Universidad Nacional del Oeste

Licenciatura en Informática

Explotación de Datos

Integrantes Grupo 1:

- Robledo Alan
- Farías Gonzalo
- Romano Diego

Docentes:

- Perez Silvia
- Mendoza Dante

21 de octubre de 2023

Índice

1. Introducción

En este informe, se presentará un análisis de dos árboles de decisión. El primero, construido utilizando el paquete rpart, tiene como objetivo predecir si un tumor es benigno o maligno basándose en ciertas características del tumor. Esta predicción es de gran utilidad para los médicos al momento de diagnosticar y seleccionar el tratamiento adecuado para el paciente. Además, proporciona a los pacientes la posibilidad de un diagnóstico temprano, lo que puede aumentar significativamente sus posibilidades de supervivencia.

El segundo árbol de decisión, desarrollado mediante el paquete randomForest, se enfoca en la clasificación del nivel de violencia en una relación experimentada por una persona. Para ello, se utiliza un conjunto de datos que recopila los resultados de pruebas realizadas a individuos de diferentes géneros y edades. Este modelo puede ser valioso en la identificación temprana de relaciones abusivas y en la intervención adecuada para garantizar la seguridad y el bienestar de las personas afectadas.

2. Clasificación del tumor con RPart

RPart, que significa Recursive Partitioning (Particionado Recursivo), es un paquete en R utilizado para llevar a cabo análisis de árboles de decisión. Los árboles de decisión son una técnica de modelado que divide un conjunto de datos en subconjuntos más pequeños en función de criterios específicos. Estos subconjuntos se asemejan a las ramas y hojas de un árbol, lo que da como resultado la estructura de un árbol de decisión.

En este contexto, utilizamos rpart para abordar la clasificación de tumores como benignos o malignos. Para lograrlo, empleamos un conjunto de datos relativamente pequeño que consta de 569 casos. Es importante destacar que rpart no es la elección óptima para conjuntos de datos grandes debido a la complejidad computacional involucrada en la construcción de árboles de decisión. El algoritmo subyacente requiere la búsqueda exhaustiva de divisiones óptimas en las características, lo que puede resultar en un proceso computacionalmente costoso y lento en conjuntos de datos de gran tamaño.

Sin embargo, en nuestro caso, la utilización de un conjunto de datos más pequeño permite que rpart sea una herramienta adecuada para abordar la clasificación de tumores. La capacidad de rpart para construir árboles de decisión claros y comprensibles es especialmente valiosa en aplicaciones médicas, ya que puede ayudar a los médicos a tomar decisiones informadas al diagnosticar y seleccionar tratamientos para los pacientes.

2.1. Conjunto de datos sobre el cáncer de mama.

El cáncer de mama es una enfermedad significativa que afecta a millones de mujeres en todo el mundo. Representa una proporción sustancial de los casos de cáncer, y la detección temprana y la clasificación precisa de los tumores son cruciales para un tratamiento efectivo y el pronóstico del paciente. El conjunto de datos de diagnóstico de cáncer de mama de Wisconsin proporciona una valiosa fuente de información para abordar estos desafíos.

Descripción del Cáncer de Mama
El cáncer de mama comienza cuando las células de la mama comienzan a crecer fuera de control. Estas células a menudo forman tumores que pueden detectarse mediante técnicas de diagnóstico como radiografías o palpitaciones, manifestándose como bultos en el área de la mama. La clasificación de estos tumores en malignos (cancerosos) o benignos (no cancerosos) es esencial para determinar el enfoque de tratamiento adecuado y el pronóstico del paciente.

Objetivo del Análisis
El objetivo del análisis es utilizar el aprendizaje automático, en particular el algoritmo RPart, para clasificar los tumores en malignos o benignos. El uso de técnicas de aprendizaje automático puede mejorar significativamente la precisión de la clasificación, lo que es esencial en la detección temprana y el tratamiento efectivo del cáncer de mama.

2.2. El Árbol de Clasificación de Tumores de Mama

El siguiente arbol de decisión nos indica que de los 100 porciento de los casos, el 63 porciento van a ser benigno y los otros malignos. Del 100 porciento, los que tienen la variable peor perimetro menor a 106, osea el 61 porciento de los casos, van a tener una probabilidad del 95 porciento de ser benignos. De ese 61 porciento los que tengan la variable peor puntos concavos con un valor menor al 0.16, el 59 porciento de los casos totales, van a tener una probabilidad del 2 porciento de ser malignos. El otro 2 porciento de los casos totales que hay (peor puntos concavos es mayor igual a 0.16) van a tener una probabilidad del 78 porciento de ser benignos. Si nos vamos por la otra rama, los que tienen un valor de peor perimetro mayor igual a 106 (el 39 porciento de los casos totales), van a tener una probabilidad del 87 porciento de ser benignos. De esos casos, lo que tengan un peor perimetro mayor igual al 117 (el 29 porciento de los casos totales) van a tener una probabilidad del 100 porciento de ser benignos. En caso que los valores de peor perimetro esten entre 106 y 117 (el 10 porciento de los casos totales) van a tener una probabilidad del 52 porciento de ser benigno. De esos, lo que tengan una peor suavidad menor a 0.14 (el 6 porciento de los casos) van a tener una probabilidad del 20 porciento de ser malignos y los casos restantes van a tener una probabilidad del 95 porciento de ser benignos.

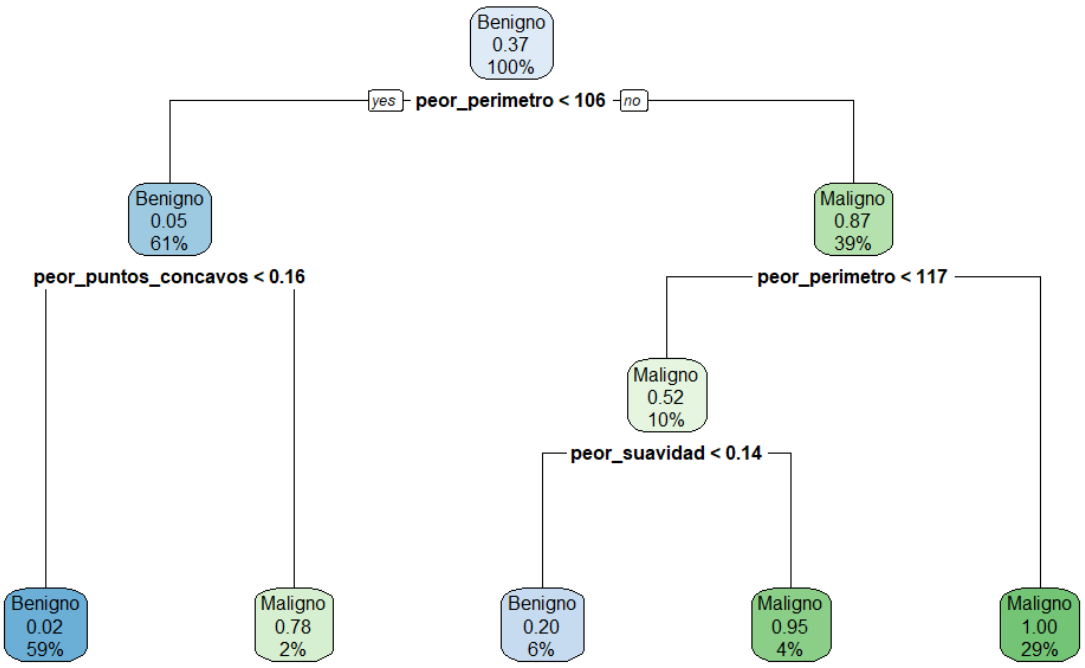

Figura 1

2.3. Matriz de Confusión de Clasificación de Tumores

```
Confusion Matrix and Statistics

          Reference
Prediction Benigno Maligno
   Benigno      87       5
   Maligno       2      48

               Accuracy : 0.9507
                 95% CI : (0.9011, 0.98)
    No Information Rate : 0.6268
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8934

 Mcnemar's Test P-Value : 0.4497

            Sensitivity : 0.9775
            Specificity : 0.9057
         Pos Pred Value : 0.9457
         Neg Pred Value : 0.9600
             Prevalence : 0.6268
         Detection Rate : 0.6127
   Detection Prevalence : 0.6479
      Balanced Accuracy : 0.9416

       'Positive' Class : Benigno
```

Figura 2

Como se puede observar, el arbol de decision para los tumores de mama es muy bueno, ya que tiene Un porcentaje de acierto del 95 porciento, esto significa que el modelo ha acertado en sus predicciones en el 95 porciento de los casos. En el contexto de la clasificación de tumores, esto indica una alta precisión en la distinción entre tumores benignos y malignos. y el coeficente kappa es cercano a 1, este nos indica que nuestro modelo esta muy alejado de clasificar los casos de forma aleatoria, osea que hay concordancia entre nuestra prediccion y la realidad .

2.4. Distribución y Porcentaje de Diagnósticos

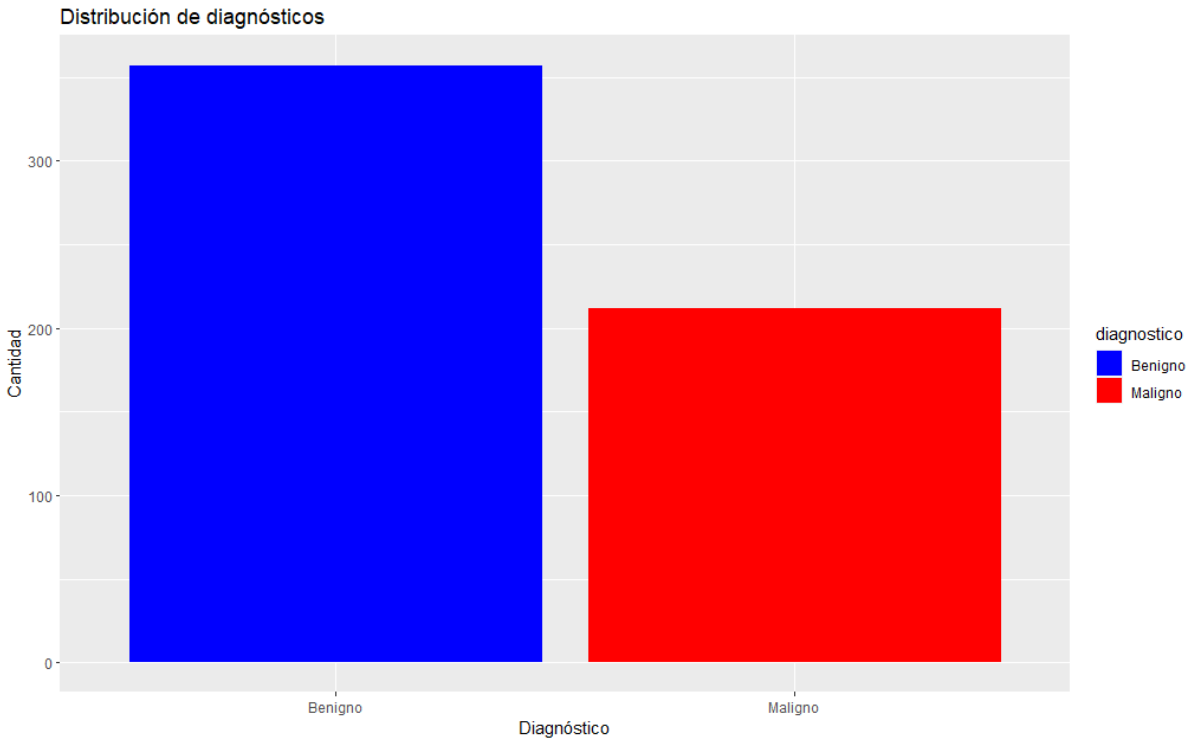

Figura 3

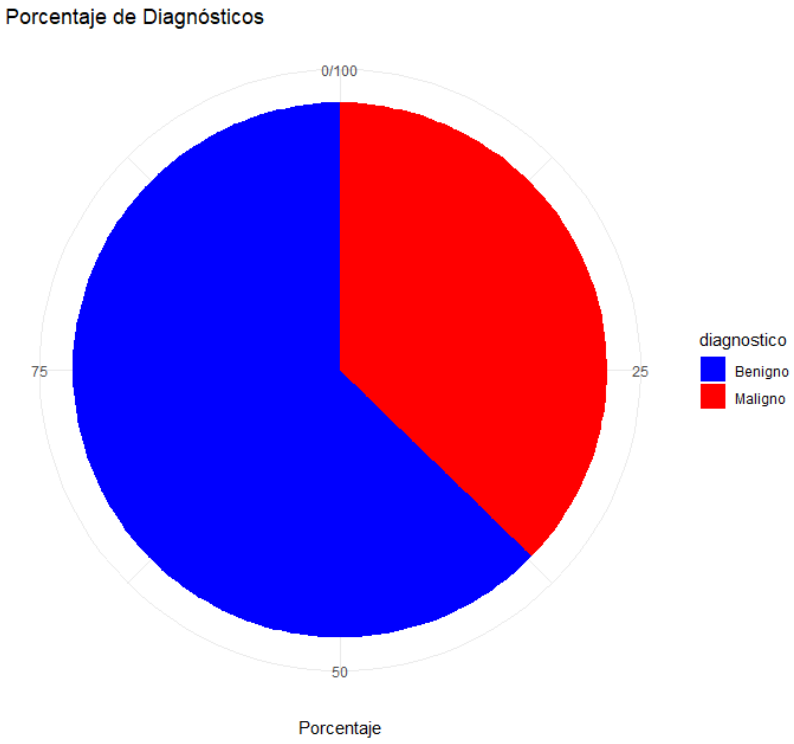

Figura 4

3. Clasificación de Nivel de Violencia en las Relaciones de Noviazgo con Random Forest

Random Forest es un algoritmo de aprendizaje automático que se utiliza tanto para tareas de clasificación como de regresión. Fue desarrollado por Leo Breiman y Adele Cutler y es una extensión del algoritmo de árbol de decisión. Random Forest se basa en la idea de crear múltiples árboles de decisión (un "bosque") y combinar sus resultados para mejorar la precisión de las predicciones. En este ejemplo lo utilizaremos para predecir y clasificar el nivel de violencia que sufre en una relación, para esto utilizaremos un dataset que contine los datos de un test que se les hizo a varias personas de diferentes edades y generos.

3.1. Test alerta sobre noviazgo violento

En este test se hicieron varias preguntas, 13 en total, a varias personas de diversos generos y edades. Algunas de estas preguntas eran:

- ¿Menosprecia en público o en privado tus opiniones?
- ¿Te dice que todo lo que hacés está mal o que no servís para nada?
- ¿Indaga o cuestiona tus noviazgos anteriores?
- Cuando no están juntos, ¿tu pareja te controla preguntándote con quién estás, dónde y qué estás haciendo mensajeándote por celular?
- ¿Revisa los mensajes de tu celular o te pidió la contraseña de tu correo electrónico, Facebook o Instagram como 'prueba de confianza'?
- ¿Te acusa de haber sido infiel o coquetear con otrxs?
- ¿Sentís que están permanentemente en tensión y que, hagas lo que hagas, se irrita o te culpabiliza de sus cambios de humor?

Estas preguntas tenian tres respuestas posibles:
'A' = 'Siempre'
'B' = 'A veces'
'C' = 'Nunca'

Las preguntas del test están diseñadas para evaluar la dinámica de una relación y detectar posibles señales de violencia o abuso emocional. Las tres opciones de respuesta ('Siempre', 'A veces' y 'Nunca') permiten a las personas que responden calificar la frecuencia con la que experimentan ciertos comportamientos en su relación. Las preguntas están destinadas a ayudar a identificar patrones de comportamiento que podrían indicar una relación perjudicial.
En un contexto de detección de violencia en las relaciones, las respuestas a estas preguntas pueden proporcionar información valiosa para evaluar si alguien está experimentando una relación abusiva. Las respuestas 'Siempre' o 'A veces' a muchas de estas preguntas podrían ser señales de alerta de que la relación tiene elementos de violencia o control. Con este mismo criterio realizamos el arbol de decisión que veremos más adelante.

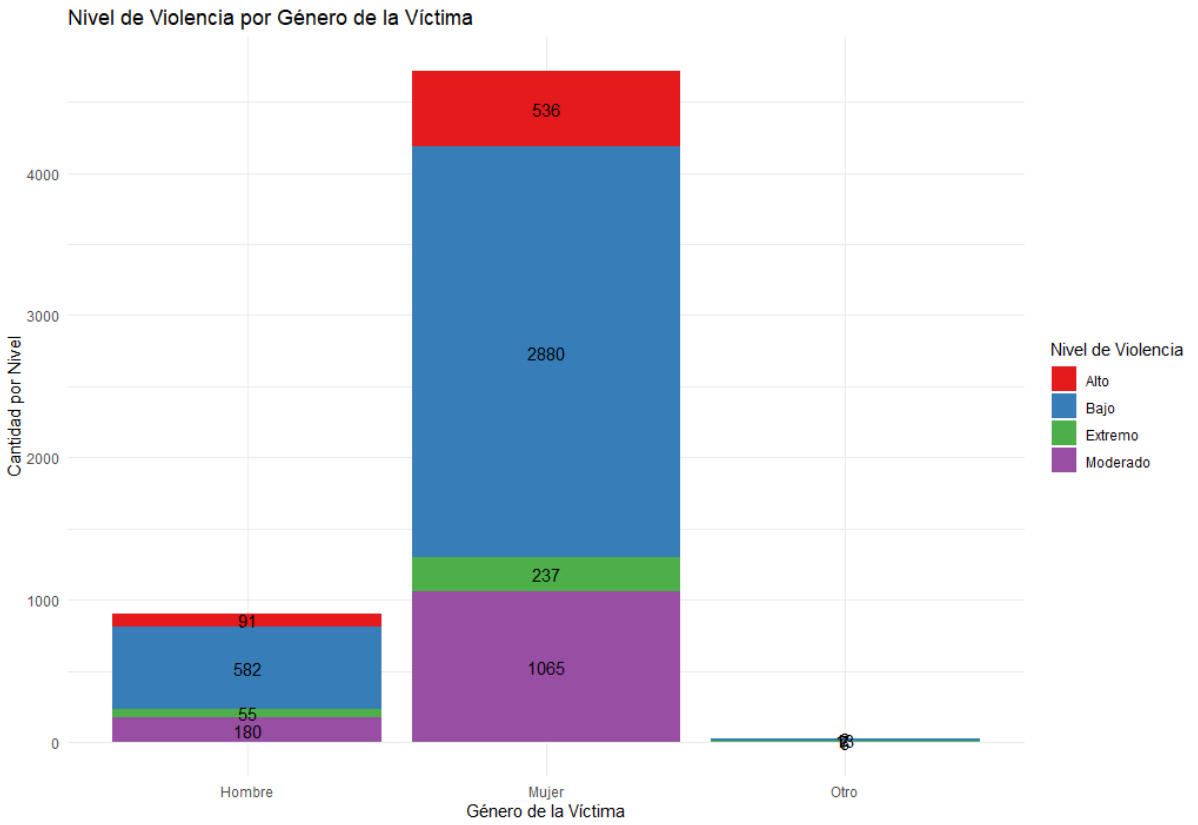

Figura 5: Nivel de violencia por genero

Es importante destacar que la violencia en una relación no se limita solo a la violencia física; el abuso emocional y psicológico también son formas graves de violencia y pueden tener un impacto significativo en la salud y el bienestar de las personas involucradas. Por lo tanto, es crucial abordar estos problemas y brindar apoyo a quienes puedan necesitarlo.

3.2. Árbol de Decisión para Detectar Violencia en Noviazgos

Para este caso se creo una variable categorica llamada nivel de violencia donde su valor varia según la cantidad de preguntas contestadas con un "siempre", si son menos 3 la preguntas contestadas de esta forma podemos decir que sufre de un nivel de violencia bajo, si son 3 o más, hasta 5 preguntas contestadas de esa forma diremos que sufre de un nivel de violencia moderado. En cambio si son entre 6 a 9 preguntas que se responde de esa forma podemos decir que sufre de un nivel de violencia alto, y por ultimo, si son más de 9 preguntas decimos que sufre de un nivel de violencia extremo.

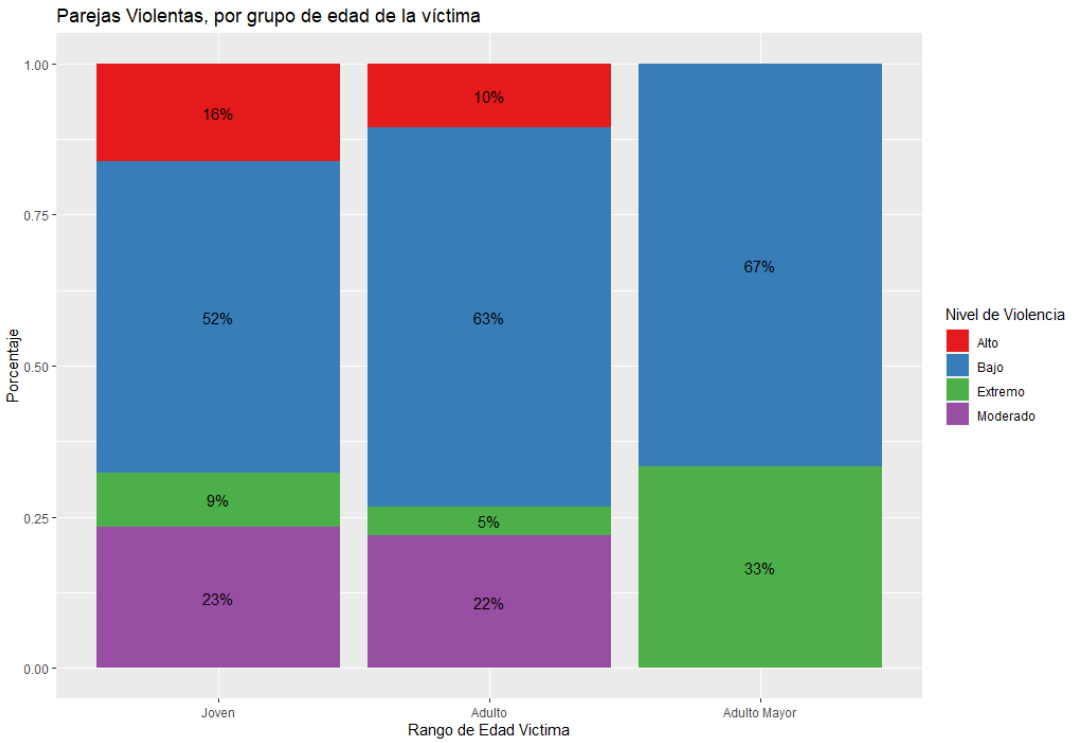

Figura 6: parejas violentas por grupo de edad.

Esta variable categórica es una herramienta valiosa para el análisis y la toma de decisiones en el contexto de relaciones abusivas o violentas, y puede ayudar a profesionales de la salud, consejeros, trabajadores sociales y otros a brindar el apoyo adecuado a las personas afectadas.

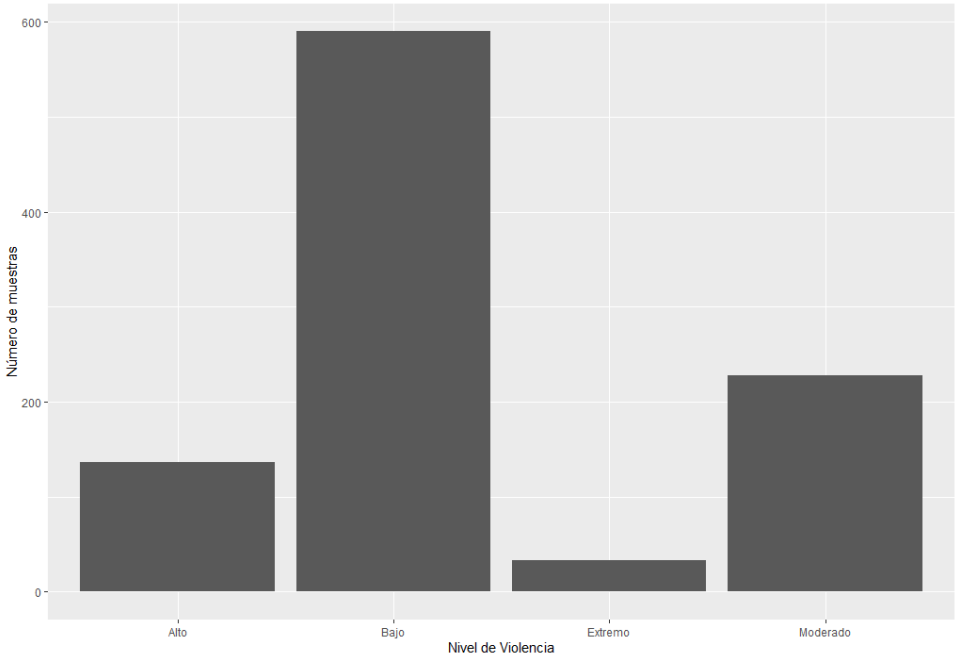

Figura 7: Esta figura muestra los casos totales divididos por el nivel de violencia

```
Call:
 randomForest(formula = nivel_violencia ~ ., data = datosTrainViolencia,      ntree = 400, mtry = 2)
               Type of random forest: classification
                     Number of trees: 400
No. of variables tried at each split: 2

        OOB estimate of  error rate: 5.31%
Confusion matrix:
         Alto Bajo Extremo Moderado class.error
Alto      360    0       1      132 0.269776876
Bajo        0 2756       0       15 0.005413208
Extremo    29    0     210        0 0.121338912
Moderado    3   60       0      957 0.061764706
```

Figura 8: Árbol de Decisión para Detectar Violencia en Noviazgos

Como podemos observar, el arbol de decisión para detectar violencia en noviazgos es muy bueno y eficaz ya que solo tenemos una estimación de error del 5 por ciento. Además, se destaca que los errores en casos particulares de cada variable, como en el extremo, son relativamente bajos, con un 12 por ciento y un 26 por ciento de error en el nivel de violencia alto. Estos resultados sugieren que el modelo de árbol de decisión tiene un buen rendimiento en la detección de violencia en noviazgos, con tasas de error generalmente bajas. Esto es una señal positiva de la utilidad del modelo en la clasificación de los niveles de violencia en las relaciones y puede ser valioso para la identificación temprana y la toma de decisiones informadas en situaciones de violencia en relaciones de pareja.

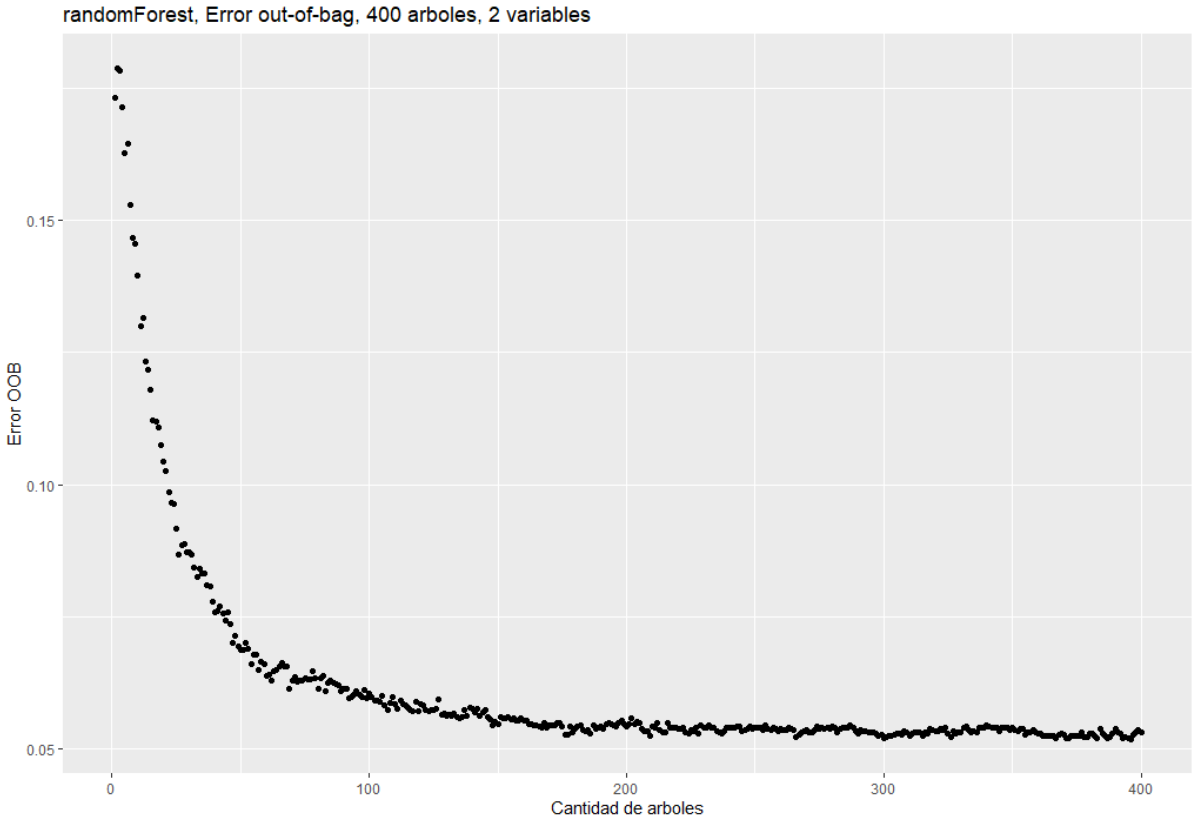

Figura 9: Grafico del error OOB, donde se observa que a partir de 200 arboles el error se estabiliza

3.3. Matriz de Confusión de Niveles de Violencia

A continuación podemos ver los resultados de la matriz de confusión, donde los resultados siguen siendo bastantes buenos y nos vuelve a confirmar de que estamos hablando de un arbol de decisión totalmente utilizable y practicamente fiel a la realidad.

```
Confusion Matrix and Statistics

          Reference
Prediction Alto Bajo Extremo Moderado
  Alto      104    0      12        0
  Bajo        0  586       0       12
  Extremo     1    0      21        0
  Moderado   31    4       0      216

Overall Statistics

               Accuracy : 0.9392
                 95% CI : (0.9224, 0.9533)
    No Information Rate : 0.5978
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8919

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Alto Class: Bajo Class: Extremo Class: Moderado
Sensitivity               0.7647      0.9932        0.63636           0.9474
Specificity               0.9859      0.9698        0.99895           0.9539
Pos Pred Value            0.8966      0.9799        0.95455           0.8606
Neg Pred Value            0.9633      0.9897        0.98756           0.9837
Prevalence                0.1378      0.5978        0.03343           0.2310
Detection Rate            0.1054      0.5937        0.02128           0.2188
Detection Prevalence      0.1175      0.6059        0.02229           0.2543
Balanced Accuracy         0.8753      0.9815        0.81766           0.9506
```

Figura 10: Matriz de confusión que evalua el rendimiento del árbol.

Un porcentaje de acierto del 94 por ciento, esto significa que el modelo ha acertado en sus predicciones en el 94 por ciento de los casos. En el contexto de los niveles de violencia en una pareja, esto indica una alta precisión en la distinción de los distintos niveles, ya sean Extremo, Alto, Moderado o Bajo. El coeficente Kappa es cercano a 1, da 89 por ciento, este nos indica que nuestro modelo esta alejado de clasificar los casos de forma aleatoria, o sea que hay concordancia entre nuestra prediccion y la realidad.

4. El Coeficiente kappa de Cohen

4.1. ¿Qué es?

El Coeficiente Kappa de Cohen es una medida estadística que ajusta el efecto del azar en la proporción de la concordancia observada. En general se cree que es una medida más robusta que el simple cálculo del porcentaje de concordancia, ya que K tiene en cuenta el acuerdo que ocurre por azar. La variable 'Kappa' en el contexto de Random Forest y otros modelos de clasificación se refiere al coeficiente kappa. El coeficiente kappa es una medida de la concordancia entre las predicciones de un modelo y las clasificaciones reales en un problema de clasificación.

El coeficiente kappa es útil cuando se trabaja con problemas de clasificación en los que las categorías pueden no estar perfectamente equilibradas y puede haber una prevalencia diferente de las clases. Ayuda a evaluar cuánto mejor está funcionando un modelo en comparación con una simple coincidencia al azar. El valor del coeficiente kappa varía entre -1 y 1, donde:

- Un valor de kappa igual a 1 indica una concordancia perfecta entre las predicciones y las clasificaciones reales.
- Un valor de kappa igual a 0 indica que el modelo está prediciendo al azar, sin concordancia significativa.
- Un valor de kappa negativo indica que el modelo está funcionando peor que una predicción al azar.

El coeficiente kappa tiene en cuenta tanto las predicciones correctas como las incorrectas, ajustando el resultado de acuerdo con lo que podría esperarse al azar. Esto lo convierte en una medida útil para problemas de clasificación donde las clases pueden estar desequilibradas.
En el contexto de Random Forest y otros modelos de clasificación, el coeficiente kappa se utiliza como una métrica de evaluación para medir la calidad de las predicciones del modelo. Un valor de kappa más cercano a 1 indica un mejor rendimiento del modelo en términos de concordancia con las clasificaciones reales.

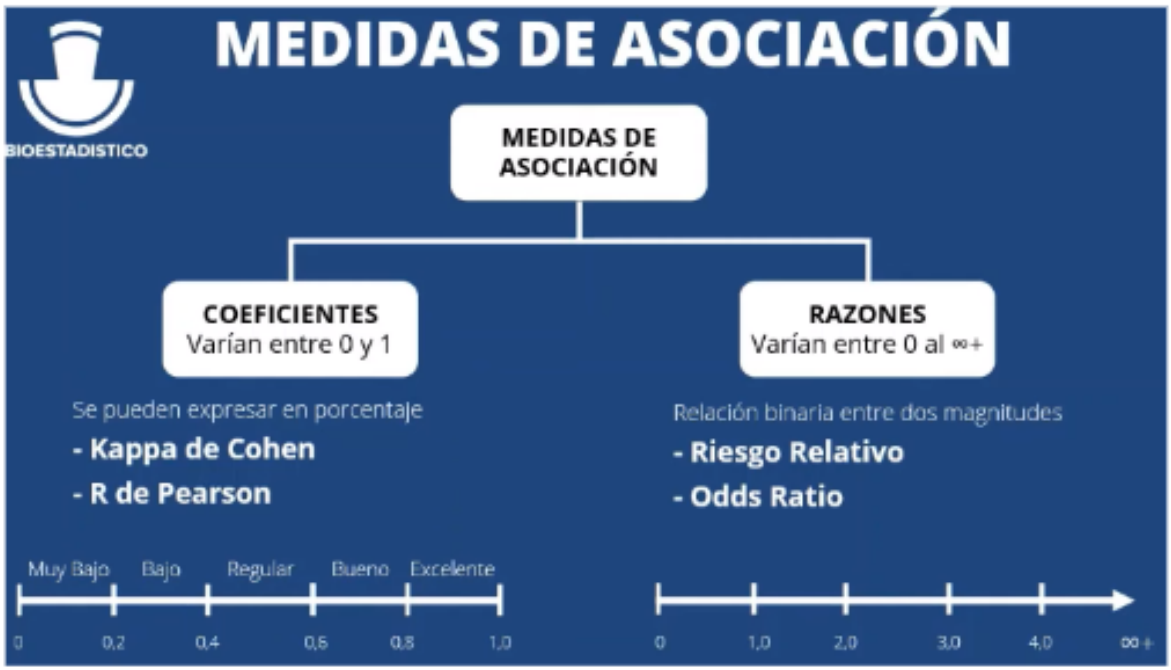

Figura 11

4.2. ¿Cómo se calcula?

La ecuación para K es:

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

$$K = \frac{\text{Concordancia observada}_{Pr(a)} \cdot \text{Concordanciia atribuible al azar}_{Pr(e)}}{\text{Total observaciones}_{1} \cdot \text{Concordancia atribuible al azar}_{Pr(e)}}$$

Figura 12: Formula para calcular el coeficiente Kappa

Donde Pr(a) es el acuerdo observado relativo entre los observadores, y Pr(e) es la probabilidad hipotética de acuerdo por azar, utilizando los datos observados para calcular las probabilidades de que cada observador clasifique aleatoriamente cada categoría. Si los evaluadores están completamente de acuerdo, entonces K = 1. Si no hay acuerdo entre los calificadores distinto al que cabría esperar por azar (según lo definido por Pr(e)), K = 0.

5. Conclusión

En este proyecto, aplicamos árboles de decisión en dos escenarios: detección de cáncer de mama y evaluación de niveles de violencia en parejas. Los resultados fueron prometedores

- En la detección de cáncer de mama, nuestro modelo de RPart mostró un rendimiento sólido en la clasificación de malignidad y benignidad. Puede ser útil en diagnósticos médicos.
- Para la evaluación de niveles de violencia en parejas con Random Forest, creamos un modelo preciso, que podría ser valioso en la identificación de situaciones de violencia doméstica.

Se ve el potencial de los árboles de decisión en diferentes contextos. Estas técnicas pueden ser herramientas útiles para tomar decisiones basadas en datos en medicina y problemas sociales.

6. Bibliografía

- Kaggle - Breast Cancer Dataset
  https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset
- Test - Señales de alerta en el noviazgo
  https://ash.buenosaires.gob.ar/desarrollohumanoyhabitat/mujer/senales-de-alerta-en-el-noviazgo
- BA Data - Test de Alerta sobre un noviazgo violento
  https://data.buenosaires.gob.ar/dataset/test-alerta-sobre-noviazgo-violento