Wisconsin Prognostic Breast Cancer
Clasificación usando Bosques Aleatorios

Dany Efrain Rubiano Jiménez
Universidad de Santiago de Chile

Abstract. La base de datos Wisconsin prognostic breast cancer alberga información acerca del seguimiento de distintos casos de cáncer de mama, enfocándose en la identificación de la recurrencia de este cáncer en específico. En esta oportunidad se hace uso del método de clasificación por medio de bosques aleatorios, buscando obtener conocimiento acerca de la recurrencia y no recurrencia del cáncer.

Keywords: Recurrencia, agrupamiento.

1 Introducción

Una recurrencia o cáncer de mama recurrente es un cáncer de mama que vuelve a aparecer después de un determinado período en el que ya no fue detectado. Es un cáncer de mama que se ha propagado a otra parte del cuerpo, las células cancerosas pueden desprenderse del tumor original de la mama y alojarse en otras partes del cuerpo usando el torrente sanguíneo o el sistema linfático, una gran red de ganglios y vasos que eliminan bacterias, virus y desechos celulares.

Los objetivos de este estudio corresponden a:

- Obtener una clasificación con el método de bosques aleatorios.
- Determinar cuáles son las variables que tienen mayor importancia en la clasificación, dadas las facilidades que brinda el mismo método.

2 Métodos y datos

2.1 Métodos utilizados

En vista de que los datos están desbalanceados, donde la clase no recurrente presenta una gran mayoría de instancias en comparación con la clase recurrente, se utilizan distintas técnicas de balance de datos tales como el submuestreo, sobremuestreo, rose y smote.
El *submuestreo (under)*, selecciona al azar un subconjunto de muestras de la clase con más instancias para que coincida con el número de muestras procedentes de cada clase.
El *sobremuestreo (over)*, genera al azar instancias adicionales en función de los datos, para que coincida con el número de muestras en cada clase.
Rose, genera las muestras equilibradas artificiales de acuerdo con un enfoque de bootstrap suavizado y permiten ayudar a las fases de estimación y evaluación de precisión de un clasificador binario en presencia de una clase rara.

Por último, *Smote* que realiza una combinación dela técnica de sobremuestreo de la clase minoritaria y submuestreo de la clase mayoritaria.

A través de los resultados de cada uno de las técnicas anteriores, se realiza una comparativa con distintas métricas tales como el área bajo la curva ROC (AUC-ROC) y el área bajo la curva de recuperación de precisión (AUPRC).

Luego, el método principal a utilizar es el de clasificación Random Forest (bosques aleatorios), método predictivo que usa la técnica de Bagging para combinar diferentes árboles, donde cada árbol es construido con observaciones y variables aleatorias.
En forma resumida este método:

1. Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes sets de datos.
2. Crea un árbol de decisión con cada set de datos, obteniendo diferentes árboles, ya que cada set contiene diferentes individuos y diferentes variables en cada nodo.
3. Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (es decir, sin podar).
4. Clasifica los datos usando el "voto mayoritario" del conjunto de árboles.

Entonces, en lo que atañe al presente estudio, se aplica Random Forest al modelo escogido según la técnica de balance de datos, para luego probar el clasificador según el conjunto OOB, que corresponde a los datos que no fueron remuestreados., de manera de encontrar los parámetros que otorgan mejores resultados. Finalmente, el clasificador hecho con Random Forest en base a dichos parámetros, es variado según la importancia de cada variable.
Por último, los clasificadores finales son comparados y seleccionados dado el nivel de error durante la clasificación de los conjuntos de prueba.

2.2 Datos utilizados

El conjunto de datos reúne diferentes registros del seguimiento de un caso de cáncer de mama. Estos pacientes fueron vistos consecutivamente por el Dr. Wolberg desde 1984, e incluyen sólo aquellos casos que presentan cáncer de mama y ninguna evidencia de metástasis a distancia al momento del diagnóstico.
El conjunto de datos cuenta con 194 instancias con 34 atributos y la clase. Esta última corresponde al resultado, recurrencia o no recurrencia del cáncer.

3 Resultados

Dado que los datos se encuentran desequilibrados, teniendo la clase recurrente un 76% del total, se procede a balancear los datos utilizando las técnicas anteriormente descritas. La elección se hace en base a la comparativa de los resultados según criterios de AUC de la ROC y el AUC de la curva de Recuperación de Precisión, los que se presentan en la Tabla 3.1. Así mismo, dado que el modelo de balanceo se

construye en base a la clasificación según Random Forest con parámetros por defecto, se presenta los resultados del modelo de clasificación en la Tabla 3.2.

Tabla 3.1: *Metricas de AUC-ROC y AUPRC para cada una de las técnicas*

Modelo	AUC ROC	AU PRC
Original	0,5787	0,2052
Under	0,5175	0,2963
Over	0,5140	0,2733
Smote	0,5140	0,2414
Rose	0,6407	0,1712

Tabla 3.2: *Modelo de clasificación para la construcción del modelo de balanceo*

Modelo	Class.error N	Class.error R	OOB
Original	0,2884	0,9090	24,09%
Under	0,4545	0,4242	43,94%
Over	0,0865	0,0288	5,77%
Smote	0,0757	0,1010	8,66%
Rose	0,2602	0,4375	34,31%

Se escoge finalmente la técnica Smote para balancear los datos. Se obtienen al final 132 instancias, distribuidas equitativamente entre las clases.
A partir de ello, se aplica Random Forest y se encuentra la configuración de parámetros, buscando disminuir la complejidad del modelo obteniendo el menor error. Los resultados del modelo construido se encuentran en la Figura 3.1, con 105 árboles y 4 predictores considerados en cada división (*ntry)*, según los resultados presentados en la Figura 3.4.

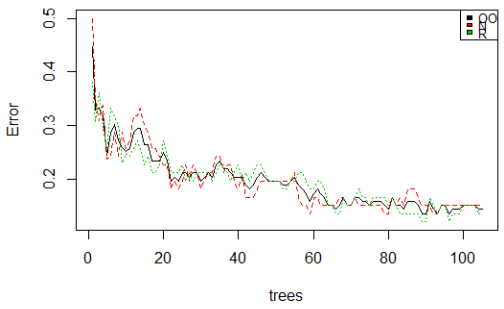

Figura 3.1*: Comportamiento del error según número de arboles*

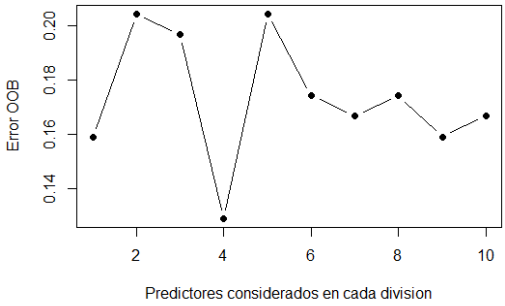

Figura 3.2*: Comportamiento del error según predictores en cada división*

Una vez obtenido el modelo anterior, la importancia de las variables dadas por el modelo de Random Forest se presenta en la Figura 3.3.

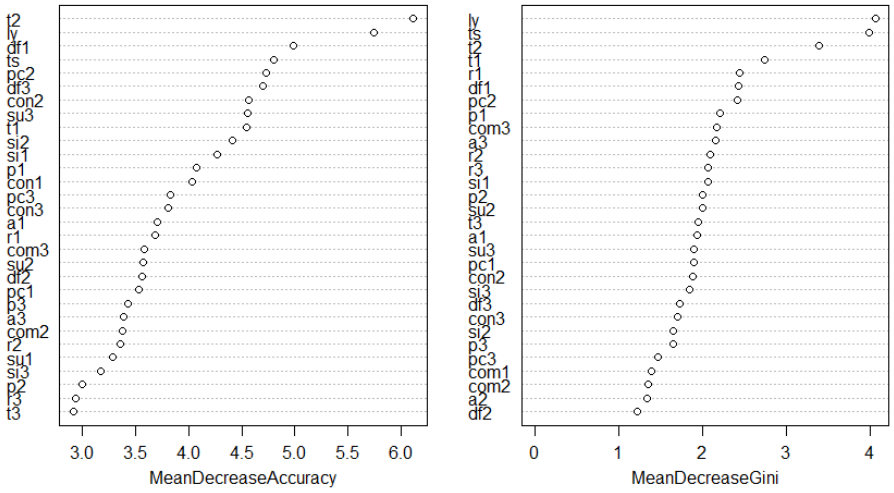

***Figura 3.3**: Importancia de variables*

Se puede observar a partir de allí, que si bien hay cierta similitud entre ambos criterios, esta no es concluyente, por lo cual se procede a obtener los modelos de clasificación para cada uno de los criterios.

A partir del criterio de Mean Decrease Accuracy (MDA), se ordenan las variables y se procede a probar iterativamente la construcción del modelo con Random Forest de manera de obtener el número de variables que otorga un menor error, lo que se refleja en la Figura 3.4. En base a ello, buscando cumplir con el principio de parsimonia se conservan las primeras 9 variables ordenadas según este criterio de importancia, y se elabora el modelo representado en la Figura 3.5, con 82 árboles y el predictor divisor igual a 4.

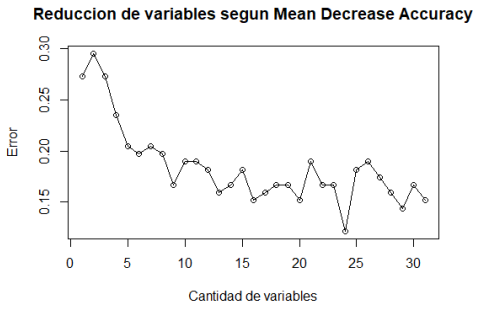

***Figura 3.4.** Reducción de variables según Mean Decrease Accuracy y RF*

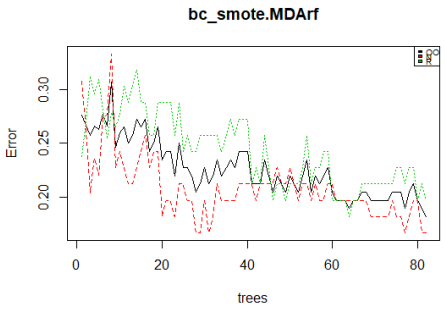

***Figura 3.5**: Comportamiento del error según número de árboles dado MDA*

El modelo generado se presenta en la Tabla 3.3

Tabla 3.3: *Modelo de Clasificación según reducción de variables con MDA.*

	N	**R**	**Error de la clase**
N	55	11	0.1667
R	16	50	0.2424
OOB tasa de error estimada = 20,45%			

Según dicha configuración, se elabora un gráfico de coordenadas paralelas a manera de obtener mayor información de las variables seleccionadas.

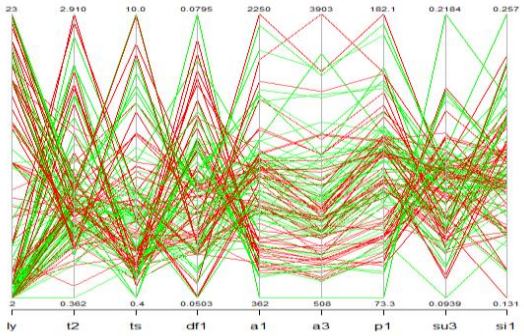

Figura 3.6*: Coordenadas paralelas de las variables Según MDA y RF.*

Luego, mediante el criterio de importancia de Mean Decrease Gini (MDG), se realiza el mismo procedimiento anterior, conservándose las 7 primeras variables dado dicho orden, según lo que se presenta en la Figura 3.7. En base a ello, se genera el modelo presentado en la Figura 3.8, con 78 árboles y el predictor divisor igual a 4.

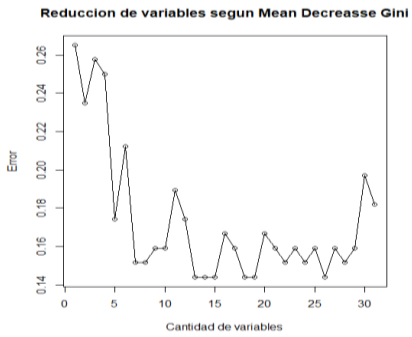

Figura 3.7*: Reducción de variables según Mean Decrease Gini y RF.*

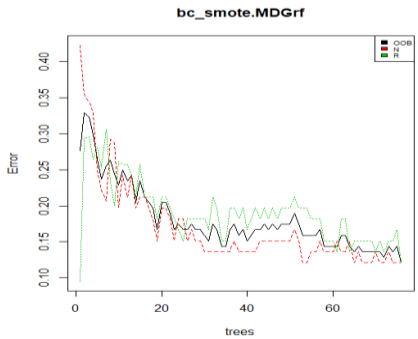

Figura 3.8*: Comportamiento del error según número de árboles dado MDGini.*

El modelo generado bajo este criterio presenta los siguientes resultados.

Tabla 3.4: *Modelo de Clasificación según reducción de variables con MDGini.*

	N	R	Error de la clase
N	58	8	0.1212
R	8	58	0.1212
OOB tasa de error estimada = 12,42%			

Por último, a manera de comparación se presenta el gráfico de coordenadas paralelas dada la configuración de variables obtenida.

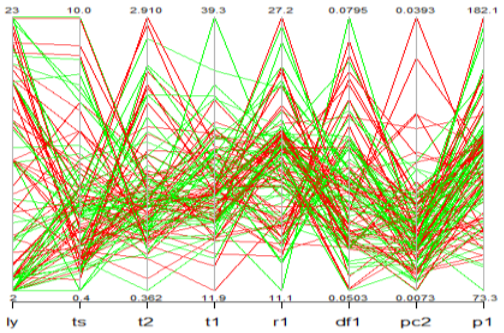

Figura 3.9*: Coordenadas paralelas de las variables Según MDGini y RF.*

4 Discusión

Al comparar las técnicas utilizadas para el balanceo de datos mediante AUC ROC y AUPRC, Tabla 3.1, se observa que todas las técnicas presentan un rendimiento bajo, destacándose la técnica de submuestreo (under) como aquella que tiene la mayor recuperación temprana de precisión y un símil con respecto a las otras técnicas del área bajo la curva ROC. Sin embargo, hay que considerar que la principal desventaja del submuestreo es que se pierde información potencialmente relevante de las muestras descartadas, cosa que se puede evidenciar en el alto error del modelo de Random Forest y su clasificación. Por otro lado, la que tiene mejor rendimiento según la curva ROC, es la técnica Rose, pero a su vez, presenta la menor recuperación temprana de precisión y no muy buenos resultados en el Random Forest.
Un punto importante a destacar es que todas las técnicas disminuyen en el AUC-ROC con respecto a los datos originales, por lo cual se recurre a considerar los resultados de cada técnica para Random Forest, de donde se escoge finalmente la técnica smote para el balance de los datos.

A partir de la importancia según el criterio de Mean Decrease Acuracy, se obtienen que las variables más importantes corresponden a la desviación estándar de la textura, estado de los ganglios linfáticos, tamaño del tumor, dimensión fractal media, área media, peor área, perímetro medio, peor suavidad y la simetría media. Por otro lado, según el criterio de Mean Decrease Gini, las variables más importantes son el estado

de los ganglios linfáticos, tamaño del tumor, desviación estándar de la textura, textura media, radio medio, dimensión fractal media, y desviación estándar de los puntos cóncavos. De ello, se puede desprender cierta diferencia entre ambos criterios, lo que puede ser debido a que las variables del conjunto de datos tienen escalas muy diferentes en sus rangos, y a que se debe tener en cuenta que Mean Decrease Gini mide la pureza de la variable, mientras que Mean Decrease Accuracy mide la idoneidad de la variable para ser un predictor.
Una vez aclarado esto, los resultados de los modelos de Random Forest bajo ambos criterios, muestran un menor error de clasificación para Mean Decrease Gini y a su vez un menor error OOB del modelo en comparación con el criterio Mean Decrease Accuracy. Sin embargo, al obtener los gráficos de las coordenadas paralelas respectivos, no es posible obtener un conocimiento substancial de la diferencia de las variables para cada una de las clases.
Finalmente, al comparar la clasificación obtenida con Random Forest para los datos originales desbalanceados con respecto los balanceados con smote, se observa como disminuye el error del modelo con estos últimos, disminuyendo a su vez considerablemente el error de clasificación, especialmente de la clase recurrente, que era la minoritaria.

5 Conclusiones

En la experimentación con el método de Random Forest, se pudo comprender de mejor manera su funcionamiento y se pudo aprovechar la evaluación de las variables según la importancia, provista por el mismo método.
Random Forest, de forma general, reduce el sesgo gracias a las ventajas provistas por la técnica de Boosting y reporta mejoras respecto a otros algoritmos como los arboles de decisión, sin embargo, esto no es siempre así. Se debe tener en cuenta el principio de parsimonia, por lo que la elección de Random Forest debe ser bien justificada, ya que la complejidad de este modelo en comparación a los árboles de decisión aumenta considerablemente.
La mayoría de los algoritmos de clasificación de aprendizaje automático son sensibles al desequilibrio en las clases, y Random Forest no es la excepción. Un modelo de aprendizaje automático entrenado y probado en un conjunto de datos de este tipo, dado que la no recurrencia es la clase más común, se evidencia que el modelo clasifica como no recurrente la mayoría de las muestras, aun cuando la precisión es muy alta. Es entonces que se puede afirmar que un conjunto de datos desbalanceado sesgará el modelo de clasificación hacia la clase más común.
Al explorar las técnicas de balance de datos, se pudo evidenciar las ventajas y desventajas que tienen cada una, explorando cual se ajusta mejor para este conjunto de datos en específico, y evitando al final el sesgo del modelo.
Según los resultados, es importante destacar que los métodos comprendidos actuarán de una forma deseada siempre y cuando el conjunto de datos esté considerando las cualidades o atributos que permitan obtener una correcta clasificación.

6 Referencias

1. Chacón, M. Taller de minería de datos avanzada, Capitulo II: Bosques Aleatorios.
2. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
3. Package 'RandomForest', https://cran.r-project.org/web/packages/randomForest/randomForest.pdf
4. William H. Wolberg. Computer-Derived Nuclear "Grade" and Breast Cancer Prognosis. Departments of Surgery, Human Oncology, and Computer Sciences University of Wisconsin, Madison.
5. Wicked Good Data, Handling Class Imbalance with R and Caret. URL https://www.r-bloggers.com/handling-class-imbalance-with-r-and-caret-caveats-when-using-the-auc/