PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Desempeño predictivo de los métodos regresión binaria potencia logística y bosque aleatorio en clasificación desbalanceada

Tesis para obtener el grado académico de Maestro en Estadística que presenta:

AUTOR

Javier Santiago Maraví Zegarra

ASESOR

Dr. Alex de la Cruz Huayanay

Lima, 2025

Declaración jurada de autenticidad

Yo, Alex de la Cruz Huayanay, docente de la Escuela de Postgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada Desempeño predictivo de los métodos regresión binaria potencia logística y bosque aleatorio en clasificación desbalanceada, del autor Javier Santiago Maraví Zegarra, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 8 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 25/08/2025.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 25 de agosto de 2025

Apellidos y nombres del asesor: Alex de la Cruz Huayanay	
DNI: 46121900	**Firma:** 
ORCID: https://orcid.org/0000-0003-0746-0803	

Dedicatoria

A mi mamá.

Agradecimentos

A Dios, siempre a El.

A mi papá por su constante aliento, a mi mamá que partió a la presencia de Dios dejándome un maravilloso legado de amor y fe.

A mi amada esposa Vanessa, a mis pequeños Santiago y Belén quienes tuvieron que aceptar, sin habérselo propuesto, el desafío que significó esta maestría.

A mis hermanos por el ejemplo de constancia y valentía que suelen transmitirme.

Al Dr. Alex de la Cruz Huayanay, mi asesor de tesis, quien no tuvo inconvenientes en dedicar muchas horas de asesoría y compartir conmigo todo su conocimiento y experiencia sobre el tema de mi trabajo. Su buena disposición, sus recomendaciones y el apoyo permanente fueron fundamentales para la culminación exitosa de esta investigación.

Resumen

Los métodos de clasificación binaria fueron diseñados bajo el supuesto de que las clases o categorias se encuentran balanceadas. Sin embargo, en la realidad observamos que las clases están desbalanceadas, esto es, hay una clase que aparece con mayor frecuencia que la otra afectando, en consecuencia, la capacidad predicitiva de los métodos de clasificación.

Para hacer frente al problema del desbalance encontramos métodos no paramétricos como el bosque aleatorio que de acuerdo a recientes estudios es el que mejor desempeño ha mostrado en su capacidad predictiva cuando se le comprara con otros métodos de aprendizaje automático. Por el lado, de los métodos paramétricos las distribuciones potencia y reversa de potencia aplicadas como funciones de enlace en métodos de regresión binaria muestran un adecuado desempeño predicitivo cuando hacen frente al problema de clases desbalanceadas.

Sin embargo en la literatura no existe un estudio que compare el desempeño predicitivo de los métodos paramétrico y no paramétrico. Esta investigación tiene ese objetivo, comparar y determinar cuál de éstos presenta la mejor performance predictiva. Para esta comparación utilizamos distintas métricas de desempeño eligiendo aquellas que resulten idóneas considerando el nivel del desbalance observado.

Los métodos utilizados son el bosque aleatorio, por el lado no paramétrico, y la regresión logística con función de enlace potencia, desde la perspectiva paramétrica. En principio, se hizo un estudio de simulación condiderando escenarios con distintos niveles de desbalance para diferentes tamaños de muestra. Luego, aplicamos los métodos a una situación real para lo cual utilizamos información de estudiantes de una institución de educación superior. Tanto para el estudio de simulación como en la aplicación real los resultados muestran que es la regresión logística con función de enlace potencia la que mejor desempeño obtuvo en términos de predicción.

Palabras clave: Desbalance de datos, bosque aleatorio, distribución logística, enlace potencia y reversa potencia, medidas de precisión.

Abstract

Binary classification methods were originally developed under the assumption that the classes or categories are balanced. However, in real world scenarios, class imbalance is frequently observed. In other words, one class occurs significantly more often than the other. This imbalance adversely affects the predictive performance of classification methods.

To address the issue of imbalance, non-parametric methods such as Random Forests have been employed. According to recent studies, Random Forests exhibit superior predictive performance when compared to other machine learning algorithms. On the parametric side, the power and reversed power distributions, when used as link functions in binary regression models, have demonstrated adequate predictive capabilities in the context of imbalanced data.

Nonetheless, the literature lacks comprehensive research comparing the predictive performance of parametric versus non-parametric methods. The aim of this research is to fill that gap by conducting a comparative analysis to determine which approach yields better predictive performance. For this purpose, we employ a variety of accuracy measures, selecting those most suitable for assessing models under different levels of class imbalance.

The methods employed in this study include Random Forest, representing the non-parametric approach, and logistic regression with a power link function, representing the parametric perspective. Initially, a simulation study was conducted under various scenarios, considering different levels of class imbalance and sample sizes. Subsequently, the methods were applied to a real-world dataset comprising student data from a higher education institution. In both the simulation and the empirical application, logistic regression with a power link function demonstrated superior predictive performance compared to Random Forest.

Keywords: Data imbalance, Random Forest, Logistic Distribution, Power and Reversed Power link functions, accuracy measures.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Planteamiento y justificación del tema

Uno de los desafíos principales de los métodos de clasificación binaria es el desequilibrio de los datos, donde una clase aparece con mucha más frecuencia que la otra. Este desequilibrio puede provocar que el clasificador tenga un rendimiento predictivo deficiente al capturar inadecuadamente la información de la clase minoritaria que, por lo general, suele ser la más relevante en muchas aplicaciones. Es importante destacar, como menciona Ali et al. (2013), que existen diferentes grados de desequilibrio, desde casos muy severos hasta conjuntos de datos perfectamente equilibrados pero en muchos casos esta situación no garantiza, necesariamente, un rendimiento óptimo del clasificador.

Según Ali et al. (2013), los algoritmos de aprendizaje automático han sido diseñados bajo la suposición de que las clases están equilibradas, cuando en realidad esto no siempre suele ser así. Esta discrepancia entre la suposición y la realidad conlleva inconvenientes en los resultados obtenidos, tanto en términos de desempeño predictivo como en las estimaciones de los parámetros en modelos de regresión. En ese sentido, como menciona Wu y Li (2024), los modelos de aprendizaje automático enfrentan limitaciones en su rendimiento predictivo cuando una clase (generalmente la clase minoritaria) está sub representada en comparación con la otra clase (la mayoritaria). Aunque el modelo puede mostrar un rendimiento global aceptable, tiende a clasificar incorrectamente a la clase minoritaria. Jafarigol y Trafalis (2023) añaden que este problema adquiere mayor relevancia cuando es la clase minoritaria la de interés en el estudio y, como suele ocurrir, esté asociada con eventos raros, entendiéndose a éstos como eventos cuya ocurrencia es significativamente menor que la de los eventos comunes conduciendo a que el clasificador, influenciado por la clase mayoritaria, produzca resultados sesgados.

Con el objetivo de mejorar el rendimiento del clasificador han sido propuestos diferentes métodos para resolver el problema de clasificación desbalanceada. Por un lado, es abordado a través de los algoritmos de aprendizaje automático, como el Bosque Aleatorio (RF: del inglés Random Forest), Support Vector Machine (SVM), así como métodos de muestreo (Qaddoura y Biltawi, 2022). Por otro lado, están los modelos de regresión binaria con enlaces asimétricos (de la Cruz Huayanay et al., 2019), los cuales intentan captar el desbalance de clases a través de un parámetro.

A pesar de los avances en el estudio del problema y de las metodologías propuestas para abordarlo, Wu y Li (2024) sostienen que sería inexacto afirmar que el problema de clasificación desbalanceada no ha sido exhaustivamente estudiado. De hecho, sostienen que aún existen numerosos desafíos por resolver en este campo.

Entre los métodos de aprendizaje automático, el método de Bosque Aleatorio, junto con la técnica de sobremuestreo mejora de forma significativa el desempeño del modelo (Bharadwaj, 2023).

En el contexto de los modelos de regresión para variables de respuesta binaria los modelos más comunes usan enlaces simétricos. Sin embargo, no son los idóneos cuando los datos presentan el problema de desbalance. En esa línea, Chen et al. (1999), sostiene que si la probabilidad de la variable de respuesta binaria se aproxima a cero a una tasa diferente de la que se aproxima a uno, no es apropiado utilizar enlaces simétricos, como es el caso de la regresión logística. Frente a esto, Bazán et al. (2017) y de la Cruz Huayanay et al. (2019) proponen el uso de funciones de enlaces asimétricos basados en distribuciones potencia y reversa de potencia para mejorar el ajuste del modelo. Estos enlaces han presentado un buen desempeño puesto que controlan la tasa de incremento o disminución de la probabilidad de éxito o fracaso de la variable de respuesta binaria. Así, Bazán et al. (2017) y de la Cruz Huayanay et al. (2019) muestran que los enlaces potencia y su reversa en regresión binaria presentan buen desempeño para determinados grados de desbalance. Un caso particular de estos modelos es la regresión binaria con enlace potencia logística, la cual es una generalización de la regresión logística.

Por lo tanto, es relevante estudiar estos dos tipos de métodos de clasificación binaria (bosque aleatorio y regresión binaria con enlace potencia logística), dado que no existe en la literatura investigaciones dedicadas al análisis comparativo del desempeño predictivo en el contexto de clases desbalanceadas. Además, porque el problema del desbalance de clases se presenta en la práctica con frecuencia en diferentes áreas. Por ejemplo, en la medicina, las finanzas, los seguros, educación, el deporte, entre otras; siendo que la clase minoritaria es, en la mayoría de las veces, la de interés.

1.2. Objetivos

El objetivo de esta investigación es el estudio comparativo del desempeño predictivo de la regresión binaria con enlace asimétrico potencia logística y el método Bosque Aleatorio en la clasificación de datos desbalanceados.

De forma específica, se pretende:

- Revisar y estudiar el método de Bosque Aleatorio y el modelo de regresión binaria con función de enlace potencia logística propuesto por Bazán et al. (2017) y de la Cruz Huayanay et al. (2019), como alternativas para clasificación binaria desbalanceada.
- Realizar la estimación de los modelos de regresión binaria con función de enlace asimétrico, bajo un enfoque bayesiano.

- Realizar un estudio de simulación que permita evaluar el desempeño predictivo de los modelos propuestos.
- Aplicar los métodos estudiados a un conjunto de datos reales vinculados a la deserción universitaria.
- Usar diferentes métricas para evaluar el desempeño predicitvo de los métodos estudiados.

1.3. Organización del trabajo

En el capítulo 2, profundizaremos sobre el problema del desbalance desarrollando en detalle algunos conceptos relacionados con el tema, haremos una exploración de recientes trabajos de investigación que nos permitan saber cuáles son los modelos de aprendizaje automático (no paramétricos) y las métricas de evaluación que se disponen para la evaluación predictiva, en el contexto de datos no balanceados. El capítulo 3, se prestará especial atención a las propiedades y principales características del Bosque Aleatorio en su versión estándar. El capítulo 4, presentamos una nueva alternativa para hacer frente al problema de desbalance. Específicamente, el modelo de Regresión Potencia Logística propuesto por Bazán et al. (2017) y de la Cruz Huayanay et al. (2019) referido a métodos paraméticos de regresión con enlace asimétrico y plantearemos su estimación bajo el enfoque bayesiano. En el capítulo 5, haremos las simulaciones para cada modelo y evaluaremos sus resultados. En el capítulo 6, aplicaremos los métodos bajo estudio a un conjunto de datos relacionados con Deserción Universitaria, analizaremos y compararemos los resultados obtenidos usando distintas métricas de evaluación. Finalmente, en el capítulo 7, presentamos las principales conclusiones del presente estudio y propondremos algunas recomendaciones para futuras investigaciones.

Capítulo 2

Conceptos Generales

2.1. Introduccion

La gran cantidad de datos disponible trajo consigo grandes desafíos a los investigadores de aprendizaje automático (del inglés machine leraning) y minería de datos (del inglés, data mining) según señala Abd Elrahman y Abraham (2013). Uno de ellos, tema central de esta investigación es el referido al desbalance que muestran las clases del conjunto de datos en evaluación, afectando directamente el desempeño predictivo de los métodos de clasificación.

El problema aparece cuando observamos en la práctica que las clases del conjunto de datos bajo estudio no están distribuidas en forma equilibrada, es decir, hay una clase que prevalece sobre la otra. Las consecuencias que trae este problema son múltiples entre los cuales tenemos:

- Cuando las clases están desbalanceadas la capacidad predictiva se ve afectada. Además si el desbalance es extremo la mala clasificación se intensifica (Jafarigol y Trafalis (2023)). A pesar de que este tipo de situaciones aparecen con frecuencia el tema del desbalance ha sido poco tratado, menos aún el desbalance severo o extremo (Krawczyk (2016)).
- El problema se agrava cuando la clase minoritaria está referida a casos o eventos raros. Jafarigol y Trafalis (2023) definen estos eventos como aquellos que presentan una frecuencia de ocurrencia significativamente menor respecto del evento con los que se compara. En esa línea, Cicak y Avci (2023) afirman que si el objetivo es identificar, por ejemplo, si un determinado sujeto padece o no de una enfermedad rara [1], el costo de una mala clasificación (clasificar, por ejemplo, a una persona sana cuando en realidad está enferma) podría ocasionar graves consecuencias. Otro ejemplo de desbalance severo de clases lo encontramos en la clasificación de imágenes de mamografías para la detección de cáncer. Por lo general, una mamografía normal presenta el 98 % de pixeles normales y el 2 % anormales Chawla (2010).
- En un conjunto de datos desbalanceados los algoritmos de aprendizaje automático pueden obtener una adecuada capacidad de predicción global pero no sucede lo mismo cuando se observa

[1] Enfermedad rara es aquella cuya prevalencia es inferior a 5 casos por cada 10.000 personas

el desempeño predictivo de la clase minoritaria, afirma Rekha et al. (2021).

- En un conjunto de datos, demás del desbalance, se pueden presentar otros problemas. Por ejemplo, el nivel de separabilidad que pueden mostrar las clases o la superposición (overlapping, en inglés), que de acuerdo a Abd Elrahman y Abraham (2013) también afectan la capacidad predictiva del algoritmo.

2.2. Clasificación binaria

En muchas situaciones reales, la variable de interés o variable respuesta presenta dos posibles clases mutuamente excluyentes. Por ejemplo, en el caso de un crédito bancario la variable de interés puede ser la conducta de pago del prestatario, es decir si éste pagará o no su obligación financiera, el método de regresión o clasificación binaria deberá discriminar entre dos clases (i) pagó o (ii) no pagó. Por convención, la clase que representa el evento de interés (éxito) asume el valor 1 y se le denomina clase positiva. Por otro lado, la clase denominada clase negativa asume el valor 0 y representa al evento complementario (fracaso). En Véliz Capuñay (2023) se señala que:

Definición 1. *La clasificación binaria ocurre cuando el algoritmo discrimina a las clases de la variable respuesta en función de una regla de clasificación que indica que un elemento determinado por* $X^t = (X_1, \ldots, X_q)$ *pertenece a la clase* $Y = 1$ *si* $P(Y = 1 \mid X) \geq t_0$ *para el umbral* $t_0 > 0$*, y que pertenece a la clase* $Y = 0$ *si* $P(Y = 1 \mid X) < t_0$*, donde,* X^t *es el vector de variables independientes y* Y *es la variable respuesta que puede tomar los valores* 0 *ó* 1*.*

La ocurrencia de alguna de las clases de la variable respuesta (en nuestro ejemplo "pagó" o "no pagó") depende de sus características particulares, éstas se observan en las variables independientes (conocidas también como variables explicativas o covariables) las cuales pueden ser continuas, discretas o categóricas.

Ahora, para lograr la clasificación binaria existen diferentes métodos, tanto paramétricos (regresión binaria) y no paramétricos (bosques aleatorios, árboles de decisión, redes neuronales y otros). En particular la regresión logística es uno de los métodos más utilizados que busca obtener el mejor ajuste entre una variable dependiente y las covariables que la explican.

En general, para obtener la probabilidad de pertenecer a una determinada clase dado una o más variables explicativas, se pueden utilizar los métodos mencionados.

2.3. Desbalance de datos en clasificación binaria

En línea con de la Cruz Huayanay (2023), diremos que una variable respuesta Y está desbalanceada, si y sólo si, $\kappa := |2\mu - 1| \geq 0.2$, donde μ es una probabilidad de éxito.

Queda evidenciado en de la Cruz Huayanay et al. (2019) que, en situaciones de datos desbalanceados, el uso de funciones de enlace simétricas no es adecuado. Por esta razón, debe considerarse el uso de funciones de enlace asimétricas.

2.4. Evaluación de desempeño predictivo de los métodos

Hemos mencionado que los modelos de clasificación de aprendizaje automático fueron diseñados bajo el supuesto de que las clases del conjunto de datos están balanceadas, en consecuencia, las métricas que se utilizan para la evaluación de su desempeño no toman en consideración lo contrario, es decir que las clases de datos se encuentran en la mayoría de las ocasiones desbalanceada. En ese sentido, es imprescindible contar con indicadores que permitan medir el desempeño predictivo del clasificador en el contexto de clases desbalanceadas más aún cuando el interés se encuentra alrededor de la clase minoritaria[2].

Sobre las métricas de evaluación Weng y Poon (2008)[3] afirman que la evaluación del rendimiento a datos no equilibrados debe realizarse a cada clase de manera individual. Cuando el análisis de la capacidad predictiva del algoritmo se realiza de forma global, como es el caso de la precisión global, sólo se refleja el desempeño de la clase mayoritaria en perjuicio de la minoritaria, situación que se agrava cuando el desequilibrio es más severo. El costo de una mala clasificación de un dato perteneciente a la clase minoritaria es mayor que un error de clasificación de un dato que corresponde a la clase mayoritaria.

De la misma manera en Abd Elrahman y Abraham (2013) opinan que las métricas de evaluación son un punto crítico en el campo del aprendizaje automático pues éstas permiten evaluar el desempeño de los algoritmos y que frente a conjunto de datos con presencia de clases desbalanceadas las métricas de evaluación comunes no son las apropiadas.

Guo et al. (2008) agrupa las métricas de evaluación en dos categorías[4], las que se derivan de la matriz de confusión como la precisión global, precisión, sensibilidad, tasa de falsos positivos, la tasa de verdaderos positivos, la curva ROC y el AUC y, un segundo grupo, que se construyen sobre la base de la precisión y la sensibilidad (recall) como el F1 y el G-mean. [5]

2.4.1. La Matriz de Confusión

La matriz de confusión es una herramienta que brinda información relevante sobre la eficacia predictiva del método de clasificación. Se registran en sus entradas los éxitos o fracasos que ocurren cuando se aplica a un conjunto de datos un método de clasificación. Es útil tanto para problemas de clasificación binaria como para clasificación multiclase.

[2]Para todos los efectos, asignaremos a la clase minoritaria el número 1 y a la clase mayoritaria el 0

[3]Los autores proponen en su investigación una nueva métrica de evaluación la AUC ponderada y concluyen que es una mejor alternativa para evaluar conjuntos de datos desbalanceados. En la presente investigación no utilizaremos este indicador.

[4]Jafarigol y Trafalis (2023) incorporan una tercera categoria que son aquellas métricas que se enfocan en la predicción de la probabilidad de una clase (Probabilty evaluation metrics)

[5]De acuerdo a los autores, el F1 es la métrica de evaluación con mayor uso para medir el desempeño predictivo de un clasificador cuando enfrentamos el problema de desbalance.

Una matriz de confusión, para una base de datos de dos clases, es:

Cuadro 2.1: Matriz de Confusión

	Valor Predicho	
Valor Observado	1	0
1	Verdaderos Positivos (VP)	Falsos Negativos (FN)
0	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Donde VP representa el número de clases positivas clasificadas correctamente, VN el número de clases negativas clasificadas correctamente, FN el número de clases positivas mal clasificadas y FP el número de clases negativas mal clasificadas.

2.4.2. Métricas de evaluación

A partir de la matriz de confusión se derivan diversas métricas que permiten la evaluación del desempeño del algoritmo. Así tenemos:

- **Precisión Global (Accuracy)**

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.1}$$

- **Tasa de Verdaderos Positivos (TPR) / Sensibilidad (Sensitivity o Recall)**: es la proporción de datos positivos del conjunto de evaluación correctamente clasificados. Mide el desempeño del algoritmo respecto de la clase minoritaria.

$$Sensibilidad = \frac{VP}{VP + FN} \tag{2.2}$$

- **Tasa de Verdaderos Negativos (TNR) / Especificidad** (Specifity): es la proporción de datos negativos del conjunto de evaluación clasificados de forma correcta.

$$Especificidad = \frac{VN}{VN + FP} \tag{2.3}$$

- **Precisión**: Proporción de datos de la clase minoritaria pertenecientes al conjunto de evaluación correctamente clasificados. Una baja precisión es síntoma de que existen un gran número de falsos positivos (FP).

$$Precision = \frac{VP}{VP + FP} \tag{2.4}$$

- $F1$: Es una métrica que incorpora a la especificidad y a la sensibilidad por lo que se ve afectada por el trade off que existe entre estos indicadores.

$$F1 = \frac{2 \times Especificidad \times Sensibilidad}{Especificidad + Sensibilidad} \tag{2.5}$$

- **G-mean** (Media Geométrica): es una métrica utilizada ante la presencia de clases desbalanceadas. Un valor alto de este indicador significa que el modelo muestra un buen desempeño en ambas clases.

$$G - mean = \sqrt{Sensibilidad \times Especificidad} \tag{2.6}$$

De acuerdo con Abd Elrahman y Abraham (2013) para los métodos de aprendizaje automático, las métricas de evaluación relacionadas con clases desbalanceadas son las ecuaciones 2.2, 2.3, 2.4, 2.5 y 2.6. Para monitorear el desempeño individual de cada clase se utiliza las ecuaciones 2.2 y 2.3. Cuando el interés es tener un alto desempeño de sólo una clase se emplea la ecuación 2.4; si el interés está enfocado en obtener un alto rendimiento de ambas clases se prefiere las ecuaciones 2.5 y 2.6.

En de la Cruz Huayanay et al. (2019), de la Cruz Huayanay (2023) y en de la Cruz Huayanay et al. (2024) se indica que las métricas ACC, TPR y TNR no son adecuadas para determinar la capacidad predictiva de un modelo de regresión binaria cuando las clases presentan desbalance. Por tal razón, en las investigaciones citadas se propone considerar otras métricas que permitirán una adecuada evaluación del desempeño predictivo cuando son aplicadas a métodos de predicción en un contexto de clases desbalanceadas. Estas métricas se muestran en el cuadro 2.2.

Cuadro 2.2: Métricas para evaluación del desempeño predictivo aplicados a métodos de regresión binaria con clases desbalanceadas

Métrica	Notación	Fórmula	Rango de valores
Índice crítico de éxito	CSI	$\frac{TP}{TP+FP+FN}$	[0 ; 1]
Índice de Sokal y Sneath	SSI	$\frac{TP}{TP+2\times FP+2\times FN}$	[0 ; 1]
Índice de confianza	FAITH	$\frac{TP+0.5\times TN}{TP+FP+FN+TN}$	[0 ; 1]
Diferencia estándar	PDIF	$\frac{4\times FP\times FN}{(TP+FP+FN+TN)^2}$	[0 ; 1]
Puntuación de habilidades de Gilbert	GS	$\frac{(TP\times TN-FP\times FN)}{(FN+FP)(TP+FP+FN+TN)+(TP\times TN-FP\times FN)}$	[0 ; 1]
Coeficiente de Correlación de Matthews	MCC	$\frac{(TP\times TN-FP\times FN)}{\sqrt{(FN+FP)(TP+FN)(TN+FP)(TN+FN)}}$	[0 ; 1]
Capa de Cohen	KAPPA	$\frac{2\times(TP\times TN-FP\times FN)}{(TP+FP)(FP+TN)+(TP+FN)(FN+TN)}$	[0 ; 1]

Con excepción de PDIF, para el resto de métricas, se debe preferir el método que presente el mayor valor sobre otros métodos posibles porque de esta forma aseguramos un alto nivel de coincidencia entre el dato observado y el predicho, según señala de la Cruz Huayanay et al. (2024).

2.4.3. La curva ROC y el AUC

La curva ROC es una herramienta que permite evaluar el desempeño de un clasificador considerando la compensación (tradeoff) que existe entre la Tasa de Verdaderos Positivos (TPR, por sus siglas en inglés) o sensitividad, y la Tasa de Falsos Positivos (1- especificidad).

La curva se va construyendo considerando distintos puntos de corte o umbrales de discriminación a cada uno de éstos le corresponderá , en consecuencia, un determinado valor de TPR y de la Tasa de Falsos Positivos .

Siguiendo a Véliz Capuñay (2023) en un modelo de clasificación binaria, es decir aquel que sólo dispone de dos categorías ($Y = 0$ ó $Y = 1$) la curva ROC permite determinar de manera adecuada el

punto de corte o umbral de discriminación entre las categoria. Señala, que el umbral o punto de corte adecuado para la clasificación es el que produce medidas de sensibilidad y especificidad para los cuales el punto que corresponde en la curva ROC está localizado a la izquierda y lo más alto posible; este corresponde al punto de máxima curvatura y es donde se produce el equilibrio entre la especificidad y sensibilidad.

La representación gráfica de la curva ROC se muestra en la figura 2.1, de acuerdo a ella el umbral ideal es el que alcanza el punto (0,1), aquí el método ha clasificado correctamente a todas las instancias de la clase positiva (+1) y, en consecuencia, su tasa de falsos positivos es cero.

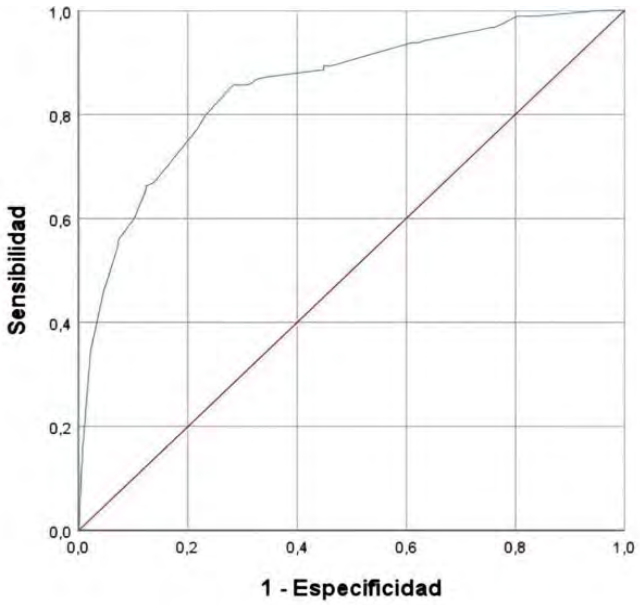


Figura 2.1: Curva ROC

2.5. Técnicas de remuestreo

Los métodos de aprendizaje automático abordan el problema de desbalance aplicando técnicas de remuestreo. Con el uso de estas técnicas lo que se pretende es equilibrar a las clases mediante la aplicación de técnicas de remuestreo (resampling) las que se clasifican en (i) técnicas de sobremuestreo (del inglés, oversampling) y (ii) técnicas de submuestreo (del inglés, undersampling). La finalidad de aplicar estas técnicas a la base de datos es lograr un adecuado balance de las clases, de tal forma que el desempeño predictivo del clasificador mejore.

En Cicak y Avci (2023) encontramos las siguientes definiciones:

Tecnicas de sobremuestreo, son aquellas que se utilizan para incrementar el número de instancias de la clase minoritaria hasta nivelarla con la clase mayoritaria, entre las cuales se tiene:

Sobremuestreo Aleatorio (Random Oversampling - ROS), permite alcanzar el balance entre clases replicando aleatoriamente datos de la clase minoritaria. Tiene como principal desventaja que en su aplicación se presente el problema de sobreajuste (overfitting).

SMOTE (Synthetic Minority Oversampling Technique), Es una técnica que genera de manera uniforme datos sintéticos en todo el espacio de características a partir de los datos existentes de la clase minoritaria. Evita el problema del sobreajuste que se presenta al aplicar el ROS.

ADASYN, esta técnica es una variante del SMOTE, su aplicación permite agregar instancias para aquellas muestras que son difíciles de clasificar de manera correcta. A diferencia del SMOTE que genera datos uniformemente, en ADASYN se asigna una ponderación a cada dato de la clase mminoritaria en función del nivel de dificultad de clasificación,

Técnicas de submuestreo, A diferencia del ROS esta es una técnica que reduce las instancias de la clase mayotiraria hasta nivelarla con las clase minoritaria. La principal desventaja de la aplicación de esta técnica es la pérdida de información valiosa.

Random Undersampling (RUS), es el retiro aleatorio de instancias pertenecientes a la clase mayoritaria hasta que el conjunto de datos se encuentre balanceado. A pesar de ser una técnica fácil y computacionalmente eficiente, su aplicación puede provocar la pérdida de información relevante afectando al desempeño de la clasificación.

Cluster Centroids fue propuesto para superar las limitaciones del RUS pues evita la pérdida de datos útiles de la clase mayoritaria. De acuerdo a esta técnica se agrupan las instancias de la clase mayoritaria en clusters para luego ser reemplazadas por su centroide.

2.6. Áreas de aplicación de los métodos en la literatura

En esta sección presentamos una revisión de algunas investigaciones que han abordado el problema del desbalance de datos en tres campos de estudio, las finanzas, la medicina y la educación. La idea de este apartado es identificar aquellos algoritmos utilizados en estas investigaciones que alcanzaron el mejor desempeño predictivo.

2.6.1. Aplicación a las finanzas

En el campo de las finanzas, los modelos de aprendizaje automático han sido utilizados desde varios enfoques Planinić y Popović-Bugarin (2024) sostiene que éstos han tenido un excepcional éxito afrontando diferentes desafíos vinculados a la actividad bancaria. Estos autores, examinan el desempeño de los que a su juicio son los algoritmos más populares en la detección de fraudes con tarjeta de crédito, esto son, la Regresión Logística, Bosque Aleatorio y Catboot apuntando que la dificultad en

la detección de este tipo de actividad fraudulenta se debe, entre otras razones[6], a que las clases del conjunto de datos están desbalanceadas.

En la aplicación de la Regresión Logística y Bosque Aleatorio utilizan diferentes técnicas de muestreo a fin de evitar que los clasificadores muestren un sesgo hacia la clase mayoritaria.

Para determinar el clasificador con mejor rendimiento predictivo utilizaron métricas estándar, la precisión global, precisión, sensibilidad y F1.

De la misma forma Bharadwaj (2023) coincide con Planinić y Popović-Bugarin (2024) en el sentido que la detección de fraudes a través de tarjeta de crédito se hace cada vez más complicada debido a que los estafadores hacen uso de formas cada vez más sofisticadas haciendo parecer a las transacciones fraudulentas como genuinas. Por tanto, es necesario, afirman, la construcción de modelos que permitan la identificación temprana de estas operaciones.

Bharadwaj (2023) compara dos modelos ampliamente utilizados en la industria bancaria, el Bosque Aleatorio y K-NN, a fin de determinar cuál de los dos presenta un mejor comportamiento predictivo. Utilizando la misma base de datos de Planinić y Popović-Bugarin (2024) correspondiente a las transacciones hechas por tarjeta habientes europeos durante el mes de setiembre del 2013, teniendo como característica que las clases (legítima y fraudulenta) están altamente desbalanceadas[7]. En su caso las métricas de medición utilizadas fueron precisión global, precisión, sensibilidad, F1 y la curva ROC. El balance de las clases de datos lo hicieron mediante la aplicación de las siguientes técnicas de balanceo de datos: submuestreo aleatorio (RUS), sobremuestreo aleatorio (ROS), SMOTE y ADASYN. El resultado de su investigación mostró que el Bosque Aleatorio tiene mejor desempeño evaluado con el F1. Además, concluyen que el Bosque Aleatorio usando ADASYN tiene mejor desempeño predictivo que el mismo clasificador pero utilizando la técnica ROS.

2.6.2. Aplicación a la medicina

En el campo de la salud el uso de técnicas de aprendizaje automático vienen siendo muy productivas[8], pero su uso también enfrenta el problema real de clases desbalanceadas. Sobre este punto Cecchini et al. (2019) realizan un estudio acerca de enfermedades metabólicas (obesidad, diabetes tipo II, cáncer, ataques al corazón que pueden causar la muerte temprana) e identifican dos clases: enfermedades heredadas y enfermedades adquiridas, siendo las primeras las que presentan una prevalencia de 1 por cada 1000 personas. Para identificar nuevos genes relacionados con enfermedades metabólicas proponen resolver el problema del desbalance utilizando la técnica SMOTE como la que permitirá el balance de los datos y entrenarla utilizando el clasificador Gradient Boosting (GBC) obteniendo

[6]Según el autor, además del problema de las clases desbalanceadas existen tres características que también dificultan una adecuada predicción. La primera, relacionada a la naturaleza dinámica de la distribución de los datos debido a la aparición de nuevos e innovadores métodos de fraudes, en segundo lugar, a la estacionalidad en el uso de la tarjeta, el mayor uso se da en días festivos y vacaciones, y. finalmente, por los cambios en los hábitos de uso.

[7]La clase fraudulenta representa el 0.172 % del total de transacciones; 284.807 operaciones efectuadas en dos días.

[8]De acuerdo a Cecchini et al. (2019) las técnicas de machine learning son una herramienta poderosa y versátil que comparada con las técnicas de experimentación tradicionales resulta ser más rápida en términos de flujos de trabajo y de resultados.

un resultado para la métrica F-score de 0.82, valor que consideran adecuado para el propósito de su investigación.

Otro estudio interesante relacionado con la medicina es el de Zhao et al. (2018) quienes utilizan distintas técnicas de remuestreo como sobremuestreo aleatorio (ROS), submuestreo aleatorio (RUS) y SMOTE a un conjunto de datos no balanceados referidos a incidentes médicos como consecuencia de errores en la prescripción de tratamientos médicos, LASA (Look Alike - Sound Alike). Utilizan en su investigación a los siguientes clasificadores: Regresión Logística, Support Vector Machine (SVM) y Árboles de Decisión (Decision Tree) obteniendo como resultado que la Regresión Logística es el modelo con mejor capacidad predictiva llegando a esa conclusión luego de comparar los resultados de las métricas precisión, sensibilidad, especificidad y F1.

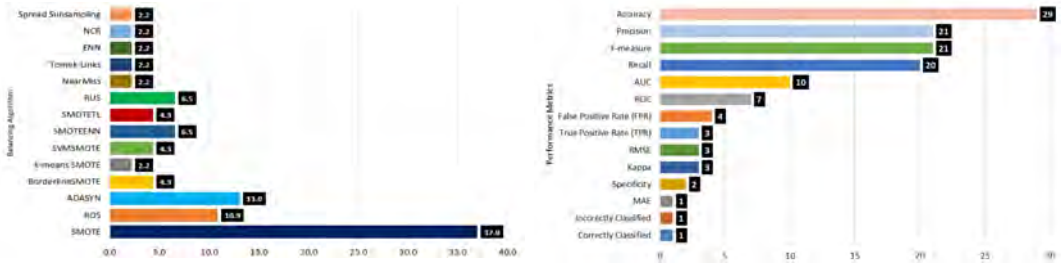
2.6.3. Aplicación a la educación

Bujang et al. (2022) realizaron una amplia y exhaustiva revisión de la literatura referida al problema del desbalance de datos aplicados a la educación. Su estudio abarcó una revisión de 43 investigaciones realizadas entre los años 2015 -2021 enfocadas en la solución de clasificación de conjuntos de clases desbalanceadas. Los autores presentan el estado del arte de cómo se ha enfrentado el problema del desbalance mostrando las mejores prácticas para el análisis de los datos, metodologías y análisis comparativos de los algoritmos propuestos. Su revisión muestra una descripción de los métodos de clasificación desbalanceada desde dos perspectivas: (i) determinar cuáles son los modelos más utilizados y que logran una alta precisión predictiva y (ii) disponer de algoritmos y métricas que permitan una adecuada evaluación del conjuntos de datos.

Los investigadores concluyeron que las técnicas de pre procesamiento de datos[9] (sobremuestreo, submuestreo y muestreo híbrido) son las más utilizadas siendo las de sobremuestreo las de mayor preferencia, específicamente SMOTE (ver figura 2.2a). Señalan, además, que existen una gran variedad de métricas para la evaluación del desempeño de los clasificadores, las más utilizadas fueron el precisión global, precisión, sensibilidad y F1 (ver figura 2.2b), sin embargo destacan que la F1 muestra una adecuada puntuación cuando se evalúa la capacidad predictva de los clasificadores en el contexto de

[9]Estas técnicas se ubican dentro de las estrategias a nivel de datos (Data Level o Data Pre Processing) y son utilizadas con más frecuencia que las estrategias a nivel de algoritmos o las estrategias híbridas, según Bujang et al. (2022)

clases desbalanceadas.



(a) (b)

Figura 2.2: Técnicas de muestreo y métricas de medición aplicados a clases desbalanceadas del área de educación - Tomado de Bujang et al. (2022)

2.7. Métodos parámetricos para enfrentar el problema de datos desbalanceados

En los modelos de regresión binaria se utilizan funciones de enlace como logit o probit. Sin embargo, éstas no resultan ser las más adecuadas cuando los datos se encuentran desbalanceadas pues la capacidad predictiva del modelo se ve afectada tal como sucede en los metodos de aprendizaje automático.

En los trabajos de Bazán et al. (2017), de la Cruz Huayanay (2023) se propone como alternativa el uso de enlaces asimétricos como los modelos de regresión binaria con función de enlace asimétrica con base en distribuciones potencia y reversa potencia los cuales incluye un parámetro de forma el cual permite modular la asimetría de la distribución.

2.8. Inferencia Bayesiana

En Wundervald (2019) se ofrece una presentación sobre el terorema de Bayes y la inferencia bayesiana, ambos conceptos son ampliamente desarrollados por los autores y los exponemos a continuación.

El teorema de Bayes permite calcular la probabilidad condicional de que ocurra un evento basándose en información previa y nueva evidencia (datos).

El teorema de Bayes se define como:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.7}$$

Donde:

- $P(A)$ es la probabilidad de que ocurra el evento A, también conocida como la probabilidad a priori.

- $P(A|B)$ es la probabilidad condicional de que ocurra el evento A, dado que B es verdadero. Esta es la probabilidad a posteriori, ya que depende de la variable B. Esto supone que el evento A no es independiente de B.
- $P(B|A)$ es la probabilidad condicional de que ocurra el evento B, dado que A es verdadero.
- $P(B)$ es la probabilidad de que ocurra el evento B.

La inferencia bayesiana tiene como fundamento el teorema descrito. Para una mejor comprensión consideremos el vector aleatorio Y con una distribución que depende de un parámetro o vector de parámetros $\theta \in \Theta$. La función de verosimilitud se define como:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{i=1}^{n} p\left(y_i \mid \boldsymbol{\theta}\right) \tag{2.8}$$

donde los y_i son los valores observados de Y con $i = 1.....n$. Además toda la información de las observaciones y_i sobre θ está incluida en 2.8

Desde la perspectiva clásica o frecuentista, la dificultad para estimar θ se convierte en un problema de optimización al maximizar la función de verosimilitud (o su logaritmo).

$$\mathcal{L}(\hat{\theta}) = \max_{\theta \in \Theta} ln[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y})] \tag{2.9}$$

En contraste, bajo la metodología bayesiana, la estimación de θ está dada por la distribución posterior conjunta, que se define mediante el teorema de Bayes.

$$p(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\mathcal{L}(\theta \mid \mathbf{y}) p(\boldsymbol{\theta})}{\int_{\Theta} \mathcal{L}(\theta \mid \mathbf{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{2.10}$$

donde Θ representa el espacio paramétrico de $\boldsymbol{\theta}$ y $p(\boldsymbol{\theta})$ corresponde a la distribución a priori. Por lo tanto, la ecuación 2.10 puede escribirse como:

$$p(\boldsymbol{\theta} \mid \mathbf{y}) \propto \mathcal{L}(\theta \mid \mathbf{y}) p(\boldsymbol{\theta}) \tag{2.11}$$

puesto que la expresión $\int_{\Theta} \mathcal{L}(\theta \mid \mathbf{y}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ es la distribución marginal de $\mathbf{y}$ y ésta no depende de θ. Por su parte, la distribución posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ proporciona toda la información que se puede tener sobre θ. Por ejemplo, es posible evaluar $p(\boldsymbol{\theta} \mid \mathbf{y})$, su media, mediana, varianza, así como otras cantidades como los cuantiles, con el objetivo de obtener estimaciones puntuales e intervalares.

Por lo general, la distribución a posteriori no tiene una forma conocida por lo que se hace necesario recurrir a métodos numéricos para su estimación.

Las ventajas de la inferencia bayesiana en comparación con la clásica son:

- Que la inferencia bayesiana considera toda la información disponible y la interpretación probabilística de las estimaciones es directa.
- El conocimiento previo se incorpora a través de la distribución a priori junto con los datos representados por la función de verosimilitud,

- Toda la inferencia se lleva a cabo con base en la distribución a posteriori.
- En el contexto bayesiano, dado un intervalo de estimación, $\boldsymbol{\theta}$ pertenece a éste con una probabilidad de $(1-\beta)\,\%$, donde β es la probabilidad de que $\boldsymbol{\theta}$ este fuera del intervalo de credibilidad.

2.9. Comentarios finales del capítulo

En este punto del estudio surgen algunas interrogantes interesantes como: (i) cuál es la mejor estrategia que nos permita un adecuado balance de clases, (ii) cuál es el mejor método de aprendizaje automático a elegir que nos brinde los mejores resultados en términos de desempeño predictivo y (iii) cuál es o son las métricas de evaluación que nos permitan determinar, entre varios métodos de clasificación, el que presente mejor desempeño predictivo.

Frente a esto, se realizó una amplia revisión de la literatura. A modo de resumen, en el Cuadro 2.3 presentamos los diferentes trabajos relacionados con el desbalance de datos que, junto con los trabajos previamente citados, nos permiten llegar a algunas conclusiones interesantes.

En principio, se evidencia que el problema del desbalance de datos es un tema de gran interés para los investigadores de aprendizaje automático (en inglés, machine learning) y mineria de datos (del inglés data mining) afirmación que podemos constatar por la vasta producción en investigaciones que han abordado el tema en cuestión. Pero también es cierto que el problema al no ser resuelto representa un desafío para las investigaciones recientes las cuales desarrollan innovadoras propuestas como las de Chen et al. (2004) que proponen dos alternativas el Bosque Aleatorio Balanceado (En inglés, Balanced Random Forrest - BRF) y Bosque Aleatorio Ponderado (En inglés, Weighted Randon Forrest WRF).

Dentro de las estrategias para afrontar el problema de desbalance las preferidas son las que abordan el problema desde la perspectiva de los datos, específicamente las técnicas de remuestreo, destacando el sobremuestreo, debido a que en comparación a las de submuestreo ofrece mejores ventajas comparativas. Destacan en particular el Random Oversampling (ROS) y el SMOTE, el uso de esta última se presenta con mayor frecuencia que el ROS.

El problema del desbalance de clases ha sido tratado desde distintos enfoques, en el sentido de que se han utilizado diferentes tipos de clasificadores en busca de seleccionar aquel que logre el mejor desempeño predictivo. Sin embargo, queda claro que aún "no existe un ganador"[10] pero lo que sí encontramos en la literatura reciente es que de la multiplicidad de clasificadores empleados son el Bosque Aleatorio y la Regresión Logística los de uso más frecuente (ver Cuadro 2.3).

Cuando el resultado observado en la aplicación de los diferentes métodos de aprendizaje automático, como el bosque aleatorio, regresión logística, árboles de clasificación y máquinas de soporte vectorial son sometidos a distintas métricas para evaluar su desempeño predictivo lo que se observa en las investigaciones (Bharadwaj (2023),Vitório y Marques (2021), Qaddoura y Biltawi (2022)) es que el bosque aleatorio presenta una mejor capacidad de predicción, incluso en situaciones en las que el nivel de severidad del desbalance ha sido alto (Goyal y Kumar (2020)).

[10] En palabras de Chen et al. (2004)

Sin embargo, como señalamos en el capítulo 1 y reafirma Chen et al. (2004) no se puede concluir que el bosque aleatorio sea el método que resuelve definitivamente el problema del desbalance de datos pero como lo evidencian las investigaciones recientes (ver Tabla 2.3) es el método más eficiente en términos de desempeño predicitivo. Por las razones expuestas pondremos a prueba el desempeño predicitvo de este método no paramétrico frente a una nueva propuesta desarrollada por Bazán et al. (2017); de la Cruz Huayanay et al. (2019) pero utilizando métodos paramétricos como la regresión potencia logistica.

Cuadro 2.3: Aplicación de métodos de aprendizaje automático a conjuntos de datos con clases no balanceadas

Área de Estudio	Modelo ML	Métricas de Evaluación				Severidad	Estrategia de Muestreo	Referencia
		A	P	R	F1			
Finanzas (Fraude TC)	Regresión Logística	0.999	0.783	0.730	0.755	0.172 %	Estándar	Planinić y Popović-Bugarin (2024)
	Bosque Aleatorio	0.999	0.930	0.81	0.866			
	CatBoost	0.999	0.953	0.837	**0.891**			
Finanzas (Fraude)	**Bosque Aleatorio**	0.999	0.900	0.830	**0.860**	0.172 %	ADASYN	Bharadwaj (2023)
	KNN	0.996	0.480	0.890	0.620			
Finanzas(Crédito)	Regresión Logística	0.870	0.810	0.980	0.880	20 %	SMOTE	Karim et al. (2022)
	XG Boost	0.850	0.810	0.910	0.860			
	Árboles de Decisión	0.810	0.820	0.810	0.810			
	Bosque Aleatorio	0.870	0.810	0.97	**0.890**			
Finanzas (Banca)	**Bosque Aleatorio**	0.813	0.813	0.813	**0.813**	11.5 %	RUS	Vitório y Marques (2021)
	Regresión Logística	0.805	0.806	0.805	0.805			
	Neural Network	0.775	0.776	0.775	0.775			
	Naive Bayes	0.754	0.755	0.754	0.754			
	AdaBoost	0.749	0.749	0.749	0.749			
Finanzas (Fraude)	Regresión Logística		0.232	0.840	0.363		SVM SMOTE	Qaddoura y Biltawi (2022)
	Bosque Aleatorio		0.975	0.800	**0.879**			
	KNN		0.119	0.260	0.164			
	Naive Bayes		0.211	0.820	0.336			
	SVM		0.330	0.720	0.453			
	Arboles de Decisión		0.812	0.780	0.796			
Salud	Gradient Boosting	0.770	0.910	0.770	**0.820**	6.36 %	SMOTE	Cecchini et al. (2019)
Salud	**Regresión Logística SMOTE**	0.837	0.597	0.757	0.665	21 %	SMOTE	Zhao et al. (2018)
	Regresión Logística	0.850	0.694	0.521	0.595			
	LSVM	0.859	0.767	0.479	0.590			
	Arboles de Decisión	0.841	0.750	0.375	0.500			
	RSVM	0.850	0.792	0.396	0.528			
Tecnología	KNN	0.822				2.46 %	RUS	Goyal y Kumar (2020)
	Gaussian Naive Bayes	0.719						
	Mult. Naive Bayes	0.684						
	Árboles de Decisión	0.848						
	Bosque Aleatorio	**0.904**						
Educación (Dropout)	**Right GBM**	0.940	0.810	0.780	**0.790**	11 %	SMOTE	Song et al. (2023)
	XG Boost	0.930	0.760	0.790	0.77			
	Regresión Logística	0.800	0.500	0.780	0.610			
	SVM	0.800	0.510	0.780	0.620			
	Bosque Aleatorio	0.930	0.790	0.730	0.750			
	Arboles de Decisión	0.890	0.630	0.700	0.660			
Educación (Dropout)	Árboles de Decisión	0.998	0.997	0.998	0.998		SMOTE	Rosly et al. (2023)
	Naive Bayes	0.999	1.000	0.997	0.999			
	MLP	0.998	1.000	0.996	0.998			
	Bosque Aleatorio	0.998	0.998	0.998	**0.998**			
	Regresión Logística	0.999	1.000	0.997	0.999			

Capítulo 3

Método de Bosque Aleatorio

El Bosque Aleatorio es un algoritmo de aprendizaje automático propuesto por Breiman (2001) compuesto por un conjunto de árboles de decisión los cuales emiten un resultado que luego se combinan para obtener, finalmente, un único resultado. En este capítulo haremos una descripción y revisión de su funcionamiento y sus principales características por lo que para tener una mejor comprensión empezaremos explicando que es un árbol de clasificación.

3.1. Introducción a los árboles de clasificación

Los árboles que forman el Bosque Aleatorio son árboles recursivos de partición binaria, Cutler et al. (2012). Es decir, a partir de un "nodo raiz", que contiene todo el espacio predictor, se realizan divisiones de forma recursiva. Durante este proceso, encontraremos "nodos terminales", es decir, aquellos que no requieren más divisiones y forman parte de la partición final del espacio predictor y, por otro lado, los nodos que aún pueden ser particionados y se denominan "nodos no terminales o intermedios" éstos se dividen según el valor de una de las variables predictoras en dos nodos descendientes, uno a la izquierda y otro a la derecha. Considerando una variable predictora continua la partición viene determinada por un punto de división o corte, si los valores que toma esta variable son menores al punto de división éstos van a la izquierda por el contrario, si los valores de la variable predictora son mayores al punto de división van a la derecha, la figura 3.1, ilustra un árbol de decisión.

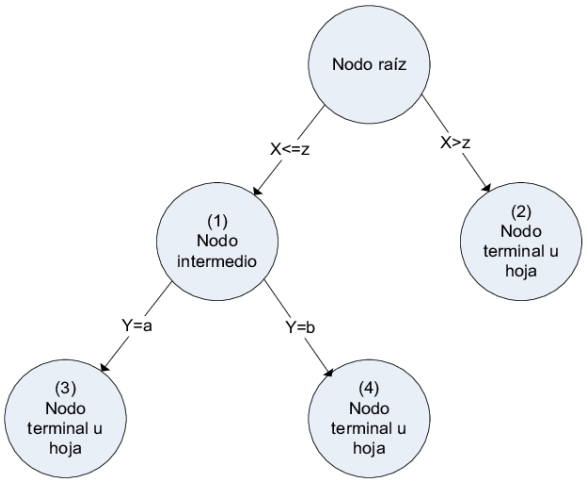


Figura 3.1: Arbol de Decisión. Tomado de Ortiz Lozano et al. (2017)

Para la elección del punto de división se utiliza como criterio el grado de impureza observado en cada nodo no terminal con la idea de que cada nodo descendiente sea lo más puro posible y, de esta forma, detener el proceso de partición. Para medir el grado de impureza de un nodo tenemos el índice de mala clasificación, el índice de Gini y el índice de Entropía.

El árbol de clasificación tiene como una de sus principales ventajas su fácil construcción e interpretación y dentro de las desventajas más relevantes su tendencia a mostrar sobreajuste afectando su capacidad predictiva.

Para mayor profundización acerca de árboles de clasificación se pueden revisar los trabajos de Breiman (2017),Véliz Capuñay (2023), Cutler et al. (2012).

3.2. Bosque Aleatorio

El método Bosque Aleatorio es un algoritmo de aprendizaje supervisado que durante su proceso de aprendizaje combina árboles de decisión aleatorios logrando mejorar su desempeño predictivo. Busca superar las desventajas de los árboles de decisión como su extrema sensibilidad a pequeños cambios en el conjunto de datos de entrenamiento. Sobre el particular, Bharadwaj (2023) afirma que con sólo cambiar la muestra puede dar lugar a un árbol diferente, por lo que señala que una forma de superar esta sensibilidad es agregando un mayor número de árboles. Por su parte, Peargin (2019) sostiene que los árboles de decisón se caracterizan por ser susceptibles a la compensación (tradeoff) entre sesgo y varianza[11].

Con este método, el proceso de agregación (bagging) de cada árbol se realiza a través de un muestreo aleatorio con reemplazo (bootstrap), lo que asegura la independencia entre los árboles pues cada

[11]Es posible que un árbol de tamaño máximo no muestre sesgo pero no será, necesariamente, robusto al ruido debido al sobreajuste. De manera similar, un árbol más pequeño puede generalizar mejor, pero estará sesgado debido a que sus nodos terminales tendrían un nivel alto de impureza

uno de ellos se construye utilizando los conjuntos de datos generados de los datos de entrenamiento. Luego, para cada árbol se eligen de manera aleatoria un determinado número de variables predictoras y dentro de éstas a la mejor para luego seleccionar el punto de corte más conveniente, Véliz Capuñay (2023).

El bosque aleatorio recibe por cada árbol construido un voto en función de alguna de las clases para, posteriormente, realizar la clasificación tomando en consideración el voto mayoritario. Es un modelo versátil y de una gran flexibilidad gracias a su capacidad de adaptarse a diferentes tipos de datos, es un algoritmo rápido y fácil de implementar, produce predicciones altamente precisas y tiene la capacidad de manejar una gran cantidad de datos sin caer en un problema de sobreajuste según refiere O Brien y Ishwaran (2019), además muestra un excelente rendimiento en entornos donde el número de variables es mucho mayor que el número de observaciones Biau y Scornet (2016), lo cual se comprueba cuando enfrenta a conjuntos de datos con presencia de clases desbalanceadas verificándose, efectivamente, su alto desempeño predictivo como se pudo constatar en el capítulo anterior.

El método de bosque aleatorio fue propuesto por Breiman (2001), quien lo define formalmente como:

Definición 2. *El clasificador con el que se construye el bosque aleatorio son árboles de clasificación que se denotan* $h(\mathbf{x}, \Theta)$*, donde* $\mathbf{x}$ *es el vector de variables predictoras y* Θ *es un vector aleatorio. Así, el bosque aleatorio es un clasificador compuesto por otros clasificadores estructurados en* k *árboles aleatorios. De tal forma que el bosque aleatorio se puede expresar como* $\{h(\mathbf{x}, \Theta_k), k = 1, \ldots\}$.

Entonces, para un conjunto de datos de entrenamiento $\mathfrak{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ *y un* θ_j *particular de* Θ_k*, un árbol ajustado se denota* $\hat{h}_j = (x, \theta_j, \mathfrak{D})$.

En la figura 3.2 tenemos la representación gráfica del Bosque Aleatorio. Donde, **X** (Base de Datos) es el conjunto de entrenamiento al que se le aplica un muestreo con reemplazo (bootstrap) con la finalidad de obtener un número determinado de árboles de clasificación $h(\mathbf{x}, \Theta_k)$. En línea con la definición de Breiman (2001) cada árbol $h(\mathbf{x}, \Theta_k)$ genera un vector Θ_k independiente de otros $\Theta_1 ... \Theta_{k-1}$. Además, considera sólo un subconjunto aleatorio de características (features) para posibles divisiones en cada nodo no terminal. Estas características del bosque aleatorio garantiza que los árboles no estén correlacionados.

El objetivo es que el Bosque Aleatorio crezca de tal forma que contenga una gran cantidad de árboles y que cada uno logre alcanzar su tamaño máximo. Al finalizar el proceso de aprendizaje cada árbol tendrá un resultado (emitirá un voto) que corresponde a alguna de las clases. La clase predicha será la que obtenga la mayor votación del total de árboles que conforman el bosque.

3.3. Fundamentos teóricos

A pesar de su popularidad, excelente rendimiento y su relativa simplicidad algorítmica algunos autores sostienen que:

Los mecanismos matemáticos que impulsan al Bosque Aleatorio aún no se comprenden bien. Más específicamente, la cuestión teórica fundamental de la consistencia, es decir, si la convergencia hacia un

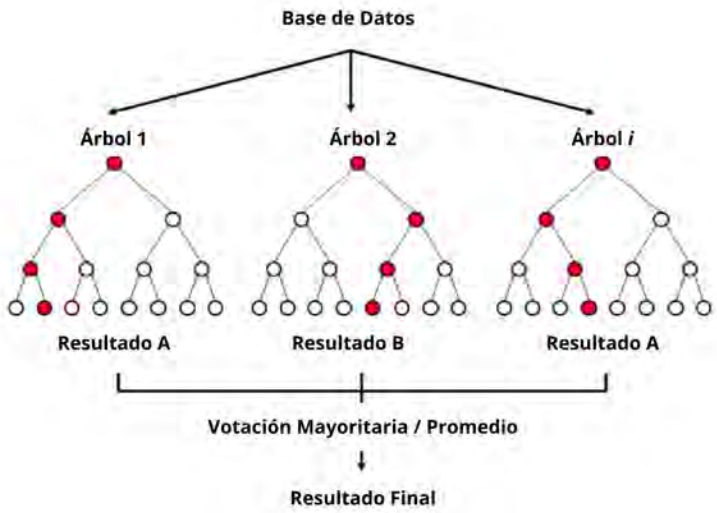


Figura 3.2: Bosque Aleatorio – Tomado de Ghosh (2024)

modelo óptimo está garantizada siempre que se disponga de un conjunto de aprendizaje infinitamente grande, sigue siendo un problema abierto y difícil, señala Louppe (2014).

Por su parte, Biau y Scornet (2016) indica que desde el punto de vista teórico, la historia del Bosque Aleatorio es menos concluyente y, a pesar de su amplio uso, se sabe poco sobre las propiedades matemáticas del método.

En la misma línea, Genuer et al. (2008) sostiene que una de las principales cuestiones abiertas sobre el Bosque Aleatorio es dilucidar desde un punto de vista matemático su comportamiento excepcionalmente atractivo.

De hecho, sus propiedades matemáticas siguen siendo en gran medida desconocidas y, hasta ahora, la mayoría de los estudios teóricos se han concentrado en partes aisladas o versiones estilizadas del algoritmo manifiesta Biau (2012).

Sin embargo, a pesar de estas observaciones el mismo Louppe (2014) hace una amplia presentación de una diversidad de investigaciones teóricas que brindan argumentos que confirman el porqué el Bosque Aleatorio tiene un buen funcionamiento en la práctica algunas de las cuales explicaremos a continuación.

Reproduciendo la propuesta de Bosque Aleatorio desarrollada por Breiman (2001) explicaremos los fundamentos teóricos sobre los que se estructura el algoritmo.

3.3.1. Convergencia

Dado un conjunto de clasificadores $h_1, h_2, ...h_k$, los cuales son entrenados con datos aleatoriamente seleccionados del conjunto de observaciones del vector aleatorio $(Y, \mathbf{X})$; donde Y es la variable respuesta y $\mathbf{X}$ son los predictores. Se puede definir a la función margen, $mg(\mathbf{X}, Y)$, como una medida que permite establecer cuán confiable es la clasificación en $\mathbf{X}$ para la clase correcta Y, frente a las demás clases $j \neq Y$, y se expresa como:

$$mg(\mathbf{X}, Y) = \frac{1}{K}\sum_{k=1}^{K} I\left(h_k(\mathbf{X}) = Y\right) - \max_{j \neq Y} \frac{1}{K}\sum_{k=1}^{K} I\left(h_k(\mathbf{X}) = j\right) \tag{3.1}$$

donde $I(.)$ es la función indicadora que devuelve 1 si la expresión entre paréntesis es verdadera y 0, en caso contrario. De la expresion 3.1 podemos concluir que a mayor margen más confianza en la clasificación.

A partir de la definición del $mg(\mathbf{X}, Y)$, Breiman (2001) propone la siguiente definición del error de generalización:

$$PE^* = P_{\mathbf{X},Y}(mg(\mathbf{X}, Y) < 0) \tag{3.2}$$

La expresión $mg(\mathbf{X}, Y) < 0$ en 3.2 indica que el margen es negativo lo cual sugiere que en promedio las clasificaciones incorrectas superan a las correctas, es un indicador de falta de confianza en la clasificación. Por lo tanto, el error de generalización PE^* mide qué tan probable es que la clasificación sea incorrecta para el conjunto (Y,X)[12].

En un Bosque Aleatorio tenemos que, $h_k(X) = h(X, \theta_k)$. Para un gran número de árboles, se desprende por la Ley Fuerte de los Grandes Números que cuando el número de árboles se incrementa, para casi seguramente todos los Θ_1, Θ_2..., el Error de Generalización, PE^*, converge a[13]:

$$P_{\mathbf{X},Y}\left(P_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(\mathbf{X}, \Theta) = j) < 0\right) \tag{3.3}$$

Para una mejor comprensión de 3.3 haremos una explicación de sus componentes. En primer lugar, la expresión $P_{\Theta}(h(\mathbf{X}, \Theta) = Y)$, corresponde a la probabilidad de que X sea clasificada correctamente en la clase Y, bajo los parámetros θ. Por su parte, la expresión $\max_{j \neq Y} P_{\Theta}(h(\mathbf{X}, \Theta) = j) < 0)$, es la probabilidad máxima de que X sea clasificada en una clase diferente de Y, en ese sentido, j puede ser cualquier otra clase que no sea Y. Entonces, si la diferencia entre ambas expresiones es negativa el clasificador pierde confiabilidad puesto que es menos probable que la clasificación sea la correcta con relación a las otras clases posibles. Si compraramos 3.1 con 3.3 observamos que en la primera se define al margen en función de promedios y en la segunda, en términos de probabilidades.

La expresión completa $P_{\mathbf{X},Y}(.)$ permite evaluar qué tan probable es que el margen sea negativo lo cual es un indicador, según hemos referido, de falta de confiabilidad en el clasificador.

Este resultado explica porqué el algoritmo no genera un sobreajuste a medida que más árboles son agregados, sino que establece un valor límite para el error de generalización. En el sentido de que un margen negativo sugiere que el modelo está aprendiendo demasiado de los datos de entrenamiento. y no clasifica adecuadamente frente a datos nuevos.

3.3.2. Fuerza y Correlación

Breiman (2001) señala que se puede establecer un límite superior para el error de generalización

[12] De acuerdo a Cutler et al. (2012) se asume que la probabilidad conjunta $P_{X,Y}$ es desconocida.
[13] La demostración se encuentra en Breiman (2001)

en función de dos parámetros. Uno permite medir qué tan preciso se muestran los clasificadores individualmente y el segundo mide el grado de dependencia entre ellos, como se comporten estos parámetros permitirá comprender el funcionamiento del Bosque Aleatorio.

Se define a este límite superior (la demostración se encuentra en el apéndice A) como:

$$PE^* \leq \bar{\rho}\left(1 - s^2\right)/s^2 \tag{3.4}$$

El parámetro $\bar{\rho}$ corresponde a la correlación entre miembros diferentes del bosque promediado sobre la distribución de Θ y Θ^* y se define como:

$$\bar{\rho} = E_{\Theta,\Theta'}\left[\rho\left(h(\cdot,\Theta), h\left(\cdot,\Theta'\right)\right)\right] \tag{3.5}$$

Donde:

$$\rho(h(\cdot,\theta), h(\cdot,\theta')) = \frac{\mathrm{Cov}(h(\cdot,\theta), h(\cdot,\theta'))}{\sigma_{h(\cdot,\theta)} \cdot \sigma_{h(\cdot,\theta')}}$$

El siguiente parámetro lo denomina la fuerza del clasificador y resulta ser el valor esperado del margen, esto es:

$$S = E_{\mathbf{X},Y} mg(\mathbf{X}, Y) \tag{3.6}$$

Observamos que la fuerza está en función de la confiabilidad del clasificador pues como indicamos cuando el margen es alto significa que los votos en la clase correcta superan a los votos en las clases incorrectas, es decir, mientras más confiable el clasificador mayor fortaleza adquiere. Lo ideal es tener una S muy grande y una correlación, $\bar{\rho}$, baja. Las estimaciones del error de generalización, fuerza y correlación se obtienen a partir de las estimaciones OOB como explicaremos en la sección 3.4.1. Para mayor abundamiento en Breiman (2001) encontramos el desarrollo de cómo se obtienen estas estimaciones.

3.4. Características y propiedades

3.4.1. Estimaciones Out-Of-Bag (OOB)

Daremos, para un mejor entendimiento, una definición de qué es el método bagging, El bagging es un método agregado (del inglés, ensemble methods) mediante el cual de un conjunto de entrenamiento inicial se forman diferentes sub conjuntos obtenidos luego de aplicar el método bootstrap (elección aleatoria con reemplazo del mismo número de elementos que tiene el conjunto de entrenamiento). Cada uno de estos subconjuntos se entrenan de forma independiente a un clasificador (por ejemplo, árbol de decisión) obteniéndose un resultado por cada sub conjunto formado. Así, obtenidas todas las predicciones, éstas se agregan y se elije dentro de ellas a la más votada, en caso el problema

que enfrentamos sea uno de clasificación. Finalmente, se puede mostrar [14] que en la aplicación del bootstrap un 37 % de las observaciones quedan fuera de los nuevos sub conjuntos formados. Las observaciones que no se encuentran en una muestra bootstrap se denominan observaciones fuera de la bolsa (OOB, por sus siglas en inglés).

En Breiman (1996, 2001), se presenta una amplia explicación acerca de las estimaciones que se hacen a partir de las OOB. El autor señala dos razones por las que se utiliza el bagging. La primera, afirma, es que el uso de bagging parece mejorar la precisión cuando se utilizan características aleatorias y la segunda es que el bagging permite realizar estimaciones continuas del error de generalización del conjunto combinado de árboles, así como estimaciones de la fortaleza y la correlación. Para una mejor comprensión tomamos la explicación de Breiman (2001) y es la que sigue a continuación.

Supongamos un método para construir un clasificador a partir de cualquier conjunto de entrenamiento. Dado un conjunto de entrenamiento específico T, se forman k conjuntos de entrenamiento bootstrap T_k y se construyen clasificadores $h(\mathbf{x}, T_k)$, luego cada uno de estos clasificadores emiten un voto para formar el predictor basado en el método bagging. Para cada par $y, \mathbf{x}$, en el conjunto de entrenamiento, se agregan los votos únicamente de aquellos clasificadores para los cuales $h(\mathbf{x}, T_k)$ no contiene a $y, \mathbf{x}$, a este se le llama el clasificador OOB. Entonces, la estimación OOB del error de generalización es la tasa de error del clasificador OOB en el conjunto de entrenamiento. La fuerza y la correlación, sostiene, también pueden estimarse utilizando métodos OOB, proporcionando estimaciones internas útiles para comprender la precisión de la clasificación y cómo mejorarla, ver Breiman (2001) para mayor detalle.

3.4.2. Uso aleatorio de predictores (features)

Con el objetivo de mejorar la precisión del algoritmo, se deben cumplir dos condiciones (i) minimizar la correlación ρ, (ii) manteniendo la fuerza (S). Cómo lo hace el Bosque Aleatorio, seleccionado de forma aleatoria una o varias combinaciones de predictores (features) al momento de llegar a cada nodo. Como consecuencia de realizar este procedimiento se generan las siguientes consecuencias favorables, en palabras de Breiman (2001):

- Es un modelo que no produce sobreajuste.
- Tiene una muy buena precisión (accuracy).
- Es relativamente robusto a la presencia de datos atípicos (outliers) y al ruido (noisy).
- Ofrece estimaciones internas útiles del error de generalización, de la fuerza y de la correlación.
- Es más rápido que el bagging o el boosting.
- Es simple y fácilmente paralelizable, es decir la independencia de los árboles permite que éstos crezcan de forma simultánea haciéndolo computacionalmente más eficiente.

[14] Si la extracción aleatoria se hace de un conjunto de entrenamiento con N muestras y sólo se extrae una, entonces, la probabilidad de que ésta sea elegida es de $\frac{1}{N}$, en consecuencia, la probabilidad de que no sea la elegida es $(1 - \frac{1}{N})$. Si repetimos el experimento infinidad de veces tendremos que $\lim_{N\to\infty}(1 - \frac{1}{N})^N \approx 0.37$.

3.4.3. Hiperparámetros

Previo al entrenamiento se deben establecer tres hiperparámetros:

1. **El tamaño del nodo**, es decir definir el número de observaciones en el nodo terminal. Si se establece un valor muy bajo, se obtienen árboles profundos, es decir que se realizan más divisiones hasta llegar al nodo final.
2. **Número de árboles (k)**, es recomendable que sea un número muy alto pues asi se garantiza la convergencia del algoritmo. Diferentes estudios muestran que $k = 500$ es un tamaño suficiente.
3. **Número de predictores (mtry)**. Es un hiperparámetro central en el modelo. Corresponde al número de variables predictoras elegidas aleatoriamente y sobre las cuales se elige el punto de división óptimo a partir del cual crecen los árboles. Según Probst et al. (2019) si se realiza el entrenamiento considerando un número bajo de predictores los árboles serán diferentes y menos correlacionados. Cuando nos encontramos frente a un problema de clasificación se sugiere que $mtry = \sqrt{p}$, donde p es el número total de variables predictoras.

Bosque aleatorio y clases desbalanceadas

Existen conjuntos de datos que se caracterizan por tener sus clases desbalanceadas, hemos tratado el tema extensamente en el capítulo 1. Señalamos que los modelos de aprendizaje automático fueron diseñados bajo el supuesto que las clases que componen los conjuntos de datos están perfectamente balanceadas, es decir estos algoritmos no están preparados para enfrentar el problema del desbalance. Ahora bien, su aplicación sin corregir este problema genera que el desempeño predictivo del algoritmo no sea el mejor. Sin embargo, con la finalidad de afrontar el problema revisamos investigaciones en las que se desarrollan diferentes propuestas de solución, en el trancurso de esta investigación revisamos las estrategias a nivel de datos que habilitan a estos algoritmos a lidiar con conjuntos de datos desbalanceados. De la revisión de la literatura pudimos determinar que el Bosque Aleatorio es el modelo de aprendizaje automático que se utiliza con mayor frecuencia pues frente al problema de datos desbalanceados ha sido el que mejor desempeño predictivo ha mostrado frente a otros clasificadores cuando, previamente, al conjunto de datos se le aplicó técnicas de sobremuestreo para resolver el problema.

Sin embargo, en Chen et al. (2004) se proponen, a partir del modelo del Bosque Aleatorio, dos alternativas distintas para afrontar el problema de clases desbalanceadas. La primera, basada en el Aprendizaje Sensible al Costo (técnica que asigna un costo mayor al error de clasificación buscando, de esta forma, minimizar el costo de clasificar una clase minoritaria como mayoritaria), los autores la denominan Bosque Aleatorio Ponderado (En inglés, Weighted Random Forest . WRF) y la segunda, vinculada a las técnicas de muestreo, la cual nombran Bosque Aleatorio Balanceado (En inglés, Balanced Random Forest - BRF).

En este trabajo evaluaremos la aplicación del método de bosque aleatorio en su versión base en el contexto de clases desbalanceadas.

Capítulo 4

Regresión binaria con enlace Potencia Logístico

Otra alternativa para el problema de clasificación binaria con la presencia de datos desbalanceados es el uso de funciones de enlace asimétricas con base en distribuciones potencia y reversa de Potencia Logística, un caso particular de las distribuciones presentados en los trabajos de Bazán et al. (2017) y de la Cruz Huayanay et al. (2019).

En los trabajos mencionados, definen una familia de distribuciones llamadas potencia y reversa de potencia, como sigue

Definición 3. *Una variable aleatoria X sigue una distribución potencia, con parámetros $\mu \in \mathbb{R}$ y $\sigma > 0$ y un parámetro de forma $\lambda > 0$, si su distribución acumulada (fda), denotada por F_P, tiene la forma*

$$F_P(x \mid \mu, \sigma, \lambda) = G\left(\frac{x-\mu}{\sigma}\right)^{\lambda}, \quad x \in \mathbb{R}, \tag{4.1}$$

y distribución reversa de potencia denotada por F_{RP}, si

$$F_{RP}(x \mid \mu, \sigma, \lambda) = 1 - G\left(-\left(\frac{x-\mu}{\sigma}\right)\right)^{\lambda}, \quad x \in \mathbb{R} \tag{4.2}$$

La función $G(\cdot)$, es considerada con la función base que pertenece a la familia de distribuciones simétricas, con soporte en la recta $\mathbb{R}$.

Observe que, para la construcción de una distribución potencia de cualquier función de distribución acumulada continua la cual se eleva a la potencia incorporando un parámetro de forma $\lambda \in \mathbb{R}^+$.

Respecto de las distribuciones línea de base, Bazán et al. (2017) y de la Cruz Huayanay et al. (2019), presentan diferentes distribuciones, por ejemplo: Logística, Normal, Cauchy, Reversa de Gumbel y Gumbel.

A partir de la definición 3, una variable aleatoria Z tiene una distribución potencia $F_P(z)$ y reversa de potencia $F_{RP}(z)$, en su forma estándar cuando $\mu = 0$ y $\sigma = 1$. En este caso, su función de

distribución acumulada (fda) es de la forma:

$$F_P(z) = G(z)^{\lambda} \quad y \quad F_{RP}(z) = 1 - G(-z)^{\lambda}, \quad z\epsilon\mathbb{R} \tag{4.3}$$

Por su parte, tenemos que la función de densidad de probabilidad (fdp) potencia de la variable aleatoria Z puede ser expresada como:

$$f_p(z) = \lambda G(-z)^{\lambda-1} g(z), \quad z\epsilon\mathbb{R} \tag{4.4}$$

y la función de densidad de probabilidad reversa de potencia como:

$$f_{RP} = \lambda G(-z)^{\lambda-1} g(z), \quad z\epsilon\mathbb{R} \tag{4.5}$$

Finalmente, los quantiles de las distribuciones potencia Q_P y reversa de potencia Q_{RP} se muestran para una probabilidad "p" dada, como :

$$Q_P(p) = G^{-1}(-p^{1/\lambda}) \tag{4.6}$$

y

$$Q_{RP}(p) = 1 - G^{-1}(-(1-p)^{1/\lambda}) = -Q_P(1-p) \tag{4.7}$$

Una conclusión importante se da cuando $\lambda = 1$, en este caso particular la distribución de X es simétrica, es decir resulta ser la misma distribución linea de base, en ese sentido, la regresión logística es un caso particular.

En este trabajo, consideramos el estudio de la distribución Potencia Logística, que tiene como distribución de linea de base la fda de una distribución logística. En este caso, si reemplazamos en 4.3, tendremos que su distribución acumulada (fda), su función de densidad de probabilidad (fdp) y su función cuantílica (FQ) son las que se observan en el Cuadro 4.1.

Cuadro 4.1: fda, fdp y FQ para la distribución potencia y reversa Potencia Logístico, con $\eta \in \mathbb{R}$, con parámetro de forma λ

Distribución	**fda**	**fdp**	**FQ**
Potencia Logístico (PL)	$\left(\frac{1}{1+e^{-\eta}}\right)^{\lambda}$	$\lambda\left(\frac{1}{1+e^{-\eta}}\right)^{\lambda-1}\frac{e^{-\eta}}{(1+e^{-\eta})^2}$	$log\left(\frac{p^{\frac{1}{\lambda}}}{1-p^{\frac{1}{\lambda}}}\right)$
Reversa de Potencia Logístico (RPL)	$1-\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{\lambda}$	$\lambda\left(\frac{1}{1+e^{\eta}}\right)^{\lambda-1}\frac{e^{-\eta}}{(1+e^{-\eta})^2}$	$-log\left(\frac{(1-p)^{\frac{1}{\lambda}}}{1-(1-p)^{\frac{1}{\lambda}}}\right)$

Podemos notar que cuando el parámetro de forma toma el valor uno, $\lambda = 1$, las fda, fdp y QF corresponden a la distribución Logística conocida. Además se comprueba que si $\lambda = 1$ la distribución es simétrica y por lo tanto no tiene distribución reversa de potencia como lo indicamos anteriormente.

4.1. Regresión binaria con función enlace Potencia Logístico

En el presente estudio utilizaremos el enfoque bayesiano basados del modelo de regresión binaria con enlace potencia logística, un caso particular de los trabajos de Bazán et al. (2005); Bazan y Millones (2008); Bazán Guzmán et al. (2010); Bazán et al. (2017); de la Cruz Huayanay (2023); de la Cruz Huayanay et al. (2024, 2019), quienes utilizan este enfoque para la estimación, entre otros, del método potencia logistico.

El modelo

Sea $\mathbf{Y} = (Y_1, ..., Y_n)^T$, un vector $n \times 1$ de variables respuesta independientes con $Y_i = \{0, 1\}$, $\mathbf{x_i} = (x_{i1}, ..., x_{ip})^T$, el vector de covariables con $i = 1, ..., n$ y $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, el vector $p \times 1$ de los coeficientes de regresión

Definimos el modelo de regresión binaria bayesiano con función de enlace Potencia Logístico:

$$\begin{aligned} Y_i \mid \boldsymbol{\beta}, \lambda &\sim Bernoulli(\mu_i) \\ \mu_i &= F_L(\mathbf{x_i^T}\boldsymbol{\beta}) \\ (\boldsymbol{\beta}, \lambda)^T &\sim \pi(\boldsymbol{\beta}, \lambda) = \pi(\boldsymbol{\beta})\pi(\lambda) \end{aligned} \tag{4.8}$$

Donde $\eta_i = \mathbf{x_i}^\top \boldsymbol{\beta}$ es el predictor lineal, $F_L(\cdot)$ es la fda de una Potencia Logístico.

Para los parámetros $\boldsymbol{\beta}$ y λ, siguiendo a Bazán et al. (2017) y de la Cruz Huayanay et al. (2019), asumimos que son independientes con la siguiente distribución a priori (i) $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \mathbf{I}\sigma_\beta^2)$ y (ii) $\delta = Log(\lambda) \sim U(-2, 2)$.

La función de verosimilitud asociada al modelo de Potencia Logístico, es dada por la siguiente expresión

$$L(\boldsymbol{\beta}, \lambda \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \left[\left(\frac{1}{1+e^{\eta_i}} \right)^{\lambda} \right]^{y_i} \left[1 - \left(\frac{1}{1+e^{\eta_i}} \right)^{\lambda} \right]^{1-y_i} \tag{4.9}$$

Incorporándola a las distribuciones a priori (por ejemplo; $Log(\lambda) \sim U(-2, 2)$) podemos construir la distribución a posteriori de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)^T$, de la siguiente forma:

$$\pi(\beta, \lambda \mid y, x) \propto \prod_{i=1}^{n} \left[\left(\frac{1}{1+e^{\eta_i}} \right)^{\lambda} \right]^{y_i} \left[1 - \left(\frac{1}{1+e^{\eta_i}} \right)^{\lambda} \right]^{1-y_i} \prod_{j=1}^{p} \exp\left\{ \frac{-\beta_j^2}{2(10)^2} \right\} \frac{1}{4\lambda} \tag{4.10}$$

Por otro lado, si consideramos la función de verosimilitud asociada al modelo de regresión con función de enlace reversa de Potencia Logístico, tiene la siguiente expresión:

$$L(\boldsymbol{\beta}, \lambda \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \left[1 - \left(\frac{e^{-\eta}}{1+e^{-\eta}} \right)^{\lambda} \right]^{y_i} \left[\left(\frac{e^{-\eta}}{1+e^{-\eta}} \right)^{\lambda} \right]^{1-y_i} \tag{4.11}$$

Considerando la función de verosimilitud expresada en 4.11 la distribución a posteriori de $\boldsymbol{\theta} =$

$(\boldsymbol{\beta}, \lambda)^\top$ queda definida como:

$$\pi(\beta, \lambda \mid y, x) \propto \prod_{i=1}^{n} \left[1 - \left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{\lambda}\right]^{y_i} \left[\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right)^{\lambda}\right]^{1-y_i} \prod_{j=1}^{p} \exp\left\{\frac{-\beta_j^2}{2(10)^2}\right\} \frac{1}{4\lambda} \tag{4.12}$$

Como se puede apreciar ambas distribuciones a posteriori no pertenecen a una familia de distribución conocida por lo que su estimación analítica no es posible, en consecuencia, debe realizarse por métodos numéricos. En esa línea, utilizaremos el método de Monte Carlo mediante Cadenas de Markov (MCMC, por sus siglas en inglés). En ese sentido, en esta investigación, como en De La Cruz Huayanay (2019), consideramos el algoritmo No-U-Turn Sampler una extensión del algoritmo MCMC denominado Monte Carlo Hamiltoniano (HMC).

El parámetro λ

En de la Cruz Huayanay et al. (2019) al desarrollar su propuesta para enfrentar el problema de clases desbalanceadas proponen el uso de funciones de enlace potencia incluyendo un parámetro de forma, en nuestro caso lo identificamos como el parámetro λ. Este parámetro permite controlar la asimetría de la distribución. Así tenemos que, si $\lambda > 1$ entonces la distribución potencia está inclinada hacia la derecha (asimetría positiva). Pero si $0 < \lambda < 1$, la distribución potencia está inclinada hacia la izquierda (asimetría negativa).

Por otro lado, Bazán et al. (2017), señala que para un intervalo de valores de η, si $\lambda < 1$ (ó $\lambda > 1$) la curva de la distribución potencia esta arriba (abajo) de la curva línea de base. Por su lado, si $\lambda < 1$ (ó $\lambda > 1$) la curva de la distribución reversa de potencia se encuentra abajo (arriba) de la curva línea de base. Esto se presenta así porque la curva de la distribución reversa de potencia es un

reflejo de la curva de la distribución potencia como puede apreciarse en la figura 4.1.

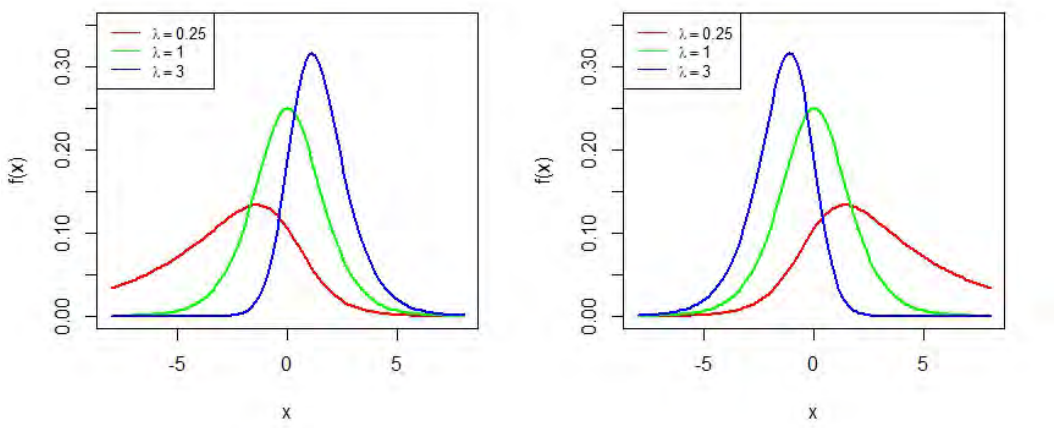


Figura 4.1: Funciones de densidad de probabilidad: Potencia Logístico para $\lambda = \{0.25, 1, 3\}$ y Reversa Potencia Logístico para $\lambda = \{3, 1, 0.25\}$

En el gráfico 4.2 se muestra el caso de una distribución línea de base Logística evaluada para $\lambda = 0.25$, y $\lambda = 3$. Se puede apreciar cuando $\lambda < 1$ la probabilidad de éxito de la distribución potencia es mayor que la probabilidad de la distribución linea de base logística. En este caso, siguiendo a (Bazán et al. (2017)), λ es un parámetro de bonificación. Por su parte, si $\lambda > 1$ la probabilidad de éxito de la distribución potencia es menor que la probabilidad de la distribución línea de base logística, por tanto λ es un parámetro que penaliza. Si el análisis lo hacemos para la reversa Potencia Logístico la

interpretación es a la inversa, es decir $\lambda < 1$ el parámetro penaliza y si $\lambda > 1$, el parámetro bonifica.

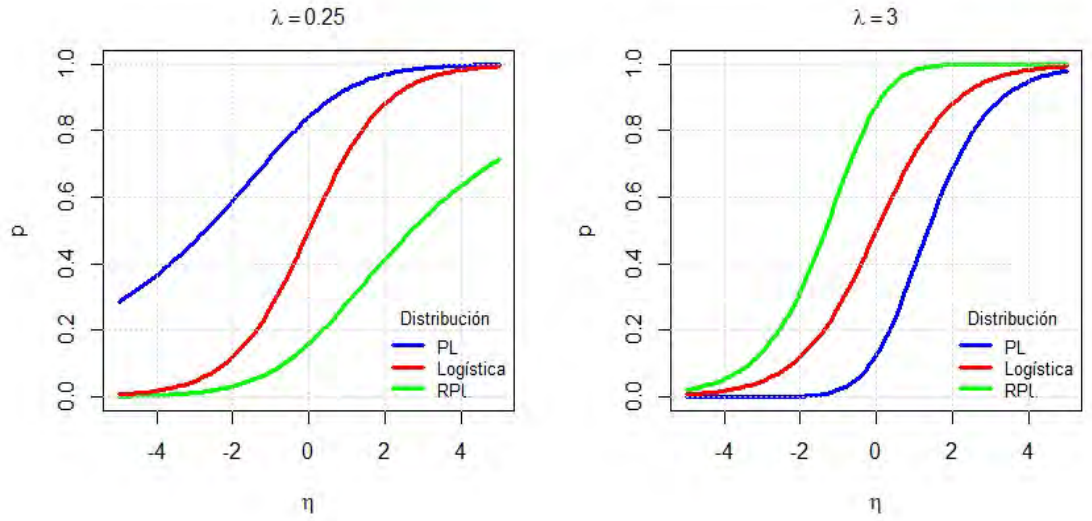


Figura 4.2: Curva respuesta de éxito para la Distribución Logística, Potencia Logístico y Reversa Potencia Logístico para $\lambda = 0.25$ y $\lambda = 3$

La función del parámetro λ en regresión binaria, demuestra que la función de enlace Potencia Logístico puede ser una buena opción para problemas de clasificación binaria desbalanceada, y puede ofrecer buenos resultados, como los mostrados en de la Cruz Huayanay et al. (2019), en los siguientes capítulos se presentarán las aplicaciones de estos modelos.

Capítulo 5

Estudio de Simulación

En este capítulo vamos a desarrollar el estudio de simulación que consistirá de dos etapas. En la primera realizaremos un análisis de sensibilidad comparando el comportamiento del parámetro de forma λ del modelo de regresión Potencia Logístico, considerando para este parámetro tres posibles distribuciones a priori; Uniforme (Bazán et al. (2017) y de la Cruz Huayanay et al. (2019)), Gamma Inversa (Lunn et al. (2000)) y Log Normal (Gelman et al. (1995)). Para la evaluación y elección de la a priori que muestre el mejor ajuste utilizaremos dos criterios, el sesgo y la raíz del error cuadrático medio (RECM).

La segunda parte estará dedicada al estudio de simulación, propiamente dicho, donde compararemos al modelo de regresión Potencia Logístico, tomando en cuenta la a priori seleccionada en la etapa previa, con el modelo de regresión logística, modelo probit y con el modelo no paramétrico bosque aleatorio para lo cual consideraremos distintos niveles de desbalance asumiendo dos posibles valores para λ para diferentes tamaño de muestra (1000, 2500 y 5000). Por último, analizaremos y concluiremos cuál de ellos tiene mejor desempeño predictivo sometiéndolos a la evaluación de doce métricas derivadas de la matriz de confusión estudiadas en el capítulo 2, sección 2.4.2.

5.1. Generación de datos desbalanceados

Para la generación de los datos desbalanceados utilizamos una distribución Potencia Logístico con un coeficiente de regresión $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ y una covariable X, además del parámetro de forma λ. De esta forma tenemos:

$$
\begin{aligned}
Y_i \mid \boldsymbol{\beta}, \lambda &\sim Bernoulli(\mu_i) \\
\mu_i &= F_L(\mathbf{X_i^T}\boldsymbol{\beta}) \\
&= \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{\lambda} \\
\boldsymbol{\beta_2} &\sim N_p(\mathbf{0}, \mathbf{I}\sigma_\beta^2)
\end{aligned}
\tag{5.1}
$$

En particular, para la simulación de los datos se asumieron las siguientes especificaciones: $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top = (-0.5, 1.5)^\top$, $\mathbf{X}_i^\top = (x_{i1}, x_{i2})^\top$ con $x_{i1} = 1$ y $x_{i2} \sim N(0, 1)$. Para obtener un desbalance del 16 % el parámetro de forma tuvo un valor de $\lambda = 3$. Como indicamos en la Sección 2.2 por convención la clase de interés (éxito) es la clase positiva 1, en consecuencia la clase que se presenta con menor frecuencia. En ese contexto, un desbalance del 16 % significa que la clase de interés es la que se presenta con menor frecuencia en la base de datos.

5.2. Análisis de sensibilidad para el parámetro de asimetría λ

Con las consideraciones descritas en la sección anterior se generó una muestra 5000 datos y se procedió, en el contexto del modelo con función de enlace Potencia Logístico, con el estudio de sensibilidad para λ, asumiendo para este parámetro tres distribuciones a priori las que detallamos a continuación:

1. A priori 1a: $\text{Log}(\lambda) \sim U(-2, 2)$. Utilizada en las investigaciones de Bazán et al. (2017), de la Cruz Huayanay et al. (2019).
2. A priori 1b: $\lambda \sim IG(2, 3)$. Lunn et al. (2000)
3. A priori 1c: $\lambda \sim LN(0, 0.5)$. Gelman et al. (1995)

La estimación de los parámetros se realiza bajo el enfoque bayesiano. Para obtener la distribución a posteriori de los parámetros estimados utilizamos el lenguaje Stan en el entorno R, empleando la libreria rstan. Se ejecutaron 100 réplicas, configurando para cada una de ellas 2 cadenas, 3000 iteraciones, descartando las primeras 1000 iteraciones y definiendo un salto igual a 2 con el objeto de reducir la autocorrelación dentro de la muestra.

El cuadro 5.1 muestra para cada a priori, la media a posteriori de las estimaciones de β_0, β_1 y λ, junto con el cálculo de sus respectivos sesgos y raíz del error cuadrático medio (RECM), criterios que nos permitirán determinar la precisión en la recuperación de los parámetros, los cuales se encuentran definidos por las siguientes expresiones:

$$\text{Sesgo}\left(\hat{\theta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_j^{(r)} - \theta_j\right), \quad \text{RECM}\left(\hat{\theta}_j\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_j^{(r)} - \theta_j\right)^2}, \quad j = 1, 2, 3 \tag{5.2}$$

donde R es el número de réplicas efectuadas en la simulación y el $\hat{\boldsymbol{\theta}}_{\mathbf{j}}^{\mathbf{r}} = \left(\hat{\beta}_j^r, \hat{\lambda}_j^r\right)$ es la media a posteriori que se obtiene en la réplica r.

Los resultados del análisis de sensibilidad se muestran en el cuadro 5.1 para las tres prioris con las que fue evaluado el parámetro de asimetría (λ).

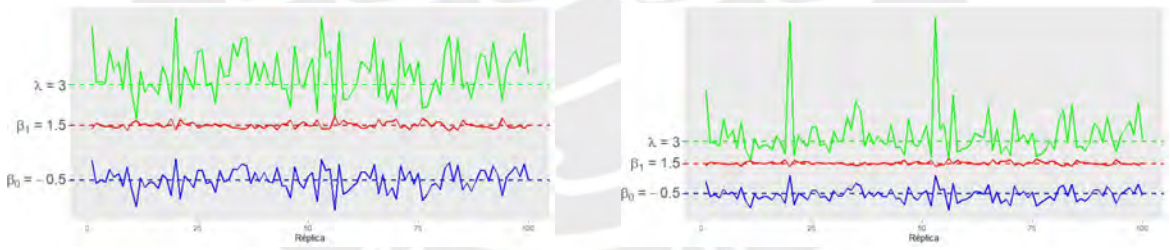
En principio, la a priori 1c es descartada por presentar un mayor valor del sesgo y RECM. Por su parte, se aprecia que es a través de las distribuciones a priori 1a y 1b donde se recuperan mejor los

Cuadro 5.1: Estimación de la media a posteriori de parámetros con diferentes a prioris para el parámetro de asimetría (λ) con un desbalance de clases del 16 % y tamaño de muestra de 5000 datos

A priori	Valor Verdadero	Estimado	Sesgo	RECM
1.a Uniforme	$\beta_0 =$ -0,5	-0,4784	0,0216	0,4020
	$\beta_1 =$ 1,5	1,5131	0,0131	0,1155
	$\lambda = 3$	3,4294	0,4294	0,9399
1.b Gamma Inversa	$\beta_0 =$ -0,5	-0,5712	-0,0712	0,4492
	$\beta_1 =$ 1,5	1,5384	0,0384	0,1244
	$\lambda = 3$	3,4408	0,4408	1,5500
1.c Log Normal	$\beta_0 =$ -0,5	-1,0042	-0,5042	0,5848
	$\beta_1 =$ 1,5	1,6396	0,1396	0,1744
	$\lambda = 3$	2,3711	-0,6289	0,7478

valores verdaderos de β_0, β_1 y λ. Se comprueba, además, que la diferencia de sus respectivos valores estimados $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\lambda}$ son mínimas. Sin embargo, la comparación de los criterios, Sesgo y RECM, nos permite concluir que el modelo Potencia Logístico es el que mejor recupera el valor verdadero cuando el parámetro de asimetría (λ) asume la a priori 1.a. En consecuencia, será esta a priori la que consideraremos en esta investigación.

Para complementar el análisis, en la figura 5.1 se observa que para el parámetro β_1, tanto para 1a como 1b, la convergencia sobre el valor verdadero se mantiene estable en todas las réplicas.



(a) A priori Uniforme (b) A priori Gamma Inversa

Figura 5.1: Réplicas para Uniforme y Gamma Inversa

Respecto de β_0 se aprecia que, en ambos casos, la variabilidad en torno a su valor verdadero es constante en el trayecto de las 100 réplicas. De la misma forma se observa que en cada réplica el valor del parámetro λ sigue el mismo comportamiento que el β_0 pero con mayor variabilidad, sin embargo los valores obtenidos se encuentran, también, alrededor de su valor verdadero.

5.3. Análisis de estudio de simulación

En esta sección evaluaremos el desempeño predictivo que tienen los métodos Logístico, Probit, Potencia Logístico, este último con $\text{Log}(\lambda) \sim U(-2, 2)$ y Bosque Aleatorio cuando se enfrentan a un conjunto de datos con clases desbalanceadas.

Las especificaciones para la simulación consideran que el parámetro λ tome dos valores posibles, 3 y 0.25, de esta manera aseguramos que el conjunto de datos generados para el estudio (proporción de 1´s) muestren un desbalance de 16 % y 76 %, respectivamente.

En el caso de los métodos parámetricos (Logístico, Probit y Potencia Logístico) se estimaron realizando 100 réplicas por cada uno de ellos considerando distintos tamaños de muestra (1000, 2500, 5000) y tomando en cuenta los valores de λ indicados.

Las estimaciones bayesianas se ejecutaron utilizando el lenguaje Stan, en el entorno R, configurando estos métodos con 2 cadenas, 3000 iteraciones, descartando las 1000 primeras y, considerando un salto igual a 2.

Para el caso particular del método Bosque Aleatorio, utilizamos la librería caret disponible en R, realizándose de igual forma 100 réplicas para los distintas niveles de desbalance previamente indicados. Para todas las estimaciones se utilizó los mismos datos generados (1000, 2500, 5000).

Para la evaluación del desempeño predictivo de los métodos indicados utilizaremos doce métricas las cuales se derivan de la matriz de confusión, estas métricas fueron estudiadas en el capítulo 2. El valor de cada una de ellas se obtuvieron al promediar el valor mostrado en cada una de las 100 réplicas luego de aplicar en cada método los tamaños de muestra previstos para cada nivel de desbalance.

5.3.1. Desempeño de métodos predictivos para clases desbalanceadas

A fin de determinar el mejor método predictivo en el marco de clases desbalanceadas hay que definir previamente cuáles son la métricas de evaluación apropiadas que permitan realizar esta selección. Esta definición debe tener en cuenta los posibles escenarios que se pueden presentar cuando el conjunto de datos muestra desbalance entre sus clases.

Iniciaremos el análisis evaluando los métodos Logístico, Probit, Potencia Logístico y Bosque Aleatorio, asumiendo un escenario con un desbalance de clases del 16 % ($\lambda = 3$), es decir, cuando la proporción de 1s es menor a 50 %.

En base a la investigación de de la Cruz Huayanay et al. (2024) donde concluyen que las métricas que presentan un mejor resultado son las que muestran el mayor valor en el rango [0,1], lo que asegura que la predicción del modelo elegido es similar al valor observado, en el contexto del desbalance indicado (16 %) son TPR, CSI, SSI, FAITH, MCC, GM, F1 y KAPPA podemos afirmar de la inspección del cuadro 5.2 que, para los tres tamaños de muestra considerados, es el método Potencia Logístico el que mejor desempeño predictivo presenta cuando lo comparamos con los otros métodos.

En contraste, cuando la proporción de 1´s es mayor al 50 %, situación que ocurre cuando el valor

Cuadro 5.2: Métricas para $\lambda = 3$ con diferentes tamaño de muestra

Datos	Modelo	Métricas											
		ACC	TPR	TNR	CSI	SSI	FAITH	GS	MCC	GM	F1	KAPPA	PDIF
N = 1000	Logístico	0.806	0.832	0.801	0.414	0.262	0.472	0.310	0.511	**0.816**	0.585	0.473	0.018
	Power Logístico	0.830	0.776	0.841	**0.430**	**0.275**	**0.479**	0.335	**0.522**	0.807	**0.601**	**0.501**	0.019
	Probit	0.796	**0.846**	0.787	0.407	0.256	0.468	0.300	0.504	0.816	0.578	0.461	0.018
	Random Forest	0.825	0.467	0.894	0.305	0.181	0.451	0.222	0.362	0.644	0.465	0.360	0.031
N =2500	Logístico	0.819	0.819	0.812	0.406	0.271	0.466	0.321	0.489	0.818	0.512	0.398	0.005
	Power Logístico	0.848	0.786	0.859	**0.440**	**0.288**	**0.490**	0.348	**0.538**	**0.821**	**0.619**	**0.511**	0.023
	Probit	0.809	**0.839**	0.802	0.406	0.271	0.466	0.311	0.499	0.788	0.592	0.458	0.005
	Random Forest	0.824	0.464	0.894	0.300	0.177	0.450	0.217	0.356	0.643	0.460	0.355	0.031
N =5000	Logístico	0.842	0.831	0.808	0.416	0.255	0.479	0.303	0.484	0.820	0.563	0.415	0.037
	Power Logístico	0.853	0.796	0.860	**0.449**	**0.283**	**0.501**	0.342	**0.538**	**0.825**	**0.623**	**0.512**	0.046
	Probit	0.851	**0.851**	0.764	0.418	0.283	0.478	0.323	0.421	0.790	0.614	0.420	0.017
	Random Forest	0.824	0.458	0.895	0.298	0.175	0.449	0.215	0.354	0.640	0.458	0.353	0.031

de λ es 0.25 y, en consecuencia, el desbalance de clases es del 76 % , las métricas que permiten una mejor discriminación son TNR, GS, MCC, GM y KAPPA (de la Cruz Huayanay et al. (2024). El cuadro 5.3 muestra los resultados obtenidos para estas métricas donde podemos comprobar que es el método Potencia Logístico el de mejor desempeño predicitvo.

Cuadro 5.3: Métricas para $\lambda = 0.25$ con diferentes tamaño de muestra

Datos	Modelo	Métricas											
		ACC	TPR	TNR	CSI	SSI	FAITH	GS	MCC	GM	F1	KAPPA	PDIF
N = 1000	Logístico	0.830	0.789	0.816	0.394	0.273	0.507	0.321	0.482	**0.818**	0.561	0.413	0.055
	Power Logístico	0.831	0.774	**0.838**	0.427	0.260	0.479	0.319	**0.516**	0.803	0.601	**0.490**	0.024
	Probit	0.838	0.848	0.761	0.416	0.280	0.475	**0.320**	0.419	0.787	0.611	0.417	0.014
	Random Forest	0.683	0.792	0.339	0.655	0.488	0.642	0.071	0.131	0.516	0.791	0.131	0.100
N =2500	Logístico	0.675	0.673	0.682	0.612	0.441	0.594	0.166	0.308	0.678	0.759	0.285	0.076
	Power Logístico	0.683	0.686	**0.720**	0.629	0.453	0.611	**0.182**	**0.348**	**0.715**	0.823	**0.322**	0.106
	Probit	0.669	0.661	0.695	0.603	0.432	0.586	0.164	0.307	0.678	0.753	0.282	0.075
	Random Forest	0.688	0.795	0.348	0.660	0.493	0.647	0.077	0.143	0.525	0.795	0.143	0.097
N =5000	Logístico	0.829	0.828	0.834	0.434	0.287	0.536	0.334	0.499	0.841	0.561	0.452	0.059
	Power Logístico	0.841	0.822	**0.874**	0.476	0.293	0.521	**0.352**	**0.551**	**0.842**	0.626	**0.539**	0.051
	Probit	0.830	0.879	0.755	0.445	0.258	0.497	0.295	0.410	0.792	0.592	0.433	0.030
	Random Forest	0.684	0.792	0.342	0.656	0.488	0.643	0.072	0.134	0.520	0.792	0.134	0.100

Por lo tanto, podemos concluir que bajo los escenarios analizados, esto es, cuando la proporción 1s es del 16 % ($\lambda = 3$) o del 76 % ($\lambda = 0.25$), el método Potencia Logístico es el que mejor desempeño predictivo muestra en la clasificación de clases desbalanceada cuando éste es comparado con el desempeño obtenido por los otros métodos evaluados (Logístico, Probit y Bosque Aleatorio).

5.3.2. Recuperación de parámetros

Para evaluar la precisión en la recuperación del valor verdadero de los parámetros para los métodos logístico, probit y Potencia Logístico, utilizaremos el criterio de la raíz del error cuadrático medio (RECM). Así, podemos observar en el gráfico 5.2 que es el método Potencia Logístico el que mejor ajuste obtiene para $\lambda = 3$ o $\lambda = 0.25$, y sus combinación con diferentes tamaños de muestra (1000, 2500, 5000), en comparación con los otros métodos mencionados.

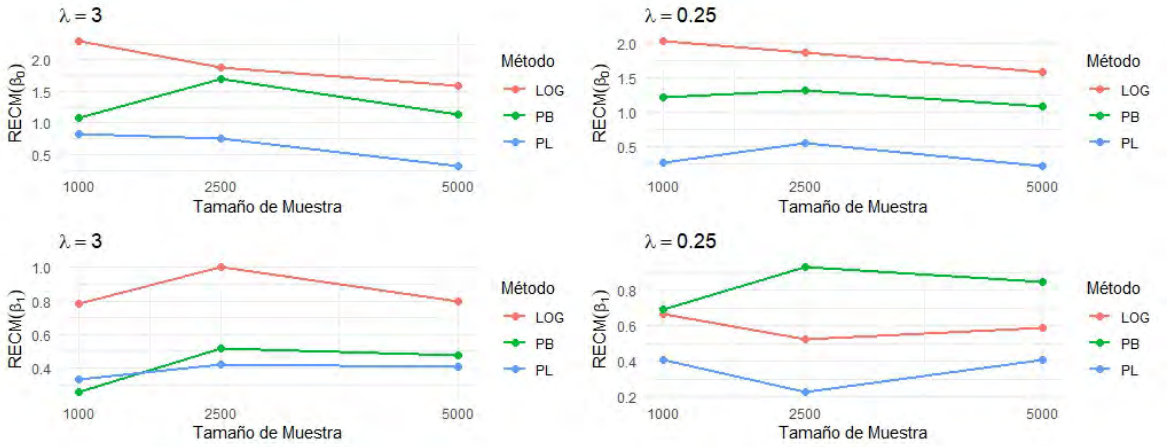


Figura 5.2: Raíz del Error Cuadrático Medio (RECM) para los parámetros β_0 y β_1 en función del tamaño de la muestra y el valor del parámetro de asimetría λ

5.3.3. Estimación de parámetros con el método Potencia Logístico

En el cuadro 5.4 se muestra el resultado de las estimaciones obtenidas para los parámetros del método Potencia Logístico. Podemos notar que para los 6 escenarios evaluados (2 valores de λ para 3 tamaños de muestra) el método recupera adecuadamente el valor verdadero de los parámetros β_0 y β_1 lo cual se corrobora al evaluar la precisión de la estimación observando los criterios de sesgo y RECM.

Cuadro 5.4: Estimación de parámetros β_0 y β_1 con $\lambda = 3$ y $\lambda = 0.25$ con diferentes tamaños de muestra (1000, 2500, 5000)

		$\lambda = 3$			$\lambda = 0.25$		
Datos	**Valor Verdadero**	**Valor Estimado**	**Sesgo**	**RECM**	**Valor Estimado**	**Sesgo**	**RECM**
N = 1000	β_0 = -0.5	-0,588	-0,088	0,819	-0,648	-0,148	0,265
	β_1 = 1.5	1,648	0,148	0,331	1,557	0,057	0,409
N = 2500	β_0 = -0.5	-0,493	0,007	0,750	-0,428	0,072	0,554
	β_1 = 1.5	1,556	0,056	0,424	1,538	0,038	0,225
N = 5000	β_0 = -0.5	-0,507	-0,007	0,317	-0,508	-0,008	0,220
	β_1 = 1.5	1,507	0,007	0,405	1,477	-0,023	0,406

5.3.4. Conclusiones del estudio de simulación

El estudio de simulación realizado nos permite arribar a las siguientes conclusiones:

- El método Potencia Logístico presenta una recuperación satisfactoria de sus parámetros cuando $Log(\lambda) \sim U(-2, 2)$.
- En comparación con los métodos Logístico, Probit y Bosque Aleatorio; el método Potencia Logístico es el que mejor desempeño predictivo presenta.

- En comparación con el método Logístico y Probit, el método Potencia Logístico es el que mejor recuperación obtiene de sus parámetros.
- El método Potencia Logístico presenta una recuperación satisfactoria de sus parámetros, es decir presenta valores razonables de sesgo y RECM.

Capítulo 6

Aplicación

Corresponde en esta etapa de la investigación realizar la aplicación de los métodos estudiados en el capítulo 5 para lo cual utilizaremos datos reales sobre deserción universitaria. Utilizaremos el enfoque bayesiano para la estimación de los modelos paramétricos y para el método no paramétrico lo haremos aplicando el método de bosque aleatorio.

Posteriormente, utilizaremos las métricas que fueron presentadas en la sección 5.3.1 para la comparación del desempeño predicitivo de los métodos estudiados y elegir al método que muestre la mejor performance.

6.1. Descripción de los datos

La base de datos contiene información de 23,108 alumnos que ingresaron a una institución de educación superior entre los años 2012 y 2017, los cuales sólo pueden presentar una de las siguientes situaciones posibles: Abandonó o No Abandonó sus estudios. Los programas tienen una duración de 10 semestres académicos lo que equivale, en la práctica, a 5 años de estudios. En consecuencia, consideraremos como alumnos que no culminaron sus estudios aquellos que los abandonaron antes de los 5 años esperados. El resto corresponde a los alumnos graduados (aquellos que culminaron satisfactoriamente sus estudios en 5 o más años registrando una fecha de egreso cierta y a aquellos que no habiéndose graduado continúan estudiando) para un mejor entendimiento la tabla 6.1 resume la composición del total de alumnos en función del año en que fueron admitidos.

Esta particularidad del conjunto de datos, que se muestra gráficamente en la figura 6.1, nos hace caer en cuenta que estamos frente a un contexto de clases desbalanceadas donde nuestra variable de interés (alumnos que abandonaron sus estudios) es la que se presenta con menor frecuencia (30 %). Estamos, entonces, en el escenario donde la proporción de unos (1´s) de la variable respuesta es menor al 50 %, lo cual se valida cuando $\hat{\kappa} := |2\hat{\mu} - 1| \geq 0.2$, para $\hat{\mu} = 0.3$ (ver sección 2.3). De esta forma, según estudiamos en la sección 5.3.1 las métricas que nos permitirán determinar el método con mejor desempeño predictivo son TPR, CSI, SSI, FAITH, MCC, GM, F1 y KAPPA.

Cuadro 6.1: Distribución de alumnos según su situación académica

Año de ingreso	Numero de alumnos			Tasa de abandono
	No Graduados (Abandonó)	Graduados (No Abandonó)	Total	
2012	1,102	2,409	3,511	31 %
2013	1,060	2,387	3,447	31 %
2014	1,102	2,553	3,655	30 %
2015	1,104	2,780	3,884	28 %
2016	1,229	3,040	4,269	29 %
2017	1,311	3,031	4,342	30 %
Total	**6,908**	**16,200**	**23,108**	**30 %**

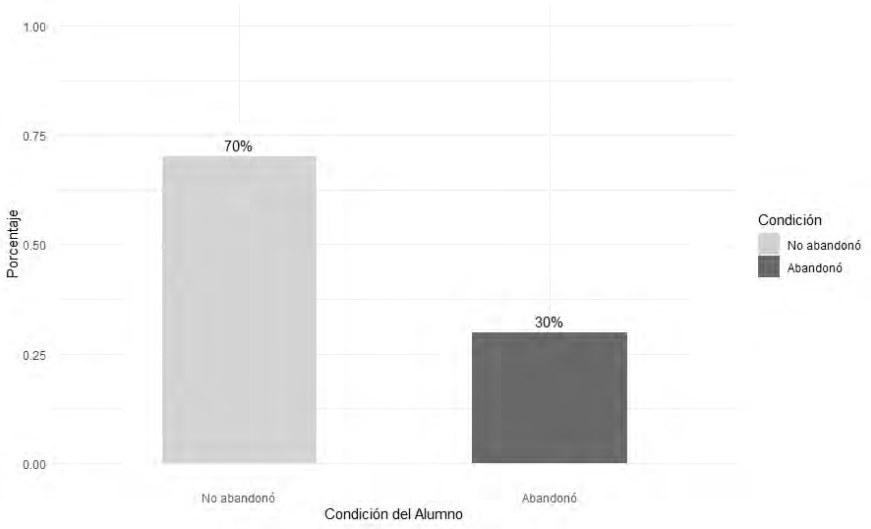


Figura 6.1: Proporción de estudiantes según su condición académica

Por otro lado, sabemos del capítulo 2 cuando hicimos referencia a Véliz Capuñay (2023) que un elemento determinado por las covariables $X^t = (X_1, \ldots, X_q)$ pertenece a la clase $Y = 1$, si $P(Y = 1 \mid \mathbf{X}) < t_0$ para el umbral $t_0 > 0$, y que pertenece a la clase $Y = 0$, si $P(Y = 1 \mid \mathbf{X}) \geq t_0$. Donde, X^t es el vector de covariables y Y es la variable respuesta la cual puede tomar los valores 0 ó 1.

En nuestro caso particular, la variable respuesta Y toma el valor 0 para los graduados o alumnos que no abandonaron sus estudios, y 1 para aquellos estudiantes que abandonaron sus estudios y, en consecuencia, no se graduaron. Por su parte, las covariables $X^t = (X_1, \ldots, X_q)$ son 25 y corresponden a los covariables de cada uno de los alumnos que explican su decisión de abandonar o no los estudios, más adelante haremos una descripción de estas covariables.

La base de datos esta compuesta, además de la variable respuesta Y (Condición), de 25 covariables ($\mathbf{X}$). Como se puede advertir del cuadro 6.2 las covariables son del tipo cuantitativo (12) y categóricas (4 ordinales y 9 nominales).

Descripción de las covariables

Las 25 covariables han sido agrupadas en las siguientes 6 categorías: situación académica, nivel de avance de estudios, economía, becas y subvenciones, antecedentes personales y, finalmente, antecedentes socio demográficos como puede apreciar en el cuadro 6.2.

Cuadro 6.2: Descripción por categoría de las variables que componen la base de datos.

Categoria	Variables	Tipo de variable	Detalle
Situación Académica	**CRAEST**	Cuantitativa	Promedio ponderado de notas acumulado
	Cred.Fal.Egreso	Cuantitativa	Número de créditos que falta al alumno para graduarse.
	N.Cred.Mat.UC	Cuantitativa	Número de créditos matriculados en último ciclo
	N.Cred.Aprob.2	Cuantitativa	Número de créditos aprobados en segunda oportunidad.
	Cred.Aprob.3	Cuantitativa	Número de créditos aprobados en tercera oportunidad.
	Cred.Aprob.4	Cuantitativa	Número de créditos aprobados en cuarta oportunidad.
	Nro.de.cred.matriculados.acumulado	Cuantitativa	Número de créditos matriculados acumulados.
	Prom.Pond.UC	Cuantitativa	Promedio ponderado obtenido en último ciclo
Nivel de avance de estudios	**Etapa**	Categórica / Ordinal	Estudios Generales o Estudios de Facultad.
	Nivel	Categórica / Ordinal	Nivel de estudios alcanzado por el alumno
	FACULTAD	Categórica / Nominal	Facultad de estudios
Economia	**Deuda**	Categórica / Nominal	Si el alumno presenta deuda o no
	Importe.total.de.la.deuda	Cuantitativa	Deuda vigente del alumno
Becas, subvenciones y otros beneficios	**Beneficio**	Categórica / Nominal	Si el alumno tiene beca, exoneracion de pago, descuentos)
	Escala.O	Categórica / Ordinal	Escala de pagos asignada inicial
	Escala.U	Categórica / Ordinal	Escala de pagos asignada final.
	Cambio.E	Categórica / Nominal	Si alumno obtuvo cambio de escala de pago.
Antecedentes personales	**Nro.postula**	Cuantitativa	Numero de postulaciones antes de ingreso a universidad
	Proceso	Categórica / Nominal	Proceso en el que ingreso el alumno
	Tipo.de.colegio	Categórica / Nominal	Si colegio de precedencia es particular o nacional.
	Genero	Categórica / Nominal	Hombre o mujer
	Edad.Egreso	Cuantitativa	Edad de egreso
	datos1_Años	Cuantitativa	Años de estudio
Antecedentes sociodemográficos	**LIMAS**	Categórica / Nominal	Distrito de procedencia
	NSE	Categórica / Nominal	Nivel socio económico del alumno según lugar de residencia.
Situación del Alumno	**Condicion**	Categórica / Nominal	Situación del alumno. Sí (Abandonó) o No (No Abandonó)

6.2. Análisis preliminar

Para determinar cuáles serán las covariables que se utilizarán en la aplicación de los métodos estudiados se requiere evaluar, inicialmente, el nivel de relación de cada una de éstas (cuantitativas y categóricas) con la variable respuesta Y y evaluar la presencia de colinealidad entre las covariables para este fin utilizaremos la prueba Chi cuadrado, la prueba V de Cramer y la prueba de Kruskal Wallis.

6.2.1. Análisis de asociación entre las variables categóricas versus la variable respuesta

Para determinar si existe independencia entre estas variables aplicaremos dos pruebas, el test estadístico Chi Cuadrado (χ^2) y el coeficiente V de Cramer, ambos pruebas permiten evaluar el nivel de asociación entre las variables categóricas y la variable respuesta Y. De esta forma, excluiremos a

las variables si no cumplen con los siguientes criterios: (i) que la variable no sea significativa, o (ii) que el valor V de Cramer no esté cercano a cero.

Cuadro 6.3: Nivel de asociación entre variables categóricas y variable respuesta

Variable	**Niveles**	**Estadística χ^2**	**valor p**	**V de Cramer**
Escala.O	14	17	**0.179**	**0.028**
Escala.U	14	39	0.000	**0.041**
Etapa	4	117	0.000	**0.072**
Nivel	12	16,575	0.000	0.855
Proceso	11	276	0.000	0.110
Tipo.de.colegio	3	49	0.000	**0.047**
Deuda	2	315	0.000	0.118
Genero	2	372	0.000	0.128
Beneficio	2	31	0.000	**0.037**
NSE	4	203	0.000	**0.095**
LIMAS	5	167	0.000	**0.086**
FACULTAD	12	654	0.000	0.170
Cambio.E	2	215	0.000	**0.097**

Test Chi Cuadrado

Si el valor p es menor que o igual al nivel de significancia, se rechaza la hipótesis nula y concluimos que hay una asociación estadísticamente significativa entre las variables categóricas y la variable respuesta Y

Para decidir si algunas de las variables categóricas son significativas o no, definimos en 1 % el nivel de significancia y lo comparamos con los valores p obtenidos. Observamos en el cuadro 6.3 que salvo la covariable Escala.O (escala de pago inicialmente asignada), el valor p del resto de covariables son menores al nivel de significancia. En consecuencia, decidimos excluir para posteriores análisis la covariable Escala.O.

Coeficiente V de Cramer

De acuerdo al coeficiente V de Cramer el criterio de decisión es el siguiente, si el coeficiente obtenido luego de aplicar la prueba presenta valores cercanos a cero es un indicador de una posible falta de asociación entre las variables, en consecuencia, éstas no serán consideradas en el estudio Así, a partir del del cuadro 6.3 las variables que decidimos excluir son: Escala.O, Escala.U, Etapa, Tipo.de.colegio, Beneficio, NSE, LIMAS, y Cambio.E.

6.2.2. Análisis de asociación entre las variables cuantitativas versus la variable respuesta. Test de Kruskal Wallis

Para definir el nivel de asociación de estas variables emplearemos el test de Kruskal Wallis que es una prueba no paramétrica que al contrario de la ANOVA no asume normalidad en los datos. Entonces, de acuerdo a esta prueba si el valor p obtenido es menor al nivel de significancia definido se rechaza la hipótesis nula de que las covariables inciden en la decisión de abandonar o no los estudios. Del cuadro 6.4 vemos que las variables que tienen su valor p mayor al nivel de siginificancia del 1 % son Cred.Aprob.2 y Cred.Aprob.4, por lo tanto, serán excluidas del análisis.

Cuadro 6.4: Nivel de relación entre variables cuantitativas y variable respuesta Y. Resultados de prueba Kruskal Wallis. Nivel de significancia $\alpha = 1\,\%$

Variable	**Estadística χ^2**	**Valor p**
CRAEST	6,697.3	0.000
Cred.Fal.Egreso	6,045.1	0.000
Nro.de.cred.matriculados.acumulado	8,371.8	0.000
N.Cred.Mat.UC	1,710.4	0.000
N.Cred.Aprob.2	0.3	**0.555**
Cred.Aprob.3	123.9	0.000
Cred.Aprob.4	4.6	**0.032**
Prom.Pond.UC	8,205.7	0.000
Edad.Egreso	5,373.7	0.000
Nro.postula	22.8	0.000
Importe.total.de.la.deuda	350.5	0.000
datos1_Años	7,821.0	0.000

6.2.3. Análisis de multicolinealidad entre las covariables

Para evaluar la presencia de colinealidad entre las variables haremos uso del factor de inflación de varianza (VIF, por sus siglas en inglés) que mide el incremento de la varianza como consecuencia de la presencia de colinealidad. De acuerdo a este criterio si una variable muestra un $VIF > 10$ es un indicador de la existencia de una alta colinealidad de la variable respecto de algunas de las otras. En el cuadro 6.5 observamos que las que superan este valor son Nivel, FACULTAD y Proceso, en consecuencia, serán excluida de la base de datos inicial.

Finalmente, luego de someter a a las variables a esta batería de pruebas estadísticas que nos permitió realizar una depuración de la base datos original, el número de covariables a emplear en las estimaciones de los métodos Bosque Aleatorio, Logístico, Potencia Logístico y Probit son las que se muestran en el cuadro 6.6, 9 variables cuantitativas (X_1 al X_{10}) y 2 categóricas (X_{11} X_{12}).

Cuadro 6.5: Factor de inflación de varianza (VIF)

Variable	VIF
CRAEST	2.6
Cred.Fal.Egreso	4.0
Nro.de.cred.matriculados.acumulado	5.7
N.Cred.Mat.UC	1.4
Cred.Aprob.3	1.9
Prom.Pond.UC	3.5
Edad.Egreso	4.2
Nivel	**14.5**
Nro.postula	2.1
Importe.total.de.la.deuda	2.1
datos1_Años	3.2
FACULTAD	**32.7**
Deuda	2.5
Proceso	**11.4**
Genero	1.2

Cuadro 6.6: Covariables para el modelo reducido

Nro.	Variables	Notación
1	CRAEST	X_1
2	Nro.de.cred.matriculados.acumulado	X_2
3	Cred.Fal.Egreso	X_3
4	N.Cred.Mat.UC	X_4
5	Cred.Aprob.3	X_5
6	Prom.Pond.UC	X_6
7	Edad.Egreso	X_7
8	Nro.postula.	X_8
9	Importe.total.de.la.deuda	X_9
10	datos1_Años	X_{10}
11	Deuda	X_{11}
12	Genero	X_{12}
13	Condicion	Y

6.3. Estimación de métodos paramétricos y bosque aleatorio

La estimación de los métodos paramétricos (Logístico, Potencia Logístico y Probit) se realizaron bajo el enfoque bayesiano utilizando la librería rstan disponible en R. Para la estimación de los modelos se considero adaptar 3000 iteraciones descartando las primeras 1000. Las 2000 iteraciones restantes son las que se tomaron en cuenta para las estimaciones, el número de cadenas definidas fueron dos y, finalmente, para evitar problemas de correlación se propuso un espaciamiento de 2.

Por su parte. el bosque aleatorio fue estimado utilizando la librería caret, disponible en R; la base de datos se dividió en datos de entrenamiento (70 %) y datos de evaluación (30 %) utilizando el método de validación cruzada, en nuestro caso particular, se tomaron en cuenta 5 subconjuntos (o folds). Además, cabe resaltar que con excepción del número de variables a considerar en cada árbol (3) y el número de árboles que conformarán el bosque definidas (200), ambos hiperparámetros definidos inicialmente, el resto de hiperparámetros considerados en la construcción del bosque son los que utiliza, por defecto, la librería.

Para poder concluir cuál es el método que muestra el mejor desempeño predictivo utilizaremos las métricas de evaluación estudiadas en el capítulo 2, teniendo en consideración sólo aquellas que resulten más idóneas para el nivel de desbalance observado en la base de datos, en nuestro caso del 30 %. De acuerdo a de la Cruz Huayanay et al. (2024) si el desbalance de clases es menor a 50 % (visto como proporción de unos) las métricas pertinentes para la evaluación del desempeño predictivo de un método de aprendizaje automático supervisado o no supervisado son TPR, CSI, SSI, FAITH, MCC, GM, F1 y KAPPA. Como en de la Cruz Huayanay et al. (2019); de la Cruz Huayanay (2023); de la Cruz Huayanay et al. (2024) tomaremos como umbral para cada método el que maximiza el valor de la métrica Kappa. Los resultados obtenidos se muestran en el cuadro 6.7 donde se aprecia que el método potencia logístico es el que muestra mejor desempeño predictivo.

Cuadro 6.7: Métricas de desempeño predictivo obtenidas según método aplicado

Modelo	Random Forest	Power Logístico	Logístico	Probit
TPR	0.813	0.863	0.870	**0.878**
CSI	0.752	**0.815**	0.799	0.783
SSI	0.602	**0.687**	0.666	0.644
FAITH	0.588	**0.597**	0.594	0.592
MCC	0.799	**0.860**	0.845	0.829
GM	0.884	**0.918**	0.916	0.914
F1	0.858	**0.898**	0.888	0.879
KAPPA	0.796	**0.858**	0.844	0.829

Estimación del modelo reducido

Habiendo concluido que el método potencia logística es el mejor en términos de desempeño predictivo mostramos a continuación en el cuadro 6.8 los resultados de las estimaciones de sus parámetros, desviación estandar e intervalo de credibilidad al 95 %.

Cuadro 6.8: Estimación de modelo reducido

Covariable	Parámetro	Estimado	Desviación Estandar	Intervalo de Credibilidad (95 %)
CRAEST	β_1	-0.234	0.049	(-0.326 , -0.138)
Nro.de.cred.matriculados.acumulado	β_2	-1.010	0.036	(-1.081 , -0.94)
Cred.Fal.Egreso	β_3	0.417	0.029	(0.36 , 0.475)
N.Cred.Mat.UC	β_4	-0.110	0.023	(-0.155 , -0.066)
Cred.Aprob.3	β_5	0.182	0.023	(0.135 , 0.229)
Prom.Pond.UC	β_6	-1.663	0.064	(-1.791 , -1.539)
Edad.Egreso	β_7	0.218	0.027	(0.165 , 0.269)
Nro.postula	β_8	-0.229	0.021	(-0.27 , -0.188)
Importe.total.de.la.deuda	β_9	0.252	0.028	(0.199 , 0.307)
datos1_Años	β_{10}	-1.452	0.044	(-1.536 , -1.372)
Deuda	β_{11}	0.167	0.072	(0.027 , 0.304)
Genero	β_{12}	0.029	0.044	(-0.057 , 0.117)
Parámetro de forma	λ	1.796	0.042	(1.716 , 1.883)

Interpretación de los resultados

En su versión reducida comprobamos que los valores estimados de los parámetros son significativos en su totalidad como se puede apreciar en el cuadro 6.8. Observamos que para todos los parámetros el intervalo de credibilidad al 95 % no contienen al cero. En particular, el parámetro de forma λ no contiene al 1, lo que explica el desbalance de los datos y su valor estimado da cuenta que el desbalance es menor al 50 % como proporción de 1´s.

La probabilidad de abandonar los estudios se reduce cuando las variables relacionadas con el rendimiento académico (CRAEST, Prom.Pond.UC) se incrementan. Por ejemplo, la covariable CRAEST mide el rendimiento académico acumulado del estudiante, observamos que tiene una relación inversa con la variable respuesta, en ese sentido, una mejora en el rendimiento del alumno, nos indica que la probabilidad de que el alumno abandone sus estudios se reduce.

Por su lado, si queremos evaluar la probabilidad de abandono en función de alguna variable vinculada a la situación económica del alumno podemos hacerlo analizando, por ejemplo, la relación entre la variable Importe.total.de.la.deuda y la variable respuesta, observamos que dicha relación es negativa. Por lo tanto, debemos esperar que en caso la deuda del alumno se incremente, la probabilidad de abandonar sus estudios también se incrementa.

Capítulo 7

Conclusiones

De la revisión de la literatura se evidenció que el problema del desbalance de datos es un tema de gran interés para los investigadores de aprendizaje automático (de inglés, machine learning) y mineria de datos (del inglés data mining).

Lo que encontramos en la literatura reciente es que de la multiplicidad de clasificadores utilizados para enfrentar el problema del desbalance son el Bosque Aleatorio y la Regresión Logística los de uso más frecuente y en el caso del primero el que mejor desempeño predictivo ha mostrado.

En esa línea, nuestra investigación abordó el problema de clasificación en un contexto de clases desbalanceadas desde dos enfoques, el paramétrico y el no paramétrico. Consideramos para el primer caso el método potencia logístico y, para el segundo, el bosque aleatorio.

Para nuestro estudio de simulación incluimos un análisis de sensibilidad y comparamos el comportamiento del método potencia logístico para diferentes prioris para el parámetro de forma λ y evaluamos su comportamiento frente a los métodos PROBIT y Logístico. Concluimos, en esta etapa, que el método potencia logístico es el que mejor ajuste presenta cuando el parámetro de forma tiene una distribución $Log(\lambda) \sim U(-2, 2)$.

Luego, para la determinación de cuál método presenta mejor desempeño predictivo hicimos la comparación de diferentes métricas de evaluación que se derivan de la Matriz de Confusión. Basándonos en el trabajo de de la Cruz Huayanay et al. (2024) empleamos las métricas adecuadas en función del desbalance observado. Así cuando la proporción de 1´s era menor al 50 % , utilizamos las métricas TPR, CSI, SSI, FAITH,MCC, GN F1 y KAPPA. Pero cuando enfrentamos la situación contraria, es decir cuando el desbalance era mayor a 1, es decir la proporción de 1´s es mayor al 50 %, lo conveniente es utilizar las métricas TNR, GS, MCC, GM y KAPPA.

El estudio de simulación desarrollado demostró que el método potencia logístico es el más apropiado para lidiar con el problema de clasificación en clases desbalanceadas y quedó evidenciado al comparar los resultados obtenidos para cada una de las métricas indicadas.

Además, en este trabajo realizamos una aplicación considerando datos reales sobre deserción universitaria con un desbalance del 30 %. Luego, hicimos el ajuste para cada uno de los métodos. En el caso de los métodos potencia logístico, logístico y probit se hicieron bajo el enfoque bayesiano,

usando la librería rstan y en el caso particular del bosque aleatorio usamos la librería caret, ambas disponibles en R.

La comparación de las métricas para evaluar el desempeño predictivo de los métodos con datos reales nos permitió concluir que el método potencia logística es el que obtiene mejores resultados, en comparación al logístico, probit y bosque aleatorio. Por su parte, en la estimación del modelo reducido, el valor estimado del parámetro de forma ($\hat{\lambda}$) presenta un intervalo de credibilidad que no contiene al 1, e incluso es mayor que este valor, lo que explica el desbalance observado en los datos.

Esta investigación podría ampliarse incluyendo otros métodos no paramétricos que han surgido como alternativa al bosque aleatorio. De hecho, encontramos investigaciones (Planinić y Popović-Bugarin (2024),Song et al. (2023)) que muestran que los métodos Gradient Boosting Machine (GBM), XGBoost y el CatBoost presentan un aceptable desempeño predictivo para hacer frente al desbalance de clases.

Frente al problema del desbalance de clases, la estrategia a nivel de datos ha sido la más utilizada. En ese sentido, un tema de investigación interesante seria la comparación del desempeño predictivo luego de la aplicación de estas técnicas en modelos paramétricos como no paramétricos.

De la misma forma, la estrategia a nivel de algoritmo ha sido muy poco tratada como posible solución de las dificultades asociadas al problema del desbalance, por lo que es una línea de investigación a profundizar.

Apéndice A

Demostración de parámetros fuerza y correlación de bosque aleatorio

Fuerza y correlación

Demostración de fuerza y correlación - tomada de Breiman (2001)

Definición 1

La función del margen (mg) para un bosque aleatorio fue definida en 3.3 como:

$$mg(X,Y) = P(h(X,\Theta) = Y) - \max_{j \neq Y} P(h(X,\Theta) = j \tag{A.1}$$

Ahora, si definimos la fuerza del conjunto de clasificadores $\{h(x,\Theta)\}$ como el valor esperado del margen, tendremos:

$$s = \mathbb{E}_{X,Y}[mg(X,Y)]. \tag{A.2}$$

Asumiendo que $s \geq 0$; a partir de la desigualdad de Chebychev se establece un limite superior para el error de generalización:

$$P(E^*) \leq \frac{\text{var}(mg)}{s^2} \tag{A.3}$$

Por otro lado, se puede tener una expresión para la varianza del mg partiendo de la siguiente expresión:

$$\hat{j}(X,Y) = \arg\max_{j \neq Y} P(h(X,\Theta) = j), \tag{A.4}$$

que corresponde a la máxima probabilidad de que el clasificador falle, entonces:

$$mg(X,Y) = P(h(X,\Theta) = Y) - P(h(X,\Theta) = \hat{j}(X,Y)), \tag{A.5}$$

$$= \mathbb{E}_\Theta \left[I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{j}(X,Y)) \right]. \tag{A.6}$$

Definición 2

La row margin rmg mide el margen para un clasificador específico Θ en lugar de su esperanza. Se define como:

$$rmg(\Theta, X, Y) = I(h(X, \Theta) = Y) - I(h(X, \Theta) = \hat{j}(X, Y)). \tag{A.7}$$

Por lo tanto, $mg(X, Y)$ es la esperanza de $rmg(\Theta, X, Y)$ con respecto a Θ. Sabemos que para cualquier función f, se cumple la identidad:

$$[\mathbb{E}_\Theta f(\Theta)]^2 = \mathbb{E}_{\Theta,\Theta'}[f(\Theta)f(\Theta')], \tag{A.8}$$

donde Θ, Θ' son independientes y tienen la misma distribución, lo que implica que:

$$mg(X, Y)^2 = \mathbb{E}_{\Theta,\Theta'}[rmg(\Theta, X, Y) \cdot rmg(\Theta', X, Y)] \tag{A.9}$$

Usando A.9, se obtiene:

$$\mathrm{var}(mg) = \mathbb{E}_{\Theta,\Theta'}\left[\mathrm{cov}_{X,Y}(rmg(\Theta, X, Y) \cdot rmg(\Theta', X, Y))\right], \tag{A.10}$$

$$= \mathbb{E}_{\Theta,\Theta'}[\rho(\Theta, \Theta') \cdot \mathrm{sd}(\Theta) \cdot \mathrm{sd}(\Theta')], \tag{A.11}$$

donde ($\rho(\Theta, \Theta')$ es la correlación entre $rmg(\Theta, X, Y)$ y $rmg(\Theta', X, Y)$ manteniendo Θ, Θ' fijos, y $\mathrm{sd}(\Theta)$ es la desviación estándar de $rmg(\Theta, X, Y)$ manteniendo Θ fijo. Entonces:

$$\mathrm{var}(mg) = \bar{\rho}(\mathbb{E}_\Theta[\mathrm{sd}(\Theta)])^2 \tag{A.12}$$

$$\leq \bar{\rho}\mathbb{E}_\Theta[\mathrm{var}(\Theta)], \tag{A.13}$$

donde $\bar{\rho}$ es el valor medio de la correlación:

$$\bar{\rho} = \frac{\mathbb{E}_{\Theta,\Theta'}[\rho(\Theta, \Theta') \cdot \mathrm{sd}(\Theta) \cdot \mathrm{sd}(\Theta')]}{\mathbb{E}_{\Theta,\Theta'}[\mathrm{sd}(\Theta) \cdot \mathrm{sd}(\Theta')]}. \tag{A.14}$$

Escribiendo:

$$\mathbb{E}_\Theta[\mathrm{var}(\Theta)] \leq \mathbb{E}_\Theta[(\mathbb{E}_{X,Y}[rmg(\Theta, X, Y)])^2] - s^2, \tag{A.15}$$

$$\leq 1 - s^2. \tag{A.16}$$

Combinando A.3, A.13 y A.16 se obtiene el límite superior para el error de generalización el cual viene dado por::

$$P(E^*) \leq \frac{\bar{\rho}(1 - s^2)}{s^2}. \tag{A.17}$$

Bibliografía

Abd Elrahman, S. M. y Abraham, A. (2013). A review of class imbalance problem, *Journal of Network and Innovative Computing* **1**: 9–9.

Ali, A., Shamsuddin, S. M. y Ralescu, A. L. (2013). Classification with class imbalance problem, *Int. J. Advance Soft Compu. Appl* **5**(3): 176–204.

Bazán Guzmán, J. L., Valdivieso Serrano, L. H. y Calderón García, A. (2010). Enfoque bayesiano en modelos de teoría de respuesta al ítem.

Bazán, J. L., Bolfarine, H. y Branco, M. D. (2005). A general skew-probit link for binary response, *Proceedings of the 9th School of Regression Models* pp. 267–81.

Bazan, J. L. y Millones, Ó. (2008). Una clasificación de modelos de regresión binaria asimétrica: el uso del bayes-pucp en una aplicación sobre la decisión del cultivo ilícito de hoja de coca, *Economía* **31**(62): 17–32.

Bazán, J., Torres-Avilés, F., Suzuki, A. K. y Louzada, F. (2017). Power and reversal power links for binary regressions: An application for motor insurance policyholders, *Applied Stochastic Models in Business and Industry* **33**(1): 22–34.

Bharadwaj, S. (2023). Credit card fraud detection using machine learning, *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, pp. 168–172.

Biau, G. (2012). Analysis of a random forests model, *The Journal of Machine Learning Research* **13**(1): 1063–1095.

Biau, G. y Scornet, E. (2016). A random forest guided tour, *Test* **25**: 197–227.

Breiman, L. (1996). Out-of-bag estimation.

Breiman, L. (2001). Random forests, *Machine learning* **45**: 5–32.

Breiman, L. (2017). *Classification and regression trees*, Routledge.

Bujang, S. D. A., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C. y Fujita, H. (2022). Imbalanced classification methods for student grade prediction: A systematic literature review, *IEEE Access* **11**: 1970–1989.

Cecchini, V., Nguyen, T.-P., Pfau, T., De Landtsheer, S. y Sauter, T. (2019). An efficient machine learning method to solve imbalanced data in metabolic disease prediction, *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, IEEE, pp. 1–5.

Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview, *Data mining and knowledge discovery handbook* pp. 875–886.

Chen, C., Liaw, A. y Breiman, L. (2004). Using random forest to learn imbalanced data, *Technical Report 666*, Department of Statistics, UC Berkley.
URL: *http://xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html*

Chen, M.-H., Dey, D. K. y Shao, Q.-M. (1999). A new skewed link model for dichotomous quantal response data, *Journal of the American Statistical Association* **94**(448): 1172–1186.

Cicak, S. y Avci, U. (2023). Handling imbalanced data in predictive maintenance: A resampling-based approach, *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, pp. 1–6.

Cutler, A., Cutler, D. R. y Stevens, J. R. (2012). Random forests, *Ensemble machine learning: Methods and applications* pp. 157–175.

De La Cruz Huayanay, A. (2019). Modelos de regressão para resposta binária na presença de dados desbalanceados.

de la Cruz Huayanay, A. (2023). Modelos alternativos para classificação em dados desbalanceados.

de la Cruz Huayanay, A., Bazan, J. L., Cancho, V. G. y Dey, D. K. (2019). Performance of asymmetric links and correction methods for imbalanced data in binary regression, *Journal of Statistical Computation and Simulation* **89**(9): 1694–1714.

de la Cruz Huayanay, A., Bazán, J. L. y Russo, C. M. (2024). Performance of evaluation metrics for classification in imbalanced data, *Computational Statistics* pp. 1–27.

Gelman, A., Carlin, J. B., Stern, H. S. y Rubin, D. B. (1995). *Bayesian data analysis*, Chapman and Hall/CRC.

Genuer, R., Poggi, J.-M. y Tuleau, C. (2008). Random forests: some methodological insights, *arXiv preprint arXiv:0811.3619* .

Ghosh, S. (2024). Comparing regular random forest model with weighted random forest model for classification problem.

Goyal, M. y Kumar, R. (2020). Machine learning for malware detection on balanced and imbalanced datasets, *2020 International Conference on Decision Aid Sciences and Application (DASA)*, IEEE, pp. 867–871.

Guo, X., Yin, Y., Dong, C., Yang, G. y Zhou, G. (2008). On the class imbalance problem, *2008 Fourth international conference on natural computation*, Vol. 4, IEEE, pp. 192–201.

Jafarigol, E. y Trafalis, T. (2023). A review of machine learning techniques in imbalanced data and future trends, *arXiv preprint arXiv:2310.07917* .

Karim, M., Samad, M. y Muntasir, F. (2022). Improving performance factors of an imbalanced credit risk dataset using smote, *2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, IEEE, pp. 1–4.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions, *Progress in Artificial Intelligence* **5**(4): 221–232.

Louppe, G. (2014). Understanding random forests: From theory to practice, *arXiv preprint arXiv:1407.7502* .

Lunn, D. J., Thomas, A., Best, N. y Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility, *Statistics and computing* **10**: 325–337.

O Brien, R. y Ishwaran, H. (2019). A random forests quantile classifier for class imbalanced data, *Pattern recognition* **90**: 232–249.

Ortiz Lozano, J. M., Rúa Vieites, A. y Bilbao Calabuig, M. P. (2017). Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas.

Peargin, G. (2019). *Random Forest for Prediction with Unbalanced Data*, PhD thesis, Hochschule für Angewandte Wissenschaften München.

Planinić, D. y Popović-Bugarin, V. (2024). Credit card fraud detection using supervised learning algorithms, *2024 28th International Conference on Information Technology (IT)*, IEEE, pp. 1–4.

Probst, P., Wright, M. N. y Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest, *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* **9**(3): e1301.

Qaddoura, R. y Biltawi, M. M. (2022). Improving fraud detection in an imbalanced class distribution using different oversampling techniques, *2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*, IEEE, pp. 1–5.

Rekha, G., Tyagi, A. K., Sreenath, N. y Mishra, S. (2021). Class imbalanced data: Open issues and future research directions, *2021 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, pp. 1–6.

Rosly, M. A. B. M., Zainudin, S. y Kassim, J. M. (2023). Analysing imbalanced dataset for postgraduate student dropout using predictive analytics, *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, pp. 1–6.

Song, Z., Sung, S.-H., Park, D.-M. y Park, B.-K. (2023). All-year dropout prediction modeling and analysis for university students, *Applied Sciences* **13**(2): 1143.

Véliz Capuñay, C. (2023). *Aprendizaje automático : introducción al aprendizaje profundo*, Pontificia Universidad Católica del Perú, Fondo Editorial.

Vitório, A. y Marques, G. (2021). Impact of imbalanced data on bank telemarketing calls outcome forecasting using machine learning, *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, IEEE, pp. 380–384.

Weng, C. G. y Poon, J. (2008). A new evaluation measure for imbalanced datasets, *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pp. 27–32.

Wu, O. y Li, M. (2024). Revisiting the effective number theory for imbalanced learning, *IEEE Transactions on Knowledge & Data Engineering* (01): 1–14.

Wundervald, B. (2019). Bayesian linear regression.

Zhao, Y., Wong, Z. S.-Y., Tsui, K. L. et al. (2018). A framework of rebalancing imbalanced healthcare data for rare events classification: a case of look-alike sound-alike mix-up incident detection, *Journal of healthcare engineering* **2018**.