Contraconceptive Method Choice

Máster Universitario en Ingeniería Informática

Pablo Alfaro Goicoechea
Carlos Morales Aguilera
Carlos Santiago Sánchez Muñoz

UNIVERSIDAD DE GRANADA

30 de Diciembre de 2020

Tratamiento Inteligente de Datos

E.T.S. de Ingenierías Informática y de Telecomunicación

Índice

Presentación

Presentación

Contraceptive Method Choice (enlace de Kaggle)

Encuesta de 1987 en Indonesia acerca del método anticonceptivo escogido por las mujeres.

Factores: educación de los integrantes del matrimonio, orientación religiosa, familia y otros.

El interés parte en el artículo *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms*.

Estado del arte: se han revisado implementaciones, como *Supervised Classification on cmc*, realizadas con el objetivo de observar los resultados obtenidos (acierto del 50 % – 60 %).

Presentación

El **objetivo de esta práctica es entender el problema y ver como se podrían predecir los diferentes comportamientos de las mujeres indonesias en base a su situación, utilizando para ello diferentes herramientas de clasificación**. Pasos a seguir:

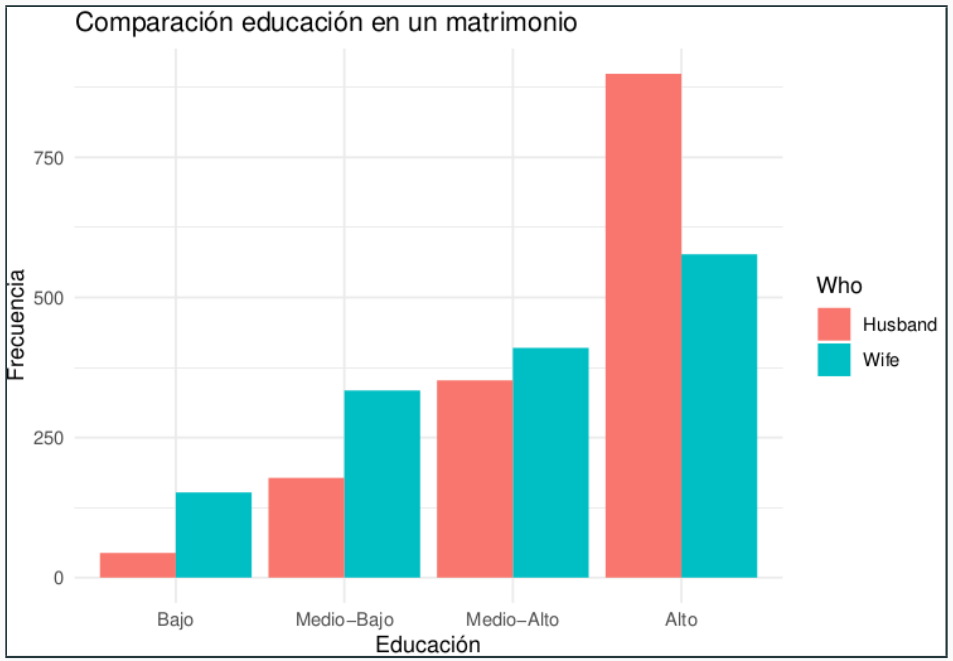
1. Leer los datos correctamente.
2. Análisis exploratorio de los datos.
3. Preprocesamiento de los datos.
4. Clasificación con diversos modelos.
5. Análisis de resultados.
6. Alternativas de planteamiento del problema.
7. Conclusiones finales.

Análisis exploratorio

Educación

Se poseen dos atributos: educación del hombre y de la mujer.

- Cuando la educación es baja, hay más mujeres que hombres.
- Por el contrario hay más hombres que mujeres con una educación alta.

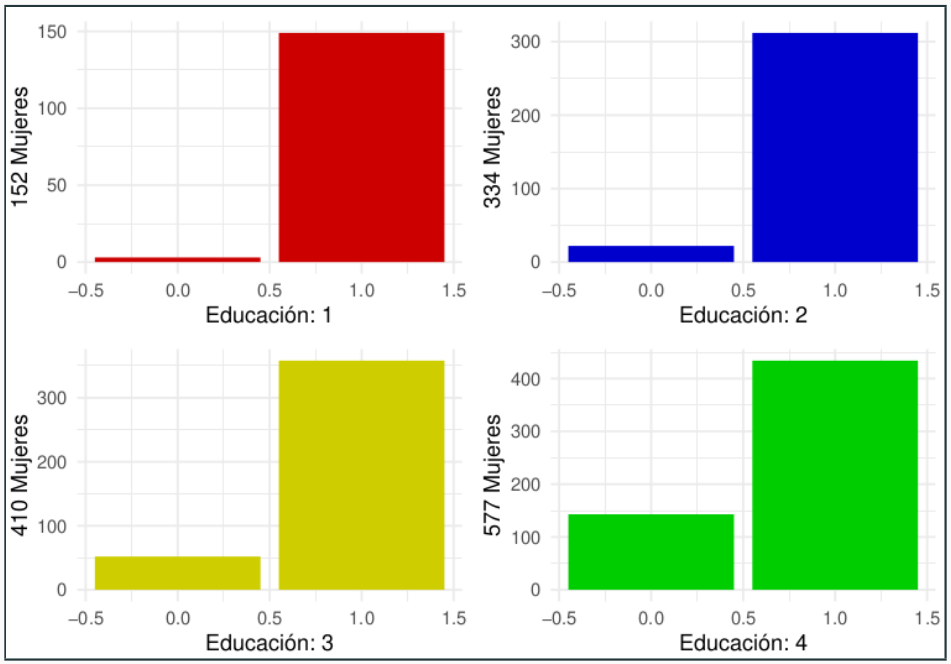


Educación Hombres vs Mujeres.

Religión a partir de educación

Comparativa que relaciona la cantidad de mujeres religiosas existen frente a las que no, para cada nivel de educación considerado:

- Proporción de mujeres creyentes es mayor a las no creyentes.
- Menor nivel de educación $\Rightarrow$ mayor proporción de creyentes.

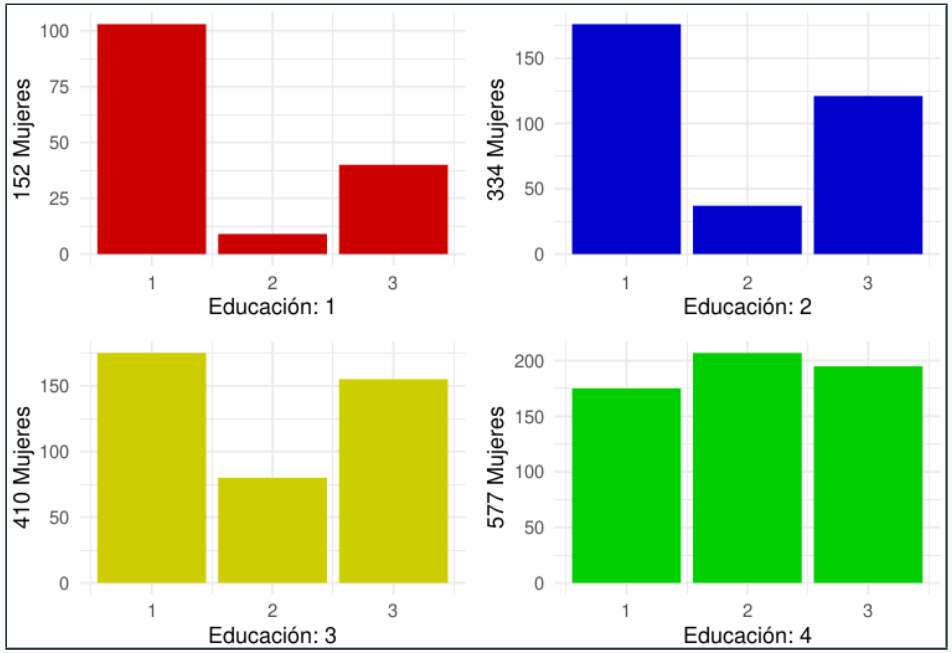


Religión para cada nivel de educación.

Método anticonceptivo a partir de educación

Comparativa que relaciona el método conceptivo para cada nivel de educación considerado:

- Métodos a largo plazo más utilizados cuanto mayor es la educación.
- Utilización de métodos anticonceptivos crece proporcionalmente a la educación de la mujer.

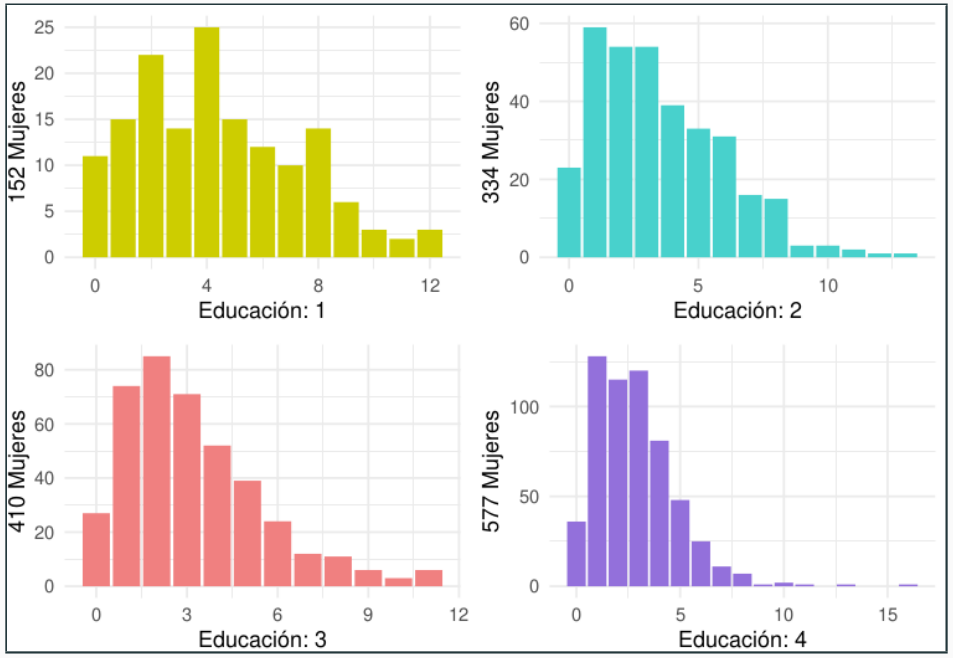


Método anticonceptivo para cada nivel de educación.

Número de hijos a partir de educación

Comparativa que relaciona el número de hijos para cada nivel de educación considerado:

- Familias con mayor número de integrantes tienden a tener mujeres con menor educación.
- Mayor educación $\Longrightarrow$ menor número de hijos.

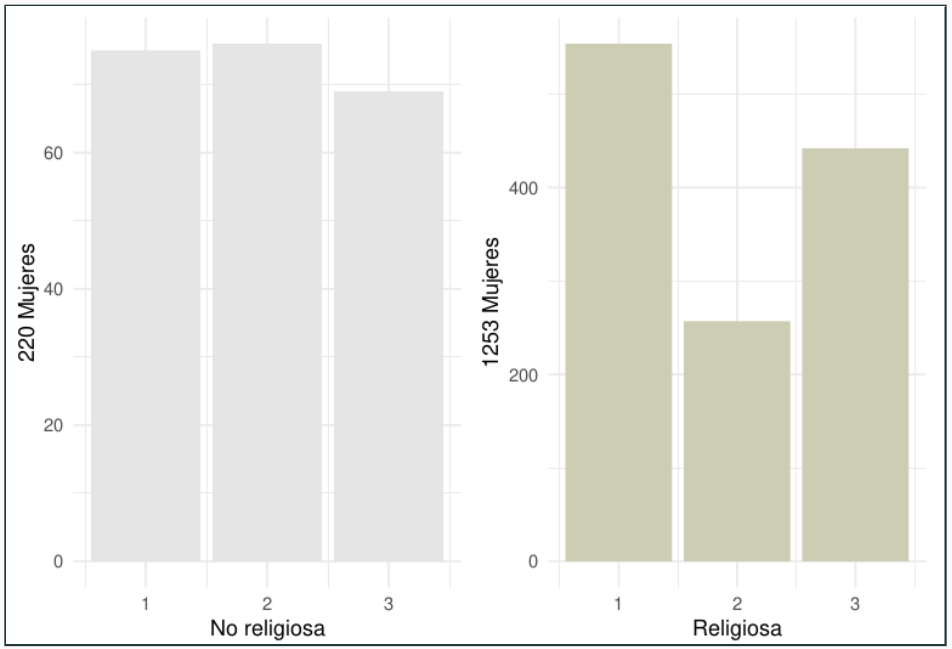


Número de hijos en función de la educación.

Método anticonceptivo a partir de religión

Comparativa que relaciona el método anticonceptivo a partir de la religión:

- Cuando una mujer es religiosa tiende a no utilizar anticonceptivos.
- En las mujeres no religiosas no existe preferencia.

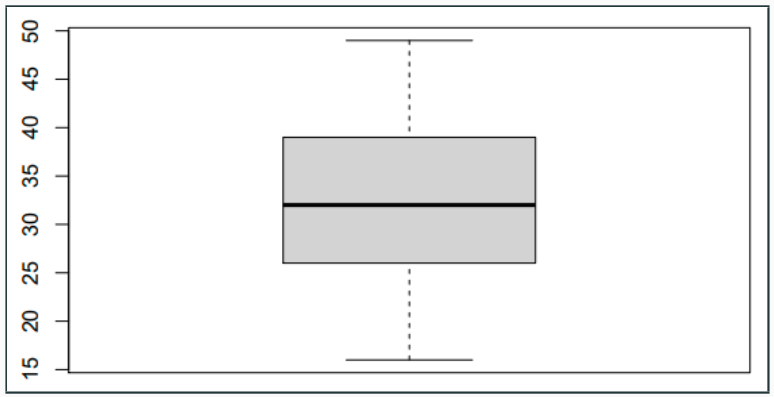


Método anticonceptivo en función de la religión.

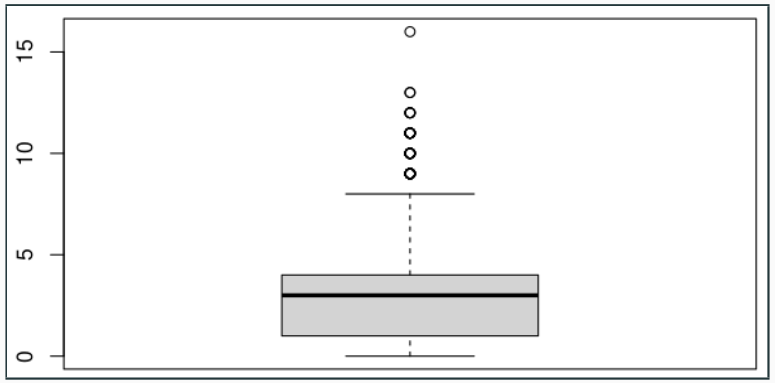
Preprocesamiento

Detección de *outliers*

- Se han aplicado gráficos *boxplots* para la detección de *outliers*.
- Se ha utilizado esta técnica para las variables numéricas de la edad de las mujeres y la del número de hijos.



Edad de las mujeres.



Número de hijos.

Detección de *outliers*

- En la variable de la edad de las mujeres no se han identificado *outliers*.
- Por el contrario, para la variable de número de hijos, si que se han encontrado algunos candidatos.
- Haciendo un análisis estadístico se ha llegado a la conclusión de que son casos especiales. Son casos en los que el número de hijos está muy por encima de la media.
- Se ha optado por eliminar estos ejemplos. Suponían el 3,05 % del conjunto total.

Discretización de variables

- Tras analizar la mayoría de edad en países árabes, y los ciclos de vida de una mujer, se han establecido los siguientes rangos:
  - Menor de edad (1): 16-20 años.
  - Joven (2): 21-30 años.
  - Adulta (3): 31-40 años.
  - Mayor (4): 41-50 años.
- Teniendo en cuenta las características estadísticas de el número de hijos, se han establecido rangos basados en la media y cuartiles:
  - Sin hijos (1): 0 hijos.
  - Pocos (2): 1-2 hijos
  - Media (3): 3-4 hijos
  - Numerosa (4): 5-8 hijos.

Normalización de variables

- Después de discretizar las variables, se ha procedido a evaluar las variables categóricas, de cara a poder ser procesadas fácilmente por los modelos, para utilizar una categorización ordinal.
- No se normaliza la variable `ContraconceptiveMethod` por ser una variable clasificatoria.
- Se ha utilizado una normalización min-max para normalizar.
- Las variables binarias se transforman en booleanas.

Selección de características

- Para seleccionar las características más importantes para el entrenamiento se han utilizado dos métodos con el fin de contrastar ambos modelos y tomar una decisión.
- El primer método es con un modelo de *Boruta*. Se consideran las tentativas y tras esto se confirma si las variables deben permanecer en el modelo o pueden ser eliminadas. Para ello se recibe previamente información sobre su importancia.
- El segundo método ha consistido en entrenar un modelo lineal y observar qué variables son necesarias para construir dicho modelo.

Selección de características

- Viendo los resultados de los anteriores métodos se ha decidido construir un modelo con las siguientes variables:

Variable	Estado
Children	Aceptada
WifeAge	Aceptada
WifeEducation	Aceptada
MediaExposure	Aceptada
StandardOfLiving	Aceptada
HusbandOcupation	Aceptada
HusbandEducation	Aceptada
WifeWorking	Rechazada
WifeReligion	Rechazada
ContraconceptiveMethod	**Clasificatoria**

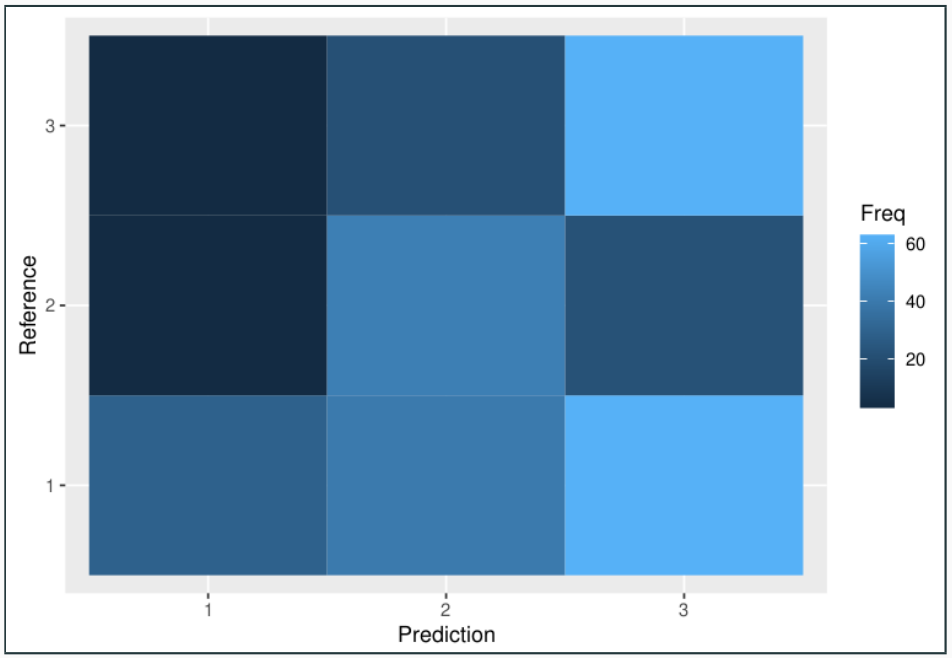
Decisión sobre las variables.

- La información que se descarta es relevante para el problema, pero no permite obtener conclusiones adicionales sobre el mismo, por lo que se procede a eliminarlas.

Clasificación

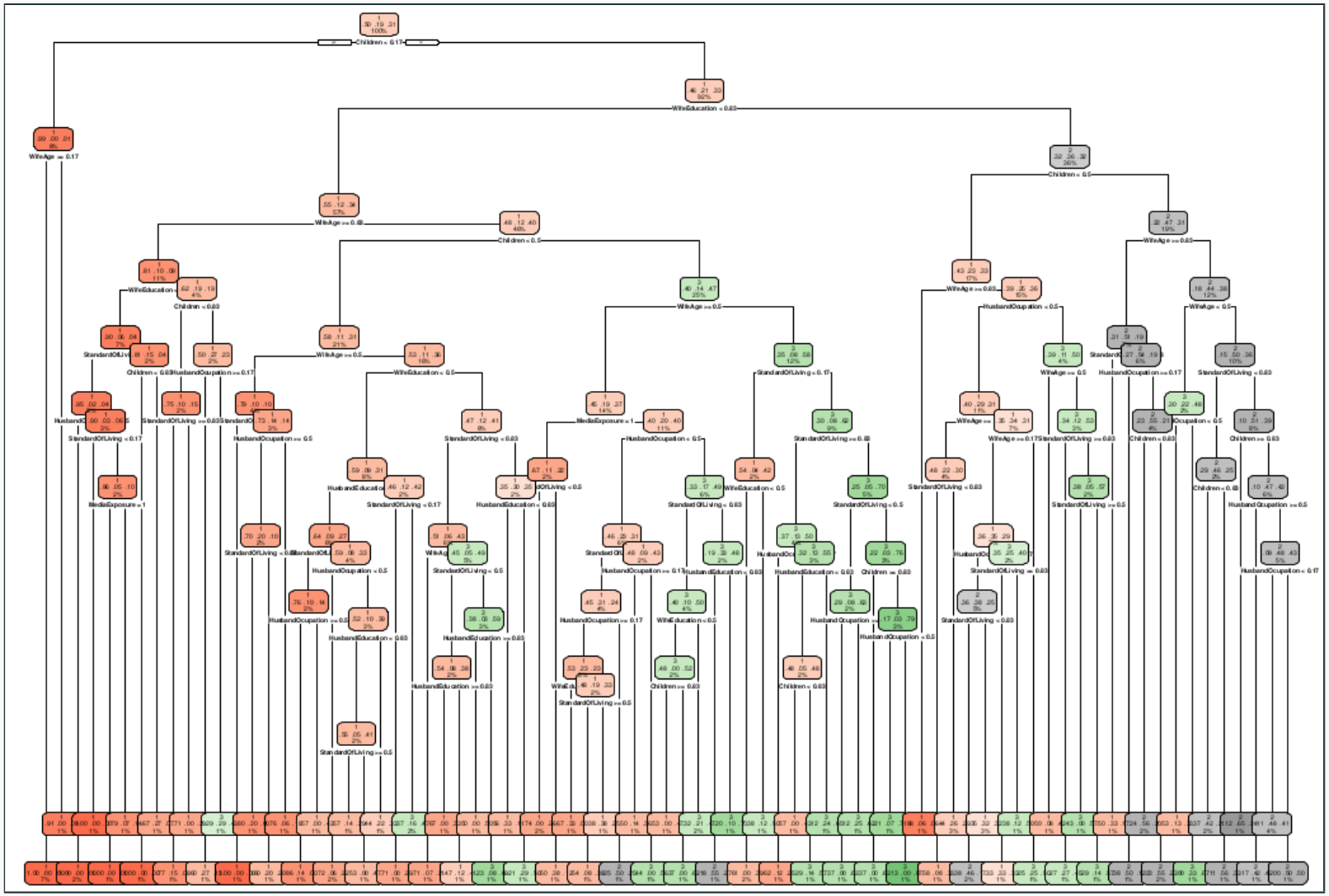
Árboles de decisión

- Árbol binario: se toma una decisión o su opuesta.
- Funciones: `rpart` y `rpart.plot`.
- Parámetro `cp=-1` para que se explore el árbol entero.
- Accuracy: $0,4685$.



Matriz de confusión del árbol de decisión.

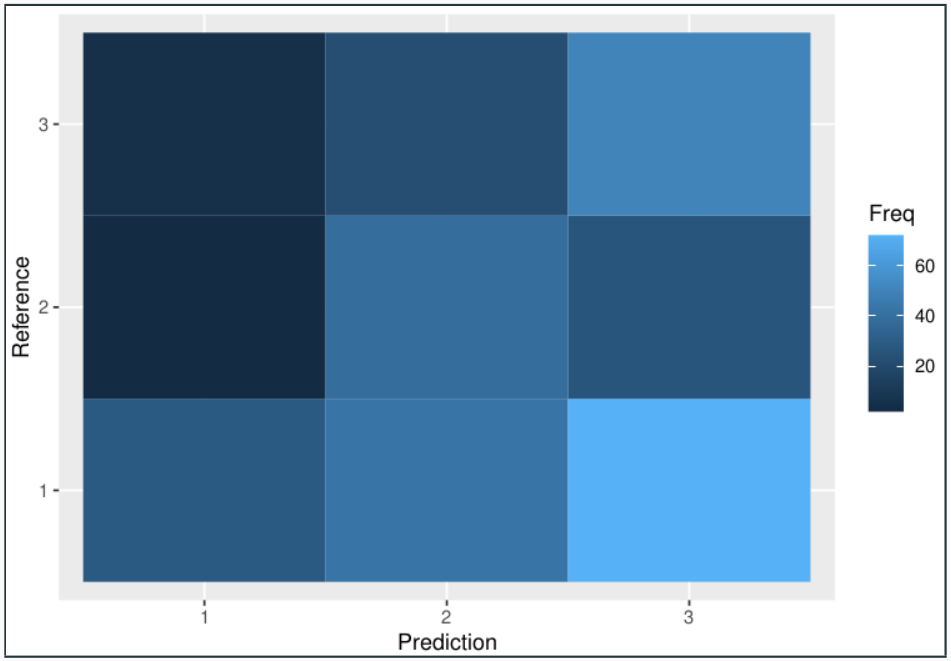
Árboles de decisión



Árbol de decisión.

Random Forest

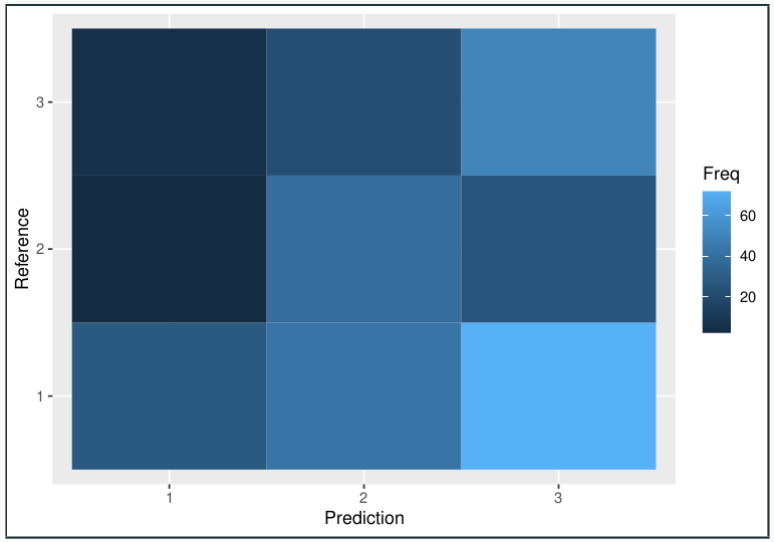
- Divide los datos por características de forma que se tome una decisión o su opuesta formando varios árboles. Votación final.
- Funciones: `randomForest`.
- Parémetros: `ntree`=500 y `mtry`=3 (variables para división).
- Accuracy: $0,4091$.



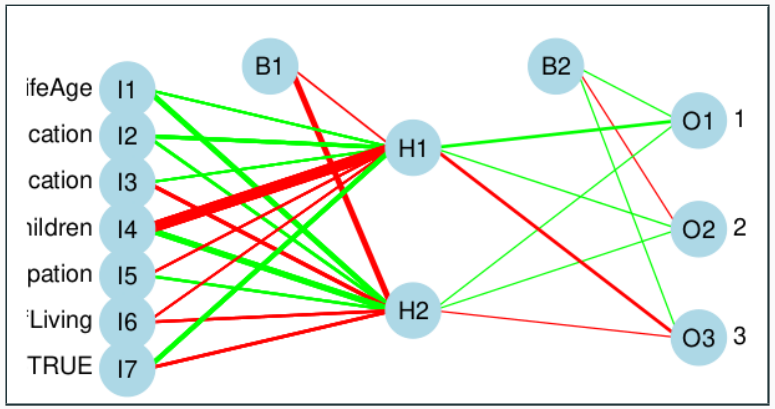
Matriz de confusión de Random Forest.

Red Neuronal Artificial

- Recibe en las neuronas de entrada un conjunto de datos, que son procesados por las capas ocultas para obtener en la salida.
- Librería: `nnet`.
- Parámetros: `size=3` (capas ocultas), `maxit=300` y permitir conexiones entre la entrada y la salida mediante `skip`.
- 2 capas ocultas $\Longrightarrow$ convergencia antes de 300 iteraciones.
- Accuracy: $0,4091$.



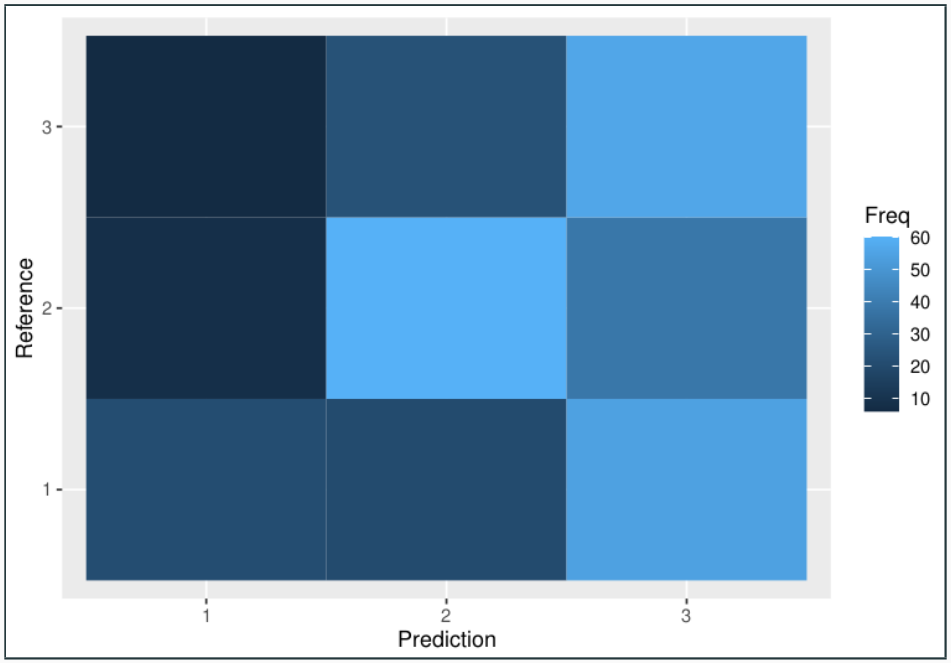
M. confusión RN.



Red Neuronal.

Naive Bayes

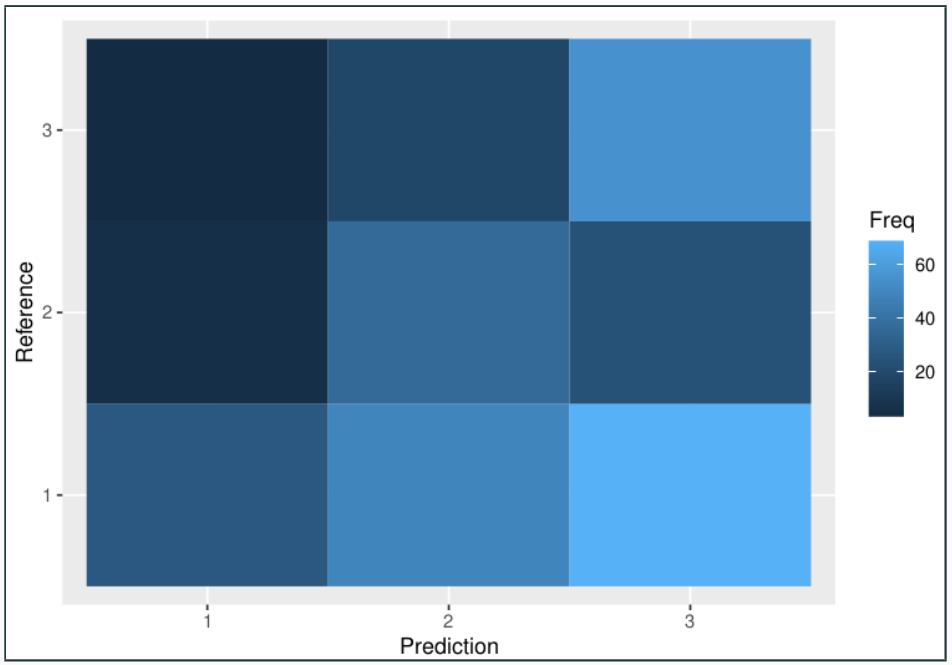
- Asume que la presencia de una característica no es dependiente de la existencia de otra, permitiendo un entrenamiento del modelo mediante las frecuencias relativas de las características del conjunto de entrenamiento.
- Funciones: `e1071` .
- Accuracy: $0,479$.



Matriz de confusión Naive Bayes.

K-Nearest

- Dado un conjunto de entrada con diferentes clases, analiza si un elemento posee dentro de su rango más vecinos de una clase u otra, y clasifica en aquella clase en la que posea más vecinos dentro de un rango con `k` vecinos.
- Funciones: `class`. Parámetros: `k=10`.
- Accuracy: $0,4126$.



Matriz de confusión KNN.

Comparativa y Resultados

Comparativa Accuracy

Resultados obtenidos

Modelos	**Accuracy**
Árbol de decisión	0.4685315
Random Forest	0.4090909
Red Neuronal Artificial	0.4090909
Naive Bayes	0.4790210
KNN	0.4125874

Análisis de resultados

- Se trata de un conjunto de datos desbalanceado, por lo que la información no posee suficientes casos que reflejen todas las posibles condiciones, por lo que es normal que se obtengan resultados que tiendan a una clase más que a otra.

- Uno de los puntos complicados que se trató fue el de selección de características. Planteamos utilizar una conversión en *Dummy* variables pero tras realizar estudio sobre diferentes opciones decidimos utilizar una categorización ordinal.

Análisis de resultados

- Es un problema complejo el decidir realmente que características son las más importantes. Se puede observar que en los diferentes modelos se hace notable que las variables `WifeEducation`, `Children` y `WifeAge` son las más representativas.

- No es de extrañar obtener en un conjunto tan desbalanceado unos resultados aparentemente no muy buenos. Las variables no siguen entre sí todas una relación lógica y existe ruido. Sin embargo, algo positivo es haber obtenido unos resultados tan similares con diferentes modelos, en los que se aproximan a las mismas conclusiones.

Alternativas

Selección aleatoria de instancias

- La primera alternativa propuesta consiste en realizar el problema sobre un subconjunto de datos escogidos aleatoriamente, considerablemente inferior al conjunto inicial (500 instancias).
- De esta manera se intenta minimizar el desbalanceo del conjunto original.

Modelos	**Accuracy**
Árbol de decisión	0.51
Random Forest	0.49
Red Neuronal Artificial	0.49
Naive Bayes	0.48
KNN	0.51

Selección aleatoria de instancias

- Como es de esperar en un modelo tan desbalanceado y con ruido, al reducir el número de instancias, los modelos son capaces de adaptarse con un mejor resultado a los distintos subconjuntos.

- Los resultados obtenidos son ligeramente mejores, aunque en este caso se reduzca a la mitad el número de instancias, se observa que tampoco existe una gran diferencia.

Transformación a un problema más sencillo

- Otra de las alternativas es trasladar el problema a un problema de clasificación en el que se evalúan dos posibilidades: Que la mujer utilice métodos anticonceptivos o no.
- Posteriormente se plantea como un problema de clasificación entre los diferentes métodos anticonceptivos.

Modelos	**Accuracy**
Árbol de decisión	0.7307692
Random Forest	0.7482517
Red Neuronal Artificial	0.7482517
Naive Bayes	0.7272727
KNN	0.7447552

Transformación a un problema más sencillo

- Se examina cuándo se escoge un método u otro, para ello sólo seleccionamos la instancias donde se utiliza un método anticonceptivo.

Modelos	**Accuracy**
Árbol de decisión	0.6121212
Random Forest	0.6000000
Red Neuronal Artificial	0.6000000
Naive Bayes	0.6787879
KNN	0.6727273

- Se observa que el problema es más sencillo de clasificar e interpretar para los modelos.
- Se comprueba que los resultados son mejores y esta podría ser una vía de estudio interesante.

Conclusiones

Conclusiones

- Las variables más interesantes de cara a la clasificación son las de la edad de la mujer, el número de hijos y la educación de la mujer, ya que son las que se encuentran más relacionadas con ella y con la elección del método.
- Realmente se puede observar que es un problema complejo que seguramente requiere de más factores para poder realizar una clasificación más precisa.
- Se destaca que la primera variable a considerar suele ser el número de hijos, lo cual es bastante significativo, por lo que ha sido finalmente una buena decisión excluir casos extremos.

Conclusiones

Existen una serie de relaciones entre las diferentes variables:

- Las mujeres religiosas tienden a no utilizar anticonceptivos.
- Mayor edad de la mujer implica no usar anticonceptivos.
- Mayor educación de la mujer implica usar anticonceptivos, especialmente de largo plazo.
- Mayor número de hijos implica usar anticonceptivos.
- Las mujeres islámicas tienden a no utilizar anticonceptivos.
- Un nivel de vida mayor implica usar anticonceptivos.
- La educación de la mujer tiende a ser más importante que la del hombre en esta decisión.
- Por lo general la opción principal es no usar anticonceptivos, mientras que la menos deseada son los de largo plazo.

Conclusiones

- En este problema la tarea de comprender el comportamiento de las mujeres indonesias es un problema realmente complejo y es más importante que realmente predecir que decisión toman, sino comprender el trasfondo de detrás del mismo.
- Es un problema interesante en la actualidad, ya que la educación sobre el tema a dichas mujeres puede suponer una ventaja que permita poseer un mayor conocimiento y poder tomas decisiones sobre la no concepción.
- El estudio de dicho problema debería ser materia de estudio en Indonesia de cara al empoderamiento de la mujer en países donde no existe dicha cultura.

Gracias por su atención