 2. Histogramas (a) edad, (b) años de escolaridad, (c) ingreso por trabajo

Se puede observar del Gráfico 1 que las variables presentan distribuciones asimétricas y que la distribución de los datos es más dispersa para la variable edad pero la variable ingreso a pesar de tener el menor rango intercuartil de las tres es la que presenta mayor cantidad de valores extremos.

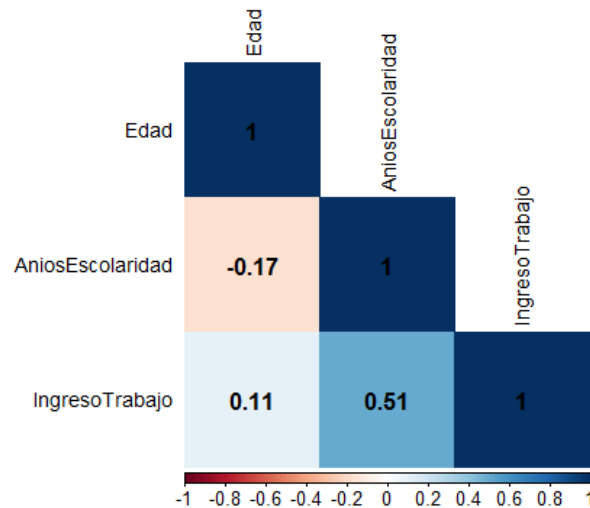


Gráfico 3. Correlograma para las variables continuas

La correlación más alta se presenta entre el ingreso y los años de escolaridad, sin embargo, ningún par de variables presenta una relación alta que permita inferir una relación directa entre variables.

El análisis descriptivo de las variables categóricas se presenta a continuación.

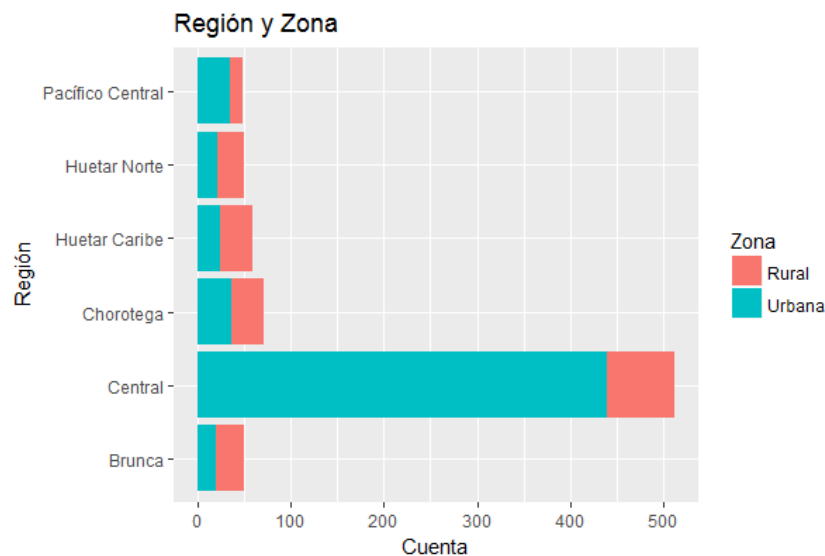


Gráfico 4. Frecuencia por región y zona

Se tiene que la cantidad de casos analizados para zona urbana es mayor que el de zonas rurales respectivamente.

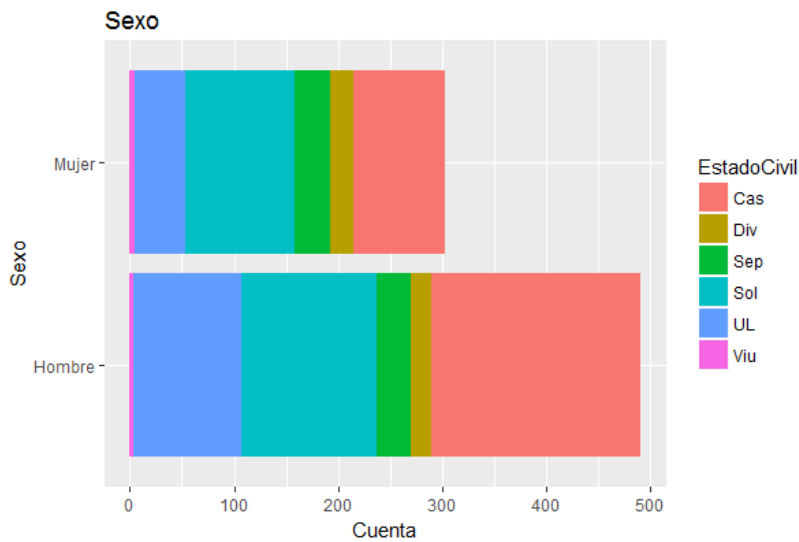


Gráfico 5. Frecuencia por sexo y estado civil

La mayoría de casos estudiados corresponden a hombre y en cuanto a estados civil la mayoría de individuos corresponde con personas casadas y la minoría de los casos es de personas viudas.

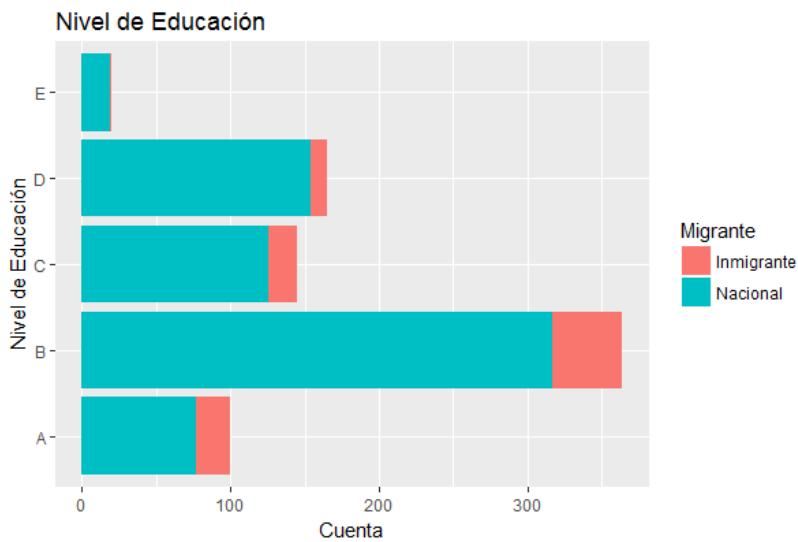


Gráfico 6. Frecuencia en migrantes y nivel de educación

La mayoría de personas son nacionales con un nivel de primaria completa o secundaria incompleta.

5.2 Árbol de decisiones (CART)

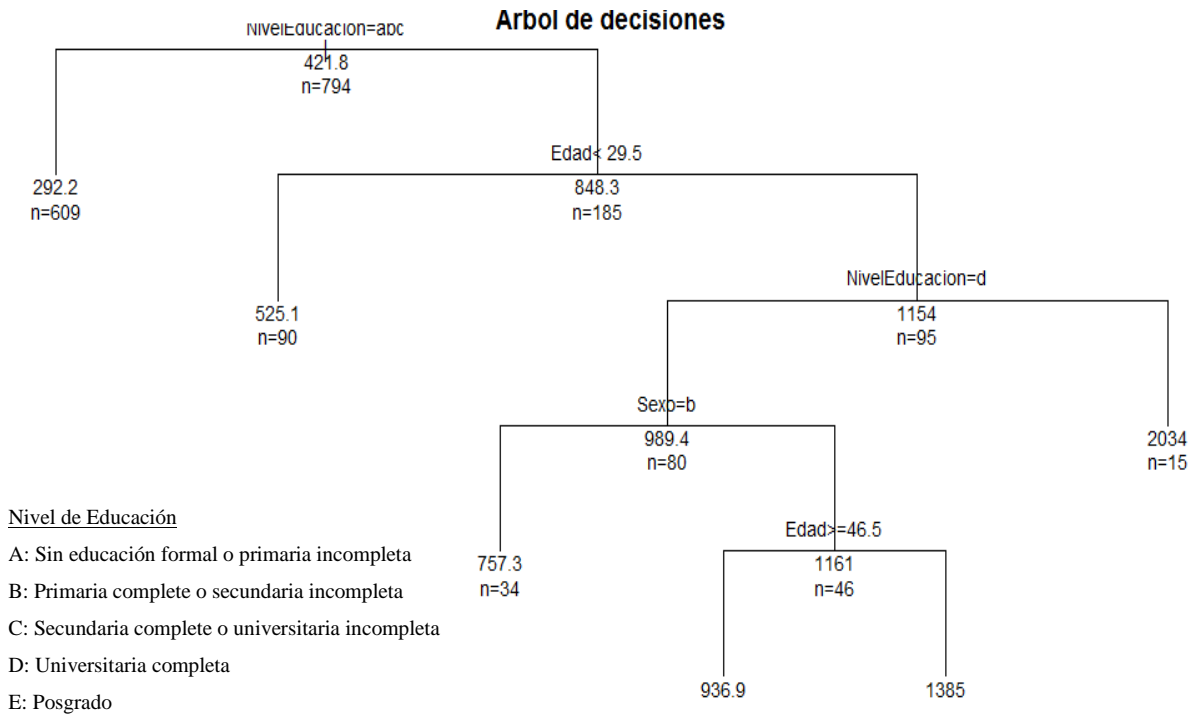


Gráfico 7. Árbol de decisiones (método CART).

Este método toma como base la reducción en la variancia como criterio de decisión para realizar las divisiones. De dicho análisis se obtienen las siguientes divisiones:

- Las personas con niveles de educación más bajos (a, b y c) reciben en promedio ₡292 200 mensuales.
- Las personas con educación universitaria completa, menores de 30 años reciben en promedio ₡525 100.
- Las personas con universidad completa de género femenino mayores a 30 años perciben ₡757 300.
- Los hombres con universidad completa cuyas edades sean entre 30 y 47 años reciben ₡1 385 000.
- Los hombres con universidad completa mayores a 47 años perciben ₡936 900.
- Las personas con posgrado mayores de 30 años reciben en promedio ₡2 034 000.

Del análisis además se obtiene que la importancia de las variables es de 60% para el nivel de educación, 21% para la edad, 10% para el estado civil, 6% para el sexo y de 2% para la región. Del árbol de decisión se puede observar que el estado civil y la región no fueron consideradas.

5.3 Árbol de decisión (inferencia condicional)

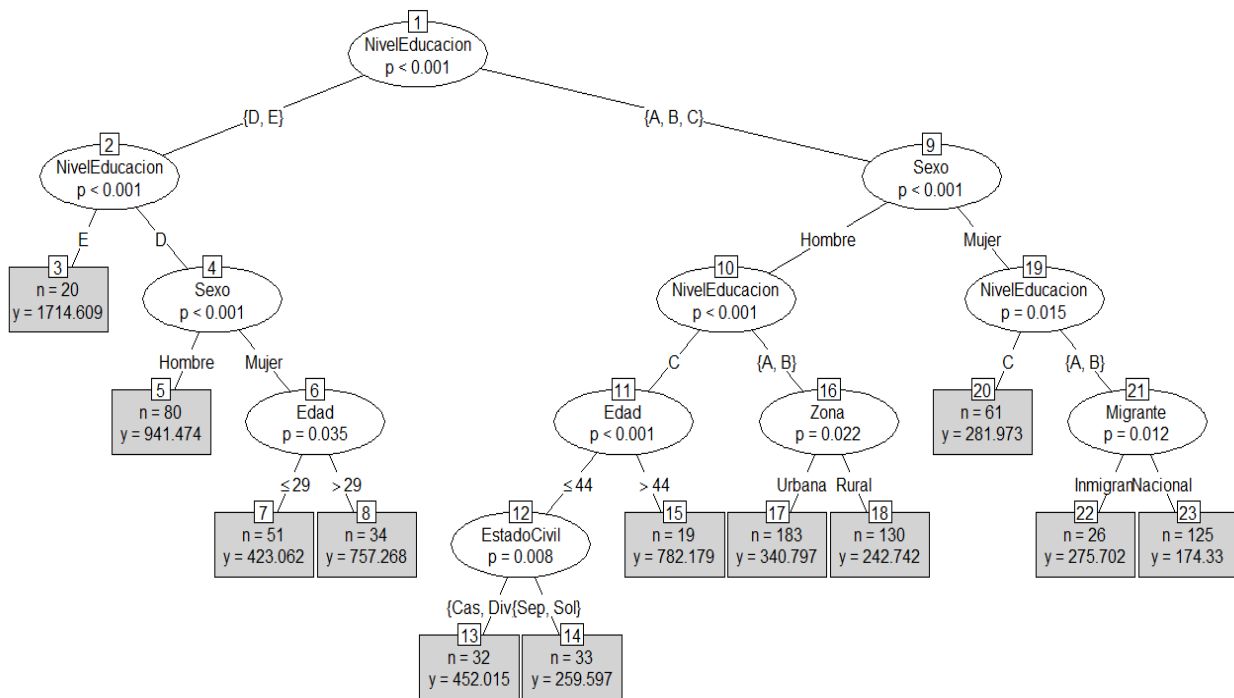


Gráfico 8. Árbol de decisión (inferencia condicional)

Este método toma como base test de permutación como prueba de significancia para realizar la división del árbol.

Los resultados que se obtienen con este método se enumeran a continuación:

- Para hombres menores a 44 años y nivel de educación con secundaria completa que estén casados/unión libre o divorciados, el ingreso mensual es de ₡452 015 en promedio.
- Para hombres menores a 44 años y nivel de educación con secundaria completa que estén separados o solteros, el ingreso mensual es de ₡259 597 en promedio.
- Para hombres mayores de 44 años y nivel de educación con secundaria completa el ingreso promedio es de ₡782 179.
- Los hombres sin educación o primaria completa de zona urbana reciben en promedio ₡340 797 mientras que los de zona rural ₡242 742.

- Las mujeres con secundaria completa reciben en promedio ₡281 973.
- Las mujeres sin educación o con primaria completa costarricenses reciben en promedio ₡174 330 mientras que las inmigrantes perciben ₡275 702.
- Los hombres con universidad completa reciben en promedio ₡941 474.
- Las mujeres con universidad completa menores de 29 años reciben en promedio ₡423 062.
- Las mujeres con universidad completa menores de 29 años reciben en promedio ₡757 268.
- Para las personas con posgrado el modelo estima que es promedio reciben ₡1 714 609.

Se evidencia según el modelo que la variable de mayor importancia al igual que en el modelo CART es el nivel de educación, seguido por características como el sexo y la edad.

5.4 Comparación entre árboles de decisión y regresión lineal

En el Cuadro 1 se presenta la comparación entre el resultado de las metodologías que se realizaron en el estudio.

Cuadro 1. Comparación de resultados según metodología

| Id | Predicción AD - CART | Predicción modelo lineal | Valor real observado en miles | 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 | 421.8 | 421.8 | 577% | 490% |
| MSE | 213450 | 222560 | | | |



Gráfico 9. Predicción y error según método CART

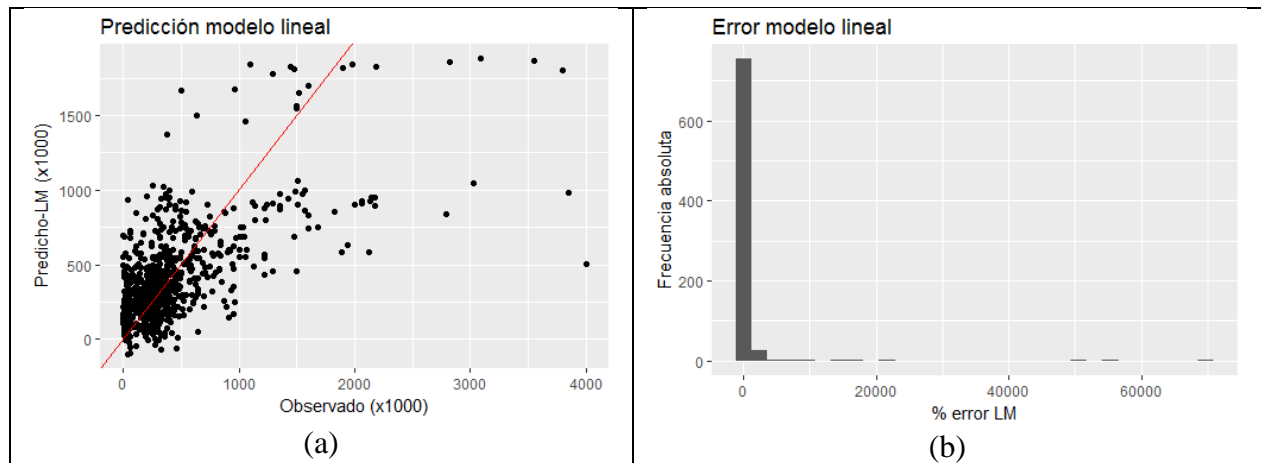


Gráfico 10. Predicción y error según modelo lineal

De la comparación entre los árboles de decisión y el modelo de regresión lineal múltiple se obtiene que, los modelo estiman una gran mayoría de los datos correctamente, sin embargo, se ve afectación en las predicciones debido a los valores extremos presentes puesto que hay valores sobrestimados o subestimados.

El error cuadrático medio para el árbol de decisión es menor que para el modelo lineal.

6. CONCLUSIÓN Y DISCUSIONES

- Las variables de nivel de educación, edad y sexo son las principales 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.
- 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.

Para mejorar el modelo, podría dividirse la población en dos o más conjuntos donde se reduzca la tenencia de valores extremos con el fin de aumentar la homogeneidad de dichos conjuntos

donde sea posible analizar la clasificación de acuerdo a las características pero a su vez se tenga una estimación de los valores predichos, lo que habría que estudiar es donde realizar la división de dichos conjuntos para realmente obtener un beneficio de la metodología.

7. BIBLIOGRAFÍA

- Calderón, F. V., & Bonilla, R. E. (9 de Noviembre de 2006). Desigualdad de salarios entre hombres y mujeres en Costa Rica. San José, Costa Rica.
- Dumas, J. (4 de Mayo de 2016). *Random Forest Analysis of Adult Income in Python*. Obtenido de <https://jasdumas.github.io/2016-05-04-RF-in-python/>
- Fernández, A., & Fernández, A. (15 de Abril de 2015). *¿Cuánto del contraste salarial de Costa Rica responde al nivel educativo?* Obtenido de Periódico El Financiero: http://www.elfinancierocr.com/economia-y-politica/contrastes-salariales-Costa-Rica-escolaridad_0_719328065.html
- Fernández, A., & Fernández, A. (2015). *Calculadora salarial*. Obtenido de El Financiero: <http://www.elfinancierocr.com/gnfactory/especiales/2015/calculadorasalarial/>
- Instituto Nacional de Estadística y Censos (INEC). (27 de Noviembre de 2015). *Acerca del INEC*. Obtenido de <http://www.inec.go.cr/acerca-del-inec>
- Instituto Nacional de Estadística y Censos (INEC). (2016). *Encuesta Nacional de Hogares Julio 2016*. San José: Instituto Nacional de Estadística y Censos.
- Instituto Nacional de Estadística y Censos (INEC). (2014). *Encuesta Nacional de Ingresos y Gastos de los Hogares 2013 (ENIGH)*. San José: Instituto Nacional de Estadística y Censos (INEC).
- Instituto Nacional de Estadística y Censos (INEC). (2016). *Encuesta Nacional de Empresas*. San José: Instituto Nacional de Estadística y Censos (INEC).
- Instituto Nacional de Estadística y Censos (INEC). (2017). *Encuesta Continua de Empleo al primer Trimestre 2017*. San José: Instituto Nacional de Estadística y Censos.
- Johnson, C., Riggs, M., & Downey, R. (1987). Fun with numbers: Alternative Models for Predicting Salary Levels. *Research in Higher Education*, 349-362.
- Lazar, A. (2004). Income prediction via support vector machine. *International Conference on Machine Learning and Applications (ICMLA)* (págs. 143-149). Louisville Kentucky: Youngstown State University.

Liu, H., Kuo, Y.-L., He, Y., & Liu, J. (2006). Bayesian Regression Model for Predicting Income.

Rodríguez, S. (23 de Marzo de 2012). Obtenido de Práctica 1: Regresión Múltiple. Análisis de la estructura salarial de una empresa:

<https://www.slideshare.net/SergioRodrguezFernnd/anlisis-de-la-estructura-salarial-de-una-empresa-anlisis-de-regresin-mltipole>