COMPARACIÓN DE TÉCNICAS BASADAS EN MUESTREO PARA SOLUCIONAR EL PROBLEMA DE DOS CLASES DESBALANCEADAS

Por

MILENA ROSA SALCEDO VILLANUEVA

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRÍA EN CIENCIAS
en
MATEMÁTICAS (ESTADÍSTICA)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGÜEZ
2011

Aprobado por:

_____	_____
Dámaris Santana Morant, Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Edgardo Lorenzo González Ph.D. Miembro, Comité Graduado	Fecha
_____	_____
Julio C Quintana Díaz, Ph.D. Presidente, Comité Graduado	Fecha
_____	_____
Linda Wessel Beaver, Ph.D. Representante de Estudios Graduados	Fecha
_____	
Silvestre Colón Ramírez M.S Director Interino Departamento de Ciencias Matemáticas	_____ Fecha

ABSTRACT

Nowadays, estimation and evaluation of classification models has increased as an area of research in pattern recognition in data bases.

One of the main problems that contribute to performance deterioration of classification methods in relation to data sets occurs when classes are unbalanced, that is one or several classes have sizes significantly bigger than the others. Here particular attention has been given to the case where data are distributed in two unbalanced classes.

In this thesis we present a comparative analysis of the effect of the use of sampling techniques to solve the problem of two unbalanced classes. The techniques analyzed were: random oversampling; oversampling "SMOTE" (Synthetic Minority Over-sampling Technique); and combinations of oversampling "SMOTE" with the cleaning methods "ENN" (Edited Nearest Neighbor), and "Tomek-Link" (these last techniques also act as undersampling procedures). We evaluated the effects of their implementations on the following classification methods: logistic regression; linear discriminant; k-nearest neighbors; and decision trees. The purpose was to establish which of these methods showed better performance based on the results of the following evaluation metrics: misclassification rate; and the measures of "Noise", "Silence", "G" (based on the geometric mean) and "F". The data sets that we used were: "CRX" and "GERMAN", located at the webpage of Dr. Edgar Acuña, and the data sets named "EST1" y "EST2".

The combination of the oversampling “SMOTE” technique with the cleaning method ENN applied to these data sets was the most efficient in those cases where unbalanced sizes between two classes were significant.

RESUMEN

En la actualidad la estimación y evaluación de modelos de clasificación se ha constituido en una línea fuerte de investigación en el reconocimiento de patrones en bases de datos.

Uno de los problemas principales que contribuyen al deterioro en el desempeño de los métodos de clasificación en relación al conjunto de datos ocurre cuando las clases están desbalanceadas. Es decir, una o varias clases tienen tamaños significativamente mayores que las otras. Se le ha dado particular atención al caso cuando los datos se encuentran distribuidos en dos clases y el tamaño de una ellas es significativamente mayor que la otra.

En esta tesis se presenta un análisis comparativo del efecto que tiene el uso de técnicas basadas en muestreo para solucionar el problema de dos clases desbalanceadas. Las técnicas analizadas fueron sobre-muestreo aleatorio, sobre-muestreo SMOTE (Synthetic Minority Over-sampling Technique) y combinaciones de sobre-muestreo "SMOTE" con los métodos de limpieza "ENN" (Edited Nearest Neighbor), y "Tomek Link" (que actúan también como técnicas de sub-muestreo). Se aplicaron estos procedimientos sobre los siguientes métodos de clasificación: regresión logística, discriminante lineal, vecinos más cercanos y árboles de decisión. El propósito fue establecer entre estos métodos de clasificación cuál mostró mejor desempeño al aplicar las siguientes métricas de evaluación: tasa de error, y las medidas: G (basada en la

media geométrica), “Ruido”, “Silencio” y “F”(basada en la media armónica). Para realizar este estudio se utilizaron los conjuntos de datos “CRX” y “GERMAN” que se encuentran en la página web del Dr. Edgar Acuña y los conjuntos de datos “EST1” y “EST2”. La combinación de la técnica de sobre-muestreo SMOTE y el método de limpieza de Wilson (ENN) en estos datos resultó ser más eficiente en aquellos casos donde el desbalance de las dos clases es muy significativo.

Derechos reservados © 2011
por
Milena R. Salcedo Villanueva

A Dios todo poderoso por su infinito amor

A mis padres Laudina Villanueva y David Salcedo J por su apoyo y amor incondicional.

A Reyes M Ortiz, por su amor, apoyo y comprensión.

A mis hermanas, sobrinos y amigos por siempre estar presentes.

AGRADECIMIENTOS

Quiero dar gracias a papito Dios porque siempre ha estado junto a mí, recordándome que las cosas se pueden lograr no importando los obstáculos que se nos presenten, por ayudarme a vencer cada uno esos obstáculos y por llevarme de su mano mostrándome siempre el camino a seguir.

Quiero dar gracias a mi familia por todo el apoyo incondicional y confianza que siempre me han brindado, pero sobre todo por su cariño; aún en la distancia, ustedes han sido el motor para el cumplimiento de todas mis metas alcanzadas.

Quiero dar gracias de manera muy especial al Dr. Julio C. Quintana, mi consejero, por su apoyo en la culminación de esta meta, por sus enseñanza de durante todos mis años de estudios, por su tiempo, paciencia, orientación durante todo este proceso.

Quiero dar gracias a mis amigos Román Kvasov y Elmer Rodríguez, por su colaboración en la realización de este trabajo, por su amistad y apoyo incondicional durante todo el tiempo de estudio de Maestría.

Quiero dar gracias a la facultad, a mis compañeros, al personal del Departamento de Ciencias Matemáticas del Recinto Universitario de Mayagüez, por hacerme sentir parte de una gran familia y brindarme siempre su apoyo al cursar mis estudios de Maestría, en especial a Carmen González.

TABLA DE CONTENIDO

ABSTRACT II

RESUMEN IV

AGRADECIMIENTOS VIII

TABLA DE CONTENIDO IX

LISTA DE FIGURAS XII

LISTA DE TABLAS XIII

LISTA DE GRÁFICOS XV

LISTA DE ALGORITMOS XVI

1 INTRODUCCION 1

1.1 JUSTIFICACIÓN 1

1.2 OBJETIVOS 3

1.3 RESUMEN DE CAPÍTULOS 3

2 REVISIÓN LITERARIA 5

2.1 CLASIFICACIÓN SUPERVISADA 6

2.2 MÉTODOS DE CLASIFICACIÓN SUPERVISADA 6

2.2.1 REGRESIÓN LOGÍSTICA 7
2.2.2 DISCRIMINANTE LINEAL 9
2.2.3 ÁRBOLES DE DECISIÓN 17
2.2.4 REGLA DEL VECINO MÁS CERCANS (k-NN) 23

3 MÉTODOS DE MUESTREO PARA SOLUCIONAR EL PROBLEMA DE DOS CLASES DESBALANCEADAS 29

3.1 PROBLEMA DE DOS CLASES DESBALANCEADAS 29

3.2 TECNICAS DE MUESTREO PARA BALANCEAR UN CONJUNTO DE DATOS. 30

3.2.1 TÉCNICAS DE SOBRE-MUESTREO (OVERSAMPLING) 30
3.2.2 SMOTE 31
3.2.3 ALGORITMO DE EDICION DE WILSON (ENN) 34
3.2.4 ALGORITMO TOMEK LINKS 35

4 METODOLOGÍA 37

4.1 CRITERIOS PARA LA EVALUACIÓN DE ALGORITMOS DE CLASIFICACIÓN 37

4.1.1 MATRIZ DE CONFUSIÓN PARA DOS CLASES 38
4.1.2 EVALUACIÓN CLÁSICA 40
4.1.3 COBERTURA Y PRECISIÓN 40
4.1.4 MEDIDA F 41
4.1.5 SENSITIVIDAD Y ESPECIFICIDAD 41
4.1.6 EVALUACIÓN BASADA EN LA MEDIA GEOMÉTRICA: 42
4.1.7 SILENCIO Y RUIDO 43

4.2 DATOS 44
4.2.1 CRX 44
4.2.2 GERMAN 45
4.2.3 EST1 45
4.2.4 EST2 46
5 RESULTADOS 47
5.1 RESULTADOS PARA LOS DATOS "CRX" 47
5.1.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO CRX 47
5.1.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN 50
5.1.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE" 53
5.1.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN) 55
5.1.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "TOMEK LINK" 58
5.2 RESULTADOS PARA LOS DATOS "GERMAN" 62
5.2.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO GERMAN 62
5.2.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN 64
5.2.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE" 67
5.2.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN) 69
5.2.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "TOMEK LINK" 71
5.3 RESULTADOS PARA LOS DATOS "EST2" 74
5.4 RESULTADOS PARA LOS DATOS "EST1" 75
5.4.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO EST1 75
5.4.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN 78

5.4.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE" 80

5.4.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN) 83

5.4.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "TOMEK LINKS" 85

6 CONCLUSIONES 89

REFERENCIAS 91

LISTA DE FIGURAS

Figuras **Páginas**

Figura 1 Proceso de clasificacion supervisada....6
Figura 2 Sensitividad y especificidad para varios niveles de probabilidad....8
Figura 3 Curva ROC (Receiver Operating Characteristic)....9
Figura 4 Ejemplo del algoritmo *k-nn*....25
Figura 5 Base de datos con clases desbalanceadas....29
Figura 6 Matriz de confusión para dos clases....38
Figura 7 Matriz de confusión-diagonal de aciertos....39

LISTA DE TABLAS

Tablas **Páginas**

Tabla 1 Matriz de confusión de los clasificadores aplicados a CRX 47

Tabla 2 Métricas de evaluación de clasificadores aplicados a CRX 49

Tabla 3 Matriz de confusión de los clasificadores aplicados a los datos CRX balanceados a través de la técnica "sobre-muestreo aleatorio con reposición" 50

Tabla 4 Métricas de evaluación de clasificadores aplicados a los datos CRX balanceados a través de la técnica "sobre-muestreo aleatorio con reposición" 52

Tabla 5 Matriz de confusión de los clasificadores aplicados a los datos CRX balanceados a través de la técnica "SMOTE" 53

Tabla 6 Métricas de evaluación de clasificadores aplicados a los datos CRX balanceados a través de la técnica "SMOTE" 55

Tabla 7 Matriz de confusión de los clasificadores aplicados a los datos CRX balanceados a través de la combinación de "SMOTE Y ENN" 55

Tabla 8 Métricas de evaluación de clasificadores aplicados a los datos CRX balanceados a través de la técnica "SMOTE Y ENN" 57

Tabla 9 Matriz de confusión de los clasificadores aplicados a los datos crx balanceados a través de la combinación de "SMOTE Y TOMEK LINKS " 58

Tabla 10 Métricas de evaluación de clasificadores aplicados a los datos crx balanceados a través de la técnica "SMOTE Y TOMEK LINKS" 60

Tabla 11 Resultados de las métricas de evaluación generales aplicadas a regresión logística sobre los datos "CRX" balanceados 60

Tabla 12 Resultados de las métricas de evaluación generales aplicadas a árboles de decisión sobre los datos "CRX" balanceados 61

Tabla 13 Matriz de confusión de los clasificadores aplicados a GERMAN 62

Tabla 14 Métricas de evaluación de clasificadores aplicados a GERMAN 63

Tabla 15 Matriz de confusión de los clasificadores aplicados a los datos GERMAN balanceados a través de la técnica "sobre-muestreo aleatorio con reposición" 64

Tabla 16 Métricas de evaluación de clasificadores aplicados a los datos GERMAN balanceados a través de la técnica "sobre-muestreo aleatorio con reposición" 66

Tabla 17 Matriz de confusión de los clasificadores aplicados a los datos GERMAN balanceados a través de la técnica "SMOTE" 67

Tabla 18 Métricas de evaluación de clasificadores aplicados a los datos GERMAN balanceados a través de la técnica "SMOTE" 68

Tabla 19 Matriz de confusión de los clasificadores aplicados a los datos GERMAN balanceados a través de la combinación de "SMOTE Y ENN" 69

Tabla 20 Métricas de evaluación de clasificadores aplicados a los datos GERMAN balanceados a través de la técnica “SMOTE y ENN” 70

Tabla 21 Matriz de Confusión de los clasificadores aplicados a los datos GERMAN balanceados a través de la combinación de “SMOTE y Tomek Link “ 71

Tabla 22 Métricas de evaluación de clasificadores aplicados a los datos GERMAN balanceados a través de la técnica “SMOTE y Tomek Link” 73

Tabla 23 Resultados de las métricas de evaluación generales aplicadas a Regresión Logística sobre los datos “GERMAN” balanceados 74

Tabla 24 Resultados de las métricas de evaluación generales aplicadas a Árboles de Decisión sobre los datos “EST2” balanceados 74

Tabla 25 Matriz de Confusión de los clasificadores aplicados a EST1. 75

Tabla 26 Métricas de evaluación de clasificadores aplicados a EST1 77

Tabla 27 Matriz de Confusión de los clasificadores aplicados a los datos EST1 balanceados a través de la técnica “sobre-muestreo aleatorio con reposición” 78

Tabla 28 Métricas de evaluación de clasificadores aplicados a los datos EST1 balanceados a través de la técnica “sobre-muestreo aleatorio con reposición” 80

Tabla 29 Matriz de Confusión de los clasificadores aplicados a los datos EST1 balanceados a través de la técnica “SMOTE” 80

Tabla 30 Métricas de evaluación de clasificadores aplicados a los datos EST1 balanceados a través de la técnica “SMOTE” 82

Tabla 31 Matriz de Confusión de los clasificadores aplicados a los datos EST1 balanceados a través de la combinación de “SMOTE y ENN” 83

Tabla 32 Métricas de evaluación de clasificadores aplicados a los datos EST1 balanceados a través de la técnica “SMOTE y ENN” 85

Tabla 33 Matriz de Confusión de los clasificadores aplicados a los datos EST1 balanceados a través de la combinación de “SMOTE y Tomek Link ” 85

Tabla 34 Métricas de evaluación de clasificadores aplicados a los datos EST1 balanceados a través de la técnica “SMOTE y Tomek Link” 87

Tabla 35 Resultados de las métricas de evaluación generales aplicadas a Vecinos cercanos sobre los datos “EST1” balanceados 87

LISTA DE GRÁFICOS

Gráficos **Páginas**

Gráfico 1 Sensitividad Y Especificidad para varios valores de ***p*** "CRX" 48

Gráfico 2 Sensitividad y Especificidad para varios valores de ***p*** "CRX" Sobre-muestreo 51

Gráfico 3 Sensitividad y Especificidad para varios valores de ***p*** "CRX-SMOTE".....54

Gráfico 4 Sensitividad y Especificidad para varios valores de ***p*** "CRX-SMOTE+ENN".....56

Gráfico 5 Sensitividad y Especificidad para varios valores de ***p*** "CRX- Smote+Tomek Links" 59

Gráfico 6 Sensitividad y Especificidad para varios valores de ***p*** "GERMAN" 63

Gráfico 7 Sensitividad y Especificidad para varios valores de ***p*** "GERMAN- Sobremuestreo" 65

Gráfico 8 Sensitividad y Especificidad para varios valores de ***p*** "GERMAN-SMOTE".....68

Gráfico 9 Sensitividad y Especificidad para varios valores de ***p*** "GERMAN-SMOTE+ENN".....70

Gráfico 10 Sensitividad y Especificidad para varios valores de ***p*** "GERMAN- Smote+Tomek Links" 72

Gráfico 11 Sensitividad y Especificidad para varios valores de ***p*** "EST1" 76

Gráfico 12 Sensitividad y Especificidad para varios valores de ***p*** "EST1" Sobre-muestreo 79

Gráfico 13 Sensitividad y Especificidad para varios valores de ***p*** "EST1-SMOTE".....81

Gráfico 14 Sensitividad y Especificidad para varios valores de ***p*** "EST1-SMOTE+ENN".....84

Gráfico 15 Sensitividad y Especificidad para varios valores de ***p*** "EST1- Smote+Tomek Links" 86

LISTA DE ALGORITMOS

Algoritmos **Páginas**

Algoritmo 1 SMOTE 33
Algoritmo 2 "ENN" WILSON 35
Algoritmo 3 TOMEK LINKS 36

1 INTRODUCCIÓN

1.1 JUSTIFICACIÓN

En la actualidad la estimación de modelos de clasificación se ha constituido en una línea fuerte de investigación en el reconocimiento de patrones en bases de datos.

El desempeño de los métodos que estiman modelos de clasificación depende, no sólo de éstos posean elementos suficientemente precisos, sino también de la calidad del conjunto de datos.

Algunos de los factores principales que deterioran el desempeño de los métodos de clasificación en relación al conjunto de datos son: presencia de datos ruidosos o atípicos, datos redundantes, bases de datos excesivamente grandes y el desbalance entre la distribución de clases.

Siendo el desbalance entre la distribución de clases uno de los factores que debilitan el desempeño de los métodos de clasificación, es importante analizar de qué manera se debe proseguir al encontrarnos con situaciones como éstas, específicamente cuando las bases de datos se encuentran distribuidas en dos clases.

Existe una gran dificultad al realizar predicciones cuando se trata de datos con dos clases desbalanceadas, es decir cuando la cantidad de instancias que corresponde a una determinada clase (clase mayoritaria) es mucho más grande que la cantidad de

instancias que tiene la otra clase (clase minoritaria). En este caso los métodos de clasificación estándares tienden a clasificar nuevas instancias en la clase mayoritaria.

Este problema de clases no balanceadas aparece frecuentemente el campo de empresas de telecomunicaciones y empresas bancarias en donde los clientes juegan un papel importante. Allí encontramos que hay clientes que permanecen activos mientras que otros abandonan el servicio por muchas razones. En este caso el desequilibrio es muy pronunciado, debido a que los clientes que abandonan el servicio (churn) son muy pocos tal vez menos del 10% de los clientes que permanecen activos (no churn) y corresponderían a la clase pequeña.

Las áreas de aplicación de reconocimiento de patrones en este tipo de dominios son muy diversas, entre éstas se pueden mencionar: medicina, específicamente en diagnóstico temprano de enfermedades; educación, en resultados de reprobación y aprobación de cursos y semestres; mercadeo, en la detección de clientes activos y pasivos; procesamiento de imágenes satelitales, reconocimientos de caracteres escritos entre otras. Durante la toma de decisiones, la asistencia inteligente de los métodos de clasificación resulta ser de vital importancia, porque de ella depende en particular el éxito o fracaso de los procesos.

Para remediar este problema de desequilibrio de dos clases se han propuestos diferentes métodos tanto a nivel de datos como a nivel de algoritmo. Este trabajo se concentró en las técnicas al nivel de datos (muestreo) tales como: sobre-muestreo

aleatorio con reposición, sobre-muestreo SMOTE y técnicas híbridas: combinación de SMOTE y métodos de limpieza “ENN” y “Tomek Links”, con el fin de aplicarlas a algunos métodos de clasificación que se adaptan a datos en este dominio: árboles de decisión, análisis de discriminante lineal, vecinos más cercanos y regresión logística.

1.2 OBJETIVOS

- Comparar los resultados obtenidos al aplicar las técnicas de muestreo: sub-muestreo aleatorio, sobre-muestreo “SMOTE” y técnicas híbridas: combinación de “SMOTE” y métodos de limpieza “ENN” y “Tomek Links”, con el fin de establecer cuál de éstas arroja mejores resultados.

- Analizar los resultados que se obtienen al implementar los métodos de clasificación en los conjuntos de datos ya balanceados con las diferentes técnicas de muestreo (sobre-muestreo e híbrido) y determinar los métodos que presentan mejor desempeño.

1.3 RESUMEN DE CAPÍTULOS

Esta tesis se encuentra organizada en seis capítulos. El primer capítulo de introducción hace referencia al problema que motivó este trabajo de investigación, los objetivos que se trazaron y se describen brevemente las partes que componen la tesis. En el segundo capítulo se hace una revisión previa de definiciones y conceptos que nos permiten tener

un mejor entendimiento del desarrollo de este trabajo de investigación. El tercer capítulo describe de manera detallada los diferentes métodos de muestreo que se utilizaron para solucionar el problema de distribución de dos clases desbalanceadas. El capítulo cuarto muestra la metodología utilizada, los criterios de evaluación para establecer el clasificador con mejor desempeño. En el capítulo cinco se muestran los resultados y en el capítulo seis las conclusiones de todo el proceso.

2 REVISIÓN LITERARIA

Con el progreso de la ciencia y la tecnología es muy común encontrar muchas empresas, agencias o instituciones que recopilan grandes cantidades de datos con diversas características; las cuales son almacenadas en bases de datos. La extracción de información oculta y predecible de dichas bases de datos es lo que se conoce como Minería de Datos. Esta disciplina se encarga de hacer específicamente las siguientes operaciones:

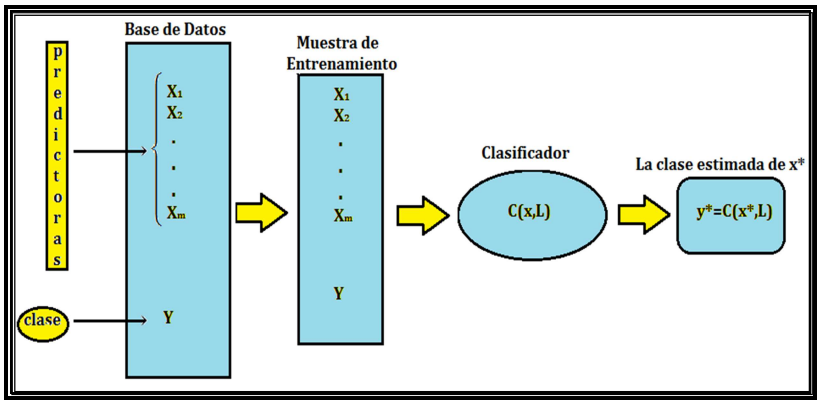
1. Realizar análisis exploratorio de datos.
2. Identificar patrones y tendencias (descubrir relaciones entre los datos).
3. Realizar predicciones a partir del análisis exploratorio y de la identificación de patrones y tendencias.

En este sentido las tareas de minería de datos pueden ser: descriptivas y predictivas. En las descriptivas se descubren las propiedades generales y características más importantes de la base de datos, las cuales se realizan por medio de técnicas tales como: formación de conglomerados "clustering", reglas de asociación, detección de datos fuera de lo común "outliers" y visualización, entre otros. En las predictivas se estima un modelo usando los datos recolectados para hacer predicciones futuras. Éstas últimas nunca son 100% precisas y lo importante es el rendimiento del modelo cuando se aplica a nuevos datos. Entre algunas tareas predictivas se pueden mencionar: regresión y clasificación supervisada.

2.1 CLASIFICACIÓN SUPERVISADA

La clasificación supervisada es un proceso en el cual se construye un modelo basado en una muestra de entrenamiento, donde se supone que cada objeto pertenece a una clase predefinida. En ciertos casos, en lugar de usar una muestra de entrenamiento se utiliza el conjunto completo en la base de datos. Este modelo se representa por medio de reglas de clasificación, árboles de decisión o fórmulas matemáticas y se usa para clasificar nuevos datos. Estos datos que se predicen son valores categóricos que representan clases, mientras que la regresión modela funciones de valores continuos para predecir valores desconocidos o faltantes también continuos (Figura 1).

Figura 1. PROCESO DE CLASIFICACIÓN SUPERVISADA



2.2 MÉTODOS DE CLASIFICACIÓN SUPERVISADA

Existen algunos métodos de clasificación supervisada que se han adaptado a bases de datos que son clasificados en dos clases entre los cuales se pueden mencionar: árboles de decisión, análisis de discriminante lineal, vecinos más cercanos y regresión logística.

2.2.1 REGRESIÓN LOGÍSTICA

La regresión logística fue sugerida inicialmente por Cox (1970) y se considera como una técnica estadística multivariante que permite estudiar la relación existente entre una variable dependiente cualitativa binaria y un grupo de variables explicativas las cuales pueden ser cuantitativas o cualitativas.

Consideremos un conjunto de datos que se encuentra clasificado en dos clases; que contiene n observaciones, de las cuales n_1 observaciones pertenecen a la clase C_1 y n_2 pertenecen a la clase C_2. Para cada observación x_j se introduce una variable binaria Y que vale 1 si es de la clase C_1 y 0 si la observación pertenece a la clase C_2.

Sea $p = P(Y = 1|x)$. El modelo de regresión logística está dado por:

$$ln\left(\frac{p}{1-p}\right) = \alpha + \beta' X \quad (Transformación\ logit) \tag{1}$$

ó

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}} = \frac{e^{\alpha+\beta' X}}{1+e^{\alpha+\beta' X}} \tag{2}$$

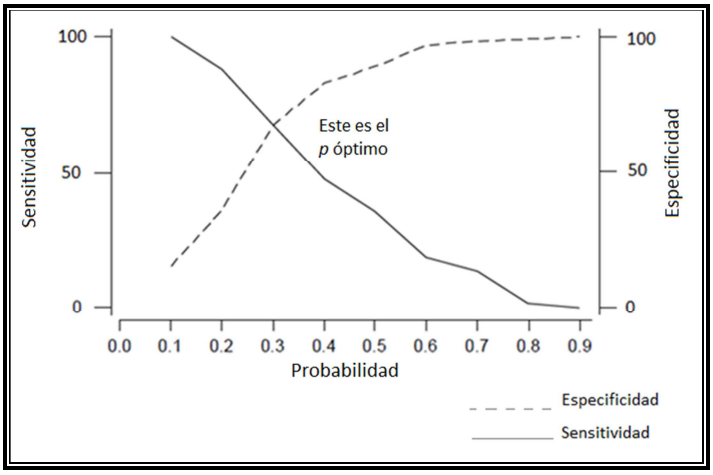
Esta última ecuación es llamada regresión logística y también cumple para otros tipos de distribuciones distintas de la normal multivariada, como es el caso de distribuciones de Bernoulli y mezclas de éstas.

2.2.2.1 Uso de la clasificación logística en la clasificación

La manera más fácil y usual para discriminar los datos con regresión logística es considerar $p > 0.5$ si la observación pertenece a la clase de interés. Pero esta manera de clasificar resulta ser un método heurístico, por no tener un criterio objetivo. Para encontrar el valor óptimo de p de manera objetiva se pueden utilizar métodos alternos como:

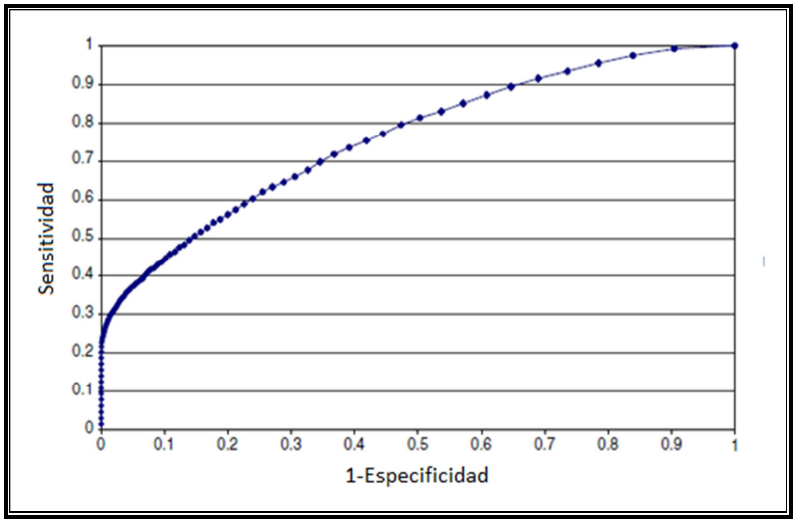
- Graficar el porcentaje de instancias que fueron clasificadas correctamente en la clase de C_1 con respecto al total de instancias que realmente pertenecen a la clase C_1 (sensitividad) versus distintos niveles de probabilidad y graficar el porcentaje de instancias que fueron clasificadas correctamente en la clase de C_2 con respecto al total de instancias que realmente pertenecen a la clase C_2 (especificidad) versus los mismos niveles de probabilidad utilizados para la gráfica de sensitividad. El punto de intersección entre las curvas de sensitividad y especificidad es el valor de p óptimo usado para clasificar una nueva instancia (Figura 2).

Figura 2. SENSITIVIDAD Y ESPECIFIDAD PARA VARIOS NIVELES DE PROBABILIDAD



- Graficar la curva ROC (Receiver Operating Characteristic), es decir los valores de "sensitividad" versus "1-especificidad". El valor de p a utilizar es aquel que se encuentra más cerca de la esquina superior izquierda [2] (Figura 3).

Figura 3. CURVA ROC (RECEIVER OPERATING CHARACTERISTIC)



2.2.2 DISCRIMINANTE LINEAL

Es una técnica estadística que permite analizar simultáneamente las diferencias entre dos ó más grupos con respecto a un conjunto de variables y asignar o clasificar instancias (objetos) en el grupo que le es más próximo [Klecka, 1982] [3].

Para llevar a cabo este análisis debemos partir de los siguientes supuestos:

- La existencia de dos o más grupos que difieran sobre varias variables y que éstas estén dadas en escala de intervalo o razón.

- El número p de variables discriminantes debe ser menor que la diferencia entre el número de objetos y el número de clases: $x_{1,}, x_2, \dots x_p$. Es decir $p < (n - k)$donde n es el número de objetos y k es el número de clases.
- La clasificación a priori de las instancias u objetos es mutuamente excluyente.
- Las variables independientes deben tener una distribución normal multivariada.
- La matriz de varianza-covarianza de las distintas poblaciones de las que provienen los grupos deben ser aproximadamente iguales.
- Ninguna variable discriminante es combinación lineal de otras variables discriminantes.

Bajo estos requisitos, el análisis discriminante se utiliza principalmente para dos fines:

1. Identificar variables que mejor describan la pertenencia al grupo. *Análisis Discriminante Descriptivo (ADD).*
2. Clasificar a las instancias en dos o más grupos de antemano conocidos. El interés en este caso es poder asignar de forma óptima una nueva instancia a cualquiera de los grupos preestablecidos*. Análisis discriminante predictivo (ADP)* [3].

El objetivo fundamental del análisis discriminante es encontrar una regla de clasificación de instancias u objetos en un conjunto de datos previamente establecidos teniendo en cuenta una serie de características (variables) observadas sobre tales instancias u objetos. En otras ocasiones se quiere obtener funciones lineales de las variables independientes, denominadas funciones discriminantes, que permitan

clasificar a los individuos en uno de los grupos establecidos por los valores de la variable dependiente [3].

2.2.2.1 Modelo matemático

A partir de q grupos donde se asigna una serie de instancias u objetos y de p variables $(x_1, x_2, \dots, x_p)$ medidas sobre éstas, se trata de obtener para cada instancia una serie de puntuaciones (score) que indican el grupo al que pertenecen $(y_1, y_2, \dots, y_m)$ de modo que sean funciones lineales de $x_1, x_2, \dots, x_p$

$$y_1 = a_{11}x_1 + \cdots + a_{1p}x_p + a_{10}$$

$$\vdots$$

$$y_m = a_{m1}x_1 + \cdots + a_{mp}x_p + a_{m0}$$

donde $m = \min(q-1, p)$, tales que discriminen o separen lo máximo posible a los q grupos. Estas combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizar la varianza dentro de los grupos [4, 7].

2.2.2.2 Análisis Discriminante para dos grupos

Sean C_1 y C_2 dos clases donde se define una variable aleatoria vectorial x de dimensión p. Supondremos que x es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Se estudia el problema de clasificar un nuevo individuo en una de estas clases cuando se observa un vector x_0.

El problema puede enfocarse desde el punto de vista de la Inferencia o de la Teoría de decisión, incluyendo además probabilidades a priori (enfoque bayesiano) o no.

A continuación, se presenta la formulación del problema más general como un problema bayesiano de decisión. Se consideran las siguientes hipótesis:

- Las probabilidades a priori π_1 y π_2, de que una instancia tomada al azar provenga de la clase C_1 y C_2 respectivamente, son conocidas y $\pi_1 + \pi_2 = 1$
- Las consecuencias asociadas a los errores de clasificación son $c(2|1)$ y $c(1|2)$, donde $c(i|j)$ es el costo de clasificar en C_i de un objeto que pertenece realmente a C_j. Estos costos se suponen conocidos.
- Las preferencias del decisor por las consecuencias de sus acciones son lineales, es decir, maximizar la función de utilidad equivale a minimizar el costo esperado o costo de oportunidad de la decisión. Por lo tanto, podemos minimizar los costos de oportunidad de la decisión mediante el criterio del valor esperado.

Las posibles decisiones en el problema son únicamente dos: asignar en C_1 ó en C_2 . Una regla de decisión es equivalente a hacer una partición del espacio muestral E_x (que en general será R^p) en dos regiones: A_1 y $A_2 = E_x - A_1$, tales que:

1. Si $x_0 \in A_1 \Rightarrow d_1$(asignar en C_1).
2. Si $x_0 \in A_2 \Rightarrow d_2$(asignar en C_2).

Una vez observado el valor x_0 podemos calcular la probabilidad a posteriori de que el elemento pertenezca a cada población.

Definamos $P(1|x_0)$ como la probabilidad a posteriori de que un elemento que ha tomado un valor igual a x_0 pertenezca a C_1. Por el teorema de Bayes esta probabilidad es:

$$P(1|x_0) = \frac{P(x_0|1)P(1)}{P(x_0|1)P(1) + P(x_0|2)P(2)} \tag{3}$$

Las probabilidades $P(x_0|1)$ y $P(x_0|2)$ son proporcionales a $f_1(x)$ y $f_2(x)$, por lo que la ecuación anterior puede escribirse de la siguiente manera para las clases C_1 y C_2 :

$$P\ (1|x_0) = \frac{f_1(x_0)\pi_1}{f_1(x_0)\pi_1 + f_2(x_0)\pi_2} \tag{4}$$

$$P\ (2|x_0) = \frac{f_2(x_0)\pi_2}{f_1(x_0)\pi_1 + f_2(x_0)\pi_2} \tag{5}$$

Si clasificamos al elemento en la clase C_2 las posibles consecuencias son:

(i) Acertar con probabilidad $P(2|x_0)$, en cuyo caso no hay ningún costo de penalización.

(ii) Equivocarnos con probabilidad $P\ (1|x_0)$, en cuyo caso incurrimos en el costo asociado $c(2|1)$ y el costo promedio o valor esperado de la decisión "clasificar x_0 en C_2" será:

$$E(d_2) = c(2|1)P(1|x_0) + 0 \cdot P(2|x_0) = c(2|1)P(1|x_0) \tag{6}$$

Análogamente, el costo esperado de la decisión "clasificar en el grupo 1" será:

$$E(d_1) = 0 \cdot P(1|x_0) + c(1|2)P(2|x_0) = c(1|2)P\ (2|x_0) \tag{7}$$

Asignaremos el elemento al grupo 2 si su costo esperado es menor, es decir, sustituyendo en las expresiones anteriores, si:

$$\frac{f_2(x_0)\pi_2}{c(2|1)} > \frac{f_1(x_0)\pi_1}{c(1|2)} \tag{8}$$

Esto indica que, siendo iguales el resto de términos, clasificaremos en la clase C_2 si:

- La probabilidad a priori es más alta
- La verosimilitud de que provenga de C_2 es más alta.
- El costo de equivocarnos al clasificarlo en C_2 es más bajo.

Si suponemos que los costos son iguales, c(1|2) = c(2|1) = c, la decisión dependerá únicamente de las probabilidades y asignaremos el elemento al grupo más probable a posteriori. Por tanto, clasificaremos al elemento x_0 en C_2 si:

$$f_2(x_0)\pi_2 > f_1(x_0)\pi_1 \tag{9}$$

Por otro lado, si las probabilidades a priori y los costos son iguales la condición de clasificar en C_2 se reduce a:

$$f_2(x_0) > f_1(x_0) \tag{10}$$

y la regla de decisión expresa simplemente que clasificamos al individuo en la clase que pueda generarlo con mayor probabilidad.

2.2.2.3 Poblaciones Normales: Función Lineal discriminante

Sea $\bar{X}_1$ el vector de medias de los p atributos en la clase 1 y sea $\bar{X}_2$ el vector de medias correspondiente a los p atributos en la clase 2.

Consideremos μ_1 y μ_2 como los vectores de medias de las respectivas clases poblacionales C_1 y C_2 , suponiendo que ambas poblaciones tienen matriz de covarianzas iguales; es decir: $\Sigma_1 = \Sigma_2 = \Sigma$ y sean f_1 y f_2 distribuciones normales con distintos vectores de medias pero idéntica matriz de varianza. Entonces su función de densidad es:

$$f_i(x) = \frac{e^{-\frac{1}{2}(x-\mu_i)'\mathrm{V}^{-1}(x-\mu_i)}}{(2\pi)^{\frac{k}{2}}\,|\Sigma|^{\frac{1}{2}}} \tag{11}$$

La manera óptima es clasificar en la población de clase 2 (C_2) sí:

$$\frac{\pi_2 f_2(x)}{c(2|1)} > \frac{\pi_1 f_1(x)}{c(1|2)} \tag{12}$$

Como ambos términos son siempre positivos, al tomar logaritmos y sustituyendo $f(x)$ por su expresión equivalente, tenemos:

$$log(\,\pi_2) - \frac{1}{2}\,(x-\mu_2)'\Sigma^{-1}(x-\mu_2) > log(\pi_1) - \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) - log\left[\frac{c(1|2)}{c(2|1)}\right] \tag{13}$$

Realizando las operaciones

$$(x-\mu_1)'\Sigma^{-1}(x-\mu_1) > (x-\mu_2)'\Sigma^{-1}(x-\mu_2) - 2\,log\left[\frac{c(1|2)}{c(2|1)} * \frac{\pi_2}{\pi_1}\right] \tag{14}$$

para $i = 1,2$ Llamemos $D(x, C_i)$ al cuadrado de la distancia de Mahalanobis de x al centro de la clase C_i .

$$D(x, C_i) \ = (x - \mu_i)' \Sigma^{-1} (x - \mu_i) \tag{15}$$

Suponiendo iguales los costos y las probabilidades a priori, la expresión resultante es clasificar en la clase 2 (C_2) si:

$$D(x, C_1) \ > D(x, C_2) \tag{16}$$

La expresión (26) puede ser escrita como:

$$(\mu_2 - \mu_1)' \Sigma^{-1} \left[x - \frac{1}{2} (\mu_1 + \mu_2) \right] > 0 \tag{17}$$

Usando la muestra de entrenamiento, μ_i puede ser estimada por $\bar{X}_i$ y Σ estimada por S (matriz de covarianza combinada), la cual se calcula:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{18}$$

donde S_1 y S_2 representan las matrices de covarianzas muestrales del vector de variables predictoras en cada clase. La expresión muestral está dada por:

$$(\bar{X}_2 - \bar{X}_1)' \mathrm{S}^{-1} \left[x - \frac{1}{2} (\bar{X}_1 + \bar{X}_2) \right] > 0 \tag{19}$$

El lado izquierdo de la expresión (19), se le conoce como función de discriminante lineal [3].

2.2.3 ÁRBOLES DE DECISIÓN

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen diagramas de construcción lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Un árbol de decisión tiene entradas, las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos. A partir de esto devuelve una respuesta, la que en último término es una decisión que se toma a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser discretos o continuos.

Un árbol de decisión realiza una prueba durante el recorrido hacia las hojas, para obtener finalmente una decisión. El árbol de decisión suele contener [8]:

- Nodos internos: contiene una prueba sobre algún valor de una de las propiedades.
- Nodos de probabilidad: indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema.
- Nodos hojas : representa el valor que devolverá el árbol de decisión
- Ramas: brindan los posibles caminos que se tienen de acuerdo con la decisión tomada.

Los árboles de decisión son herramientas que se usan para realizar elecciones adecuadas entre muchas posibilidades. Su estructura permite seleccionar una y otra vez

diferentes opciones para explorar las diferentes alternativas posibles de decisión. Los árboles de decisión son guías jerárquicas multi-vía donde los valores de las características son el criterio diagnóstico para evaluar la calidad de la solución y determinar el uso más apropiado. La jerarquía se refiere a que la toma de una decisión o camino lleva a otra, hasta que todos los factores o características involucradas se hayan tomado en cuenta. Es multi-vía porque pueden existir más de dos opciones y es una guía porque al responder una pregunta se llega a una decisión [8].

El uso de árboles de decisión tuvo su origen en las ciencias sociales con los trabajos de Sonquist y Morgan el año 1964 y Morgan y Messenger el año 1979, ambos realizados en la Universidad de Michigan. El programa para la "Detección de Interacciones Automáticas", creada el año 1971 por los investigadores Sonquist, Baker y Morgan, fue uno de los primeros métodos de ajuste de los datos basados en árboles de clasificación. En estadística, en el año 1980, Kass introdujo un algoritmo recursivo de clasificación no binario, llamado "Detección de Interacciones Automáticas Chi-cuadrado". Hacia el año 1984, los investigadores Breiman, Friedman, Olshen y Stone, introdujeron un nuevo algoritmo para la construcción de árboles y los aplicaron a problemas de regresión y clasificación. El método es conocido como "Árboles de clasificación y regresión". Casi al mismo tiempo el proceso de inducción mediante árboles de decisión comenzó a ser usado por la comunidad de "Aprendizaje automático" [8].

C4.5 es uno de los algoritmos de árboles de decisión introducido por Quilan (1983) dentro de la comunidad de "Aprendizaje Automático" [9]. RPART (Particionamiento Recurrente) es una versión de CART disponible en el lenguaje de programación "R". Los

árboles de decisión son útiles para la exploración inicial de datos y apropiados cuando hay un número elevado de datos, y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo. [10].

2.2.3.1 Procedimiento para particionar los nodos

Suponemos que se dispone de una muestra de entrenamiento que incluye la información del grupo al que pertenece cada caso y que sirve para construir el criterio de clasificación. Se comienza con un nodo inicial y nos preguntamos cómo dividir el conjunto de datos disponibles en dos partes más homogéneas utilizando una de las variables. Esta variable se escoge de modo que la partición de datos se haga en dos conjuntos lo más homogéneos posibles [10]. Se elige, por ejemplo, la variable x_1 y se determina un punto de corte, llamémosle c, de modo que se puedan separar los datos en dos conjuntos: aquellos con $x_1 \leq c$ y los que tienen $x_1 > c$. De este nodo inicial saldrán ahora dos: uno al que llegan las observaciones con $x_1 \leq c$ y otro al que llegan las observaciones con $x_1 > c$. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes más homogéneas. El proceso termina cuando se hayan clasificado todas las observaciones correctamente en su grupo.

Para decidir qué variable va a utilizarse para hacer la partición en un nodo, se calcula primero la proporción de observaciones que pasan por el nodo para cada uno de los grupos. Si se denomina a los nodos como $t = 1, 2, \ldots, T$ en el caso de árboles de

clasificación, la proporción de observaciones en un nodo "t" se representa de la siguiente manera [9]:

$$p(t) = \frac{\{\#(i \leq N): X_i \in t\}}{N} \tag{20}$$

Por otro lado, la proporción de observaciones que lleguen al nodo "t" que pertenecen a la clase "g" ($g = 1, \dots, G$) donde G es el número de clases está dada por:

$$P(g/t) = \frac{\{\#(i \leq N): X_i \in s \ \ y \ Y_i = g\}}{\{\#(i \leq N): X_i \in s\}} \tag{21}$$

2.2.3.2 Medidas de Impureza

Para árboles de clasificación se pueden usar las siguientes medidas de impureza [9].

a) **El Coeficiente de Gini**. Para el nodo t se define por

$$i_{Gini}(t) = \sum_{g} \sum_{k} p(g/t)\, p(k/t) = \sum_{g}^{G} p(g/t)[1 - p(g/t)] \tag{22}$$

Si hay dos clases, entonces G=2 y el coeficiente de Gini sería:

$$i_{Gini}(t) = 2p(1-p) \tag{23}$$

donde p es la proporción de observaciones en la primera clase.

El coeficiente de Gini se puede interpretar como aquel donde se clasifica cada observación de un nodo a una clase g con probabilidad $p(g/t)$ en lugar de clasificar todas las observaciones a la clase con mayor número de observaciones.

b) La Entropía Cruzada o Devianza o Impureza de Información, definida por

$$I_E(t) = -\sum_{g=1}^{G} p(g/t) * log\,[p(g/t)] \tag{24}$$

la cual es máxima cuando:

$$p(g|t) = 1/G \tag{25}$$

La variable que se introduce en un nodo es la que minimiza la heterogeneidad o impureza que resulta de la división en el nodo. El logaritmo que se toma es por lo general en base 2, pero cuando se aplica entropía a distribuciones continuas se aplica logaritmo natural.

Para dos clases se utiliza la siguiente expresión:

$$I_E(t) = -2p * log(p) - 2(1-p) * log(1-p) \tag{26}$$

La clasificación de las observaciones en los nodos terminales se hace asignando todas las observaciones del nodo al grupo más probable en ese nodo, es decir, al grupo con máxima $p(g|t)$. Si la impureza del nodo es cero, todas las observaciones pertenecerían al mismo nodo, en caso contrario puede haber cierto error de clasificación. La impureza Gini se usa automáticamente en la librería *rpart* del lenguaje de programación "R", pero tiene la opción de utilizar la impureza de información.

c) La tasa de Mala clasificación, definida como

$$I_{MC}(t) = 1 - max_g p(g/t) \tag{27}$$

Cuando se tienen dos clases la expresión para calcular la tasa de error es:

$$I_{MC}(t) = 1 - max(p, 1-p) \tag{28}$$

Cuando el número de variables es grande, el árbol puede contener un número excesivo de nodos por lo que se hace necesario definir procedimientos de poda o simplificación del mismo. Es conveniente encontrar un tamaño óptimo del árbol de modo que se encuentre el equilibrio entre la complejidad del árbol y el ajuste a las observaciones [10]. En la librería *"rpart"*, se define el parámetro de complejidad "*cp"*. Se obtiene un árbol óptimo en función de este parámetro.

2.2.3.3 Estimacion del error de mala clasificación

Para encontrar la tasa de error de mala clasificación es recomendable usar validación cruzada 10 para estimar el error de clasificación [11]. No se recomienda usar "bootstrapping" ya que se ha mostrado que el sesgo se reduce en menos del 40% del verdadero sesgo [11]; aunque, para el caso de muestras pequeñas el "bootstrapping" podría ser usado.

La función *xpred.rpart* es muy útil puesto que nos muestra las predicciones de la variable de respuesta usando validación cruzada para valores dados del parámetro de complejidad.

2.2.4 REGLA DEL VECINO MÁS CERCANOS (k-NN)

El algoritmo de clasificación *k-nn* (K nearest neighbors) es un método de clasificación supervisada publicado por Fix y Hodges en 1951, es una de las primeras investigaciones que proporciona reglas basadas en métodos no paramétricos para la manipulación del conjunto de datos [13]. Se usa para estimar la función de densidad $f(x \,/\, C_j)$ de las predictoras x por cada clase C_j.

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de instancias. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras [14].

2.2.4.1 Estimación de la función de densidad multivariada

El estimado de la función de densidad para datos multivariados tiene la siguiente forma:

$$\hat{f}(x) = \frac{k}{n v_k(x)} \tag{29}$$

dónde $v_k(x)$ es el volumen de un elipsoide centrado en x de radio $r_k(x)$, que a su vez es la distancia de x al k-ésimo punto más cercano.

2.2.4.2 Clasificador k-nn

Desde el punto de vista de clasificación supervisada el método *k-nn* es bien simple de aplicar. En efecto, si las funciones de densidades condicionales f($\boldsymbol{x}$ / C_i) de la clase C_i que aparecen en la ecuación

$$P(C_i/\boldsymbol{x}) = \frac{\mathrm{f}(\boldsymbol{x}/\ C_i)\pi_i}{f(\boldsymbol{x})} \tag{30}$$

son estimadas por *k-nn*. Entonces, para clasificar un objeto, con mediciones dadas por el vector x, en la clase C_i se debe cumplir que:

$$\frac{k_i\pi_i}{n_i\ v_k(\boldsymbol{x})} > \frac{k_j\pi_j}{n_j\ v_k(\boldsymbol{x})} \tag{31}$$

para $j \neq i$, donde k_i y k_j son los k vecinos de x en las clase C_i y C_j respectivamente. Asumiendo a priori proporcionales a los tamaños de las clases (n_i/n y n_j/n respectivamente) lo anterior es equivalente a:

$$k_i > k_j \text{ para } j \neq i$$

Sea $\boldsymbol{X}$ el conjunto de entrenamiento de dimensión $n * p$, los ejemplos o patrones de entrenamiento son vectores en un espacio característico multidimensional, cada ejemplo está descrito en términos de p atributos considerando j clases para la clasificación. Los valores de los atributos del i-ésimo ejemplo se representan por el vector p-dimensional

$$\boldsymbol{x_i} = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right) \in X$$

donde $i = 1, \dots, n$. El proceso de este método se basa en el supuesto de considerar a las instancias más cercanas, como aquellas que tienen la mayor probabilidad de pertenecer

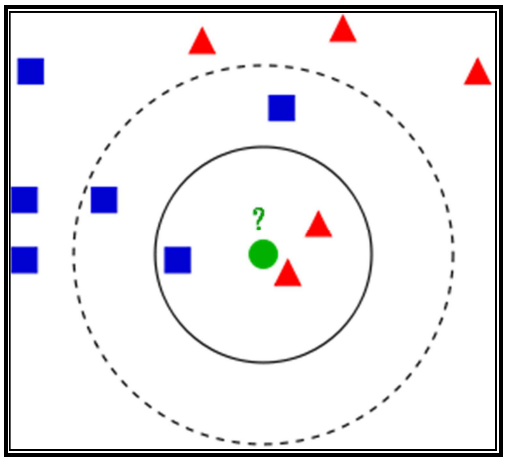
a la misma clase. Para esto se desea clasificar un nuevo caso $\boldsymbol{x_m}$, se realiza el cálculo de la distancia entre $\boldsymbol{x_m}$ y las instancias $\boldsymbol{x_i}$ en el conjunto de entrenamiento, y se asigna la etiqueta correspondiente al patrón que obtuvo la menor distancia (el más cercano) a $\boldsymbol{x_m}$ [13]. Es decir:

$$d(\boldsymbol{x_i}, \boldsymbol{x_m}) = min_i[d(\boldsymbol{x_i}, \boldsymbol{x_m})] \quad (32)$$

donde d es cualquiera medida de similaridad, x_m es un patrón no conocido por el clasificador y x_i es una instancia o patrón de entrenamiento. El espacio es particionado en regiones por localizaciones y etiquetas de los ejemplos de entrenamiento. Un punto en el espacio es asignado a la clase C_j si esta es la clase más frecuente entre los k ejemplos de entrenamiento más cercano.

Por ejemplo en la Figura 4 se observa que el ejemplo que se desea clasificar es el círculo verde. Para $k = 3$ este es clasificado con la clase *triángulo*, ya que hay solo un cuadrado y 2 triángulos, dentro del círculo que los contiene. Si $k = 5$, éste es clasificado con la clase *cuadrado*, ya que hay 2 triángulos y 3 cuadrados, dentro del círculo externo[15].

Figura 4. EJEMPLO DEL ALGORITMO KNN [15]



De acuerdo al ejemplo anterior, el procedimiento de clasificación por *k-nn* de manera más general se podría plantear de la siguiente manera:

1. Hallar los k objetos que están a una distancia más cercana al objeto $\boldsymbol{x}$, k usualmente es un número impar.
2. Si la mayoría de esos k objetos pertenecen a la clase C_i entonces el objeto $\boldsymbol{x}$ es asignado a ella.

En caso de que haya el mismo número de objetos en ambas clases, entonces se procede a clasificar aleatoriamente el nuevo objeto.

2.2.4.3 Medidas de distancia

Uno de los problemas en el método k-nn, es la elección de la distancia o métrica para hallar los k vecinos cercanos. Veamos algunas medidas de similaridad y distancias que podrían utilizarse.

Sean: $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})$ y $\boldsymbol{x_j} = (x_{1j}, x_{2j}, \dots, x_{pj})$ dos vectores en un espacio vectorial real n-dimensional con un sistema de coordenadas cartesianas fijo.

- **Distancia Euclideana :** Está definida por:

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2} \tag{33}$$

Esta es la métrica más elemental que se puede elegir. Sin embargo, puede causar problemas si las variables predictoras han sido medidas en unidades muy distintas entre sí. Algunos optan por re-escalar los datos antes de aplicar el método.

- **Distancia Manhattan:** Otra distancia que comúnmente se utiliza es la distancia Manhattan definida por :

$$d_1(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{r=1}^{p} |x_{ir} - x_{jr}| \tag{34}$$

- **Distancia Gower :** La distancia Gower está definida

$$d_G(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{r=1}^{p} \frac{|x_{ir} - x_{jr}|}{pc_r} \tag{35}$$

dónde $c_i = max(col_i) - min(col_i)$.

Esta distancia se puede usar cuando se tienen variables predictoras mixtas.

2.2.4.4 Elección de k

Otro de los problemas que se presenta en el método *k-nn* es la elección de *k.* La mejor elección de *k* depende fundamentalmente de los datos; generalmente, valores grandes de *k* reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas. Un buen *k* puede ser seleccionado mediante una optimización de uso [13].

Por otro lado, Enas y Choi (1996) usando simulación hicieron un estudio para determinar el k óptimo cuando solo hay dos clases presentes y determinaron que si los

tamaños muestrales de las dos clases son comparables entonces $k = n^{3/8}$ si había poca diferencia entre las matrices de covarianzas de los grupos y $k = n^{2/8}$ si había bastante diferencia entre las matrices de covarianzas[14].

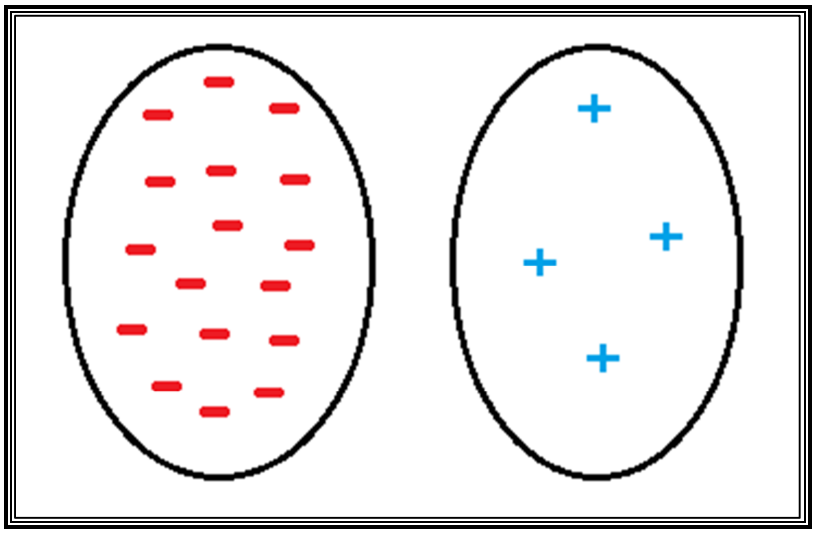
El sesgo del error de clasificación aumenta a medida que k aumenta, en tanto que la varianza disminuye [14].

3 MÉTODOS DE MUESTREO PARA SOLUCIONAR EL PROBLEMA DE DOS CLASES DESBALANCEADAS

3.1 PROBLEMA DE DOS CLASES DESBALANCEADAS

Un conjunto de datos cuyos objetos o ejemplos se clasifican en dos grupos se dice que se encuentra desbalanceado cuando la cantidad de ejemplos de una clase es mucho más grande que la de la otra (Figura 5).

Figura 5. BASE DE DATOS CON CLASES DESBALANCEADAS



En un caso ideal todos los datos pertenecientes a cada clase se encuentran agrupados entre ellos y claramente diferenciables del resto de las clases. La realidad es bien distinta y con frecuencia los datos presentan diferentes problemas que dificultan el desempeño de los clasificadores y disminuyen la calidad de la clasificación realizada [16]. Uno de los problemas más destacables en conjuntos de datos es el desbalance de clases que ocurre cuando el número de instancias entre las clases es muy diferente. Este

problema se presenta en aplicaciones tales como clasificación de imágenes de satélite [17], análisis de riesgos [18], clasificación de datos de tele-observación [19], detección de llamadas telefónicas fraudulentas [20], detección de derrames de petróleo a partir de imágenes de radar [21], recuperación y filtrado de información [22], Monitoreo de fallas en la caja de velocidades de helicópteros durante el vuelo [23], y especialmente en aplicaciones médicas [24].

El problema en base de datos no balanceados consiste en que los algoritmos de clasificación presentan una tendencia hacia la clase mayoritaria. Es decir, resultan muy bien clasificados los elementos de la clase mayoritaria, mientras que los de la clase minoritaria tiende a estar mal clasificados. Esto ocurre debido a que el clasificador intenta reducir el error global, de forma que el error de clasificación no tiene en cuenta la clasificación de los datos [25].

3.2 TÉCNICAS DE MUESTREO PARA BALANCEAR UN CONJUNTO DE DATOS.

Existen diferentes opciones para abordar el problema de dos clases desbalanceadas, entre las cuales se encuentran:

3.2.1 TÉCNICAS DE SOBRE-MUESTREO (OVERSAMPLING)

Consiste en balancear la distribución de las clases incrementando ejemplos de la clase minoritaria realizando re-muestreo de la clase pequeña (completamente al azar, con

reposición) hasta que haya aproximadamente la misma cantidad de ejemplos tanto de la pequeña como de la grande. Presenta inconvenientes cuando existen instancias con ruido en la clase minoritaria, debido a que puede repetir muestras con ruido generando sobre ajuste.

Este método de muestreo, fue previamente propuesto por Ling y Li [26] y Domingo [27] pero no fueron comparados sistemáticamente.

3.2.2 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) [25][29] es un algoritmo de sobre-muestreo (oversampling) que incrementa el tamaño del conjunto de ejemplos de la clase minoritaria generando instancias sintéticas o artificiales para equilibrar el conjunto de datos desbalanceados. En su implementación se hace uso de la regla del vecino más cercano y se establece el índice de incremento deseado.

Para cada una de las instancias minoritarias se buscan las instancias minoritarias vecinas (más cercanas) y se crean N instancias entre la línea que une la instancia original y cada una de las vecinas. El valor N depende del tamaño del sobre-muestreo deseado. Por ejemplo si las instancias de la clase mayoritaria resultan 5 veces mayor que el número de instancias de la clase minoritaria el porcentaje de incremento sería de aproximadamente 500%; entonces por cada instancia de la clase minoritaria se crea 5 nuevas instancias genéricas.

SMOTE es un algoritmo de sobre-muestreo de ejemplos utilizados para la clase minoritaria:

- Crea ejemplos sintéticos en lugar de hacer sobre-muestreo con reemplazo.
- Opera en el espacio de atributos, en lugar del espacio de datos.
- Crea un ejemplo sintético a lo largo de los segmentos de línea que unen algunos o todos los k vecinos más cercanos de manera aleatoria.
- SMOTE utiliza típicamente k=5

El algoritmo SMOTE realiza los siguientes pasos:

- Recibe como parámetros el porcentaje de ejemplos a muestrear.
- Calcula el número de ejemplos que tiene que generar.
- Calcula los k vecinos más cercanos de los ejemplos de la clase minoritaria.
- Genera los ejemplos siguiendo el proceso:
    - Para cada ejemplo de la clase minoritaria elige aleatoriamente el vecino a utilizar para crear el nuevo ejemplo.
    - Para cada atributo del ejemplo a sobre-muestrear, calcula la diferencia entre el vector de atributos muestra y el vecino elegido.
    - Multiplica esta diferencia por un número aleatorio entre 0 y 1.
    - Suma este último valor al valor original de la muestra.
    - Devuelve el conjunto de ejemplos sintéticos.

Algoritmo 1. SMOTE [24]

Entradas:
M: Conjunto de m patrones de la clase minoritaria $\{ M_i \ / \ i = 1, \dots, m\}$
Inicio:
Seleccionar de forma aleatoria un patron x de entrenamiento
Obtener los k vecinos de x en la misma clase.
Desde $s = 0, \dots, inc$
$v =$ un k vecino seleccionado aleatoriamente
Para todo i // Patrones de la clase minoritaria
Obtener R_n // Número aleatorio entre 0 y 1
// Creación del patrón sintético
$p = M_{ij} + R_n * (v_j - M_{ij})$ para todo j // Para cada atributo
$M = M \cup p$ // Incluye el patrón sintético
Fin para
Fin desde
Fin de algoritmo

SMOTE es conocido como el algoritmo que más se utiliza para realizar sobre-muestreo, pero este puede presentar inconvenientes tales como:

- Puede generar muchos ejemplos artificiales cuyas semillas son ejemplos con ruido.
- Al generar un nuevo ejemplo interpola entre dos ejemplos de la clase minoritaria, sin embargo pueden existir muchos ejemplos cercanos inclusive entre ellos de la clase mayoritaria generando modelos incorrectos [29]
- Solapamiento de clases, entre otros.

Con la finalidad de incrementar la calidad del conjunto de datos y minimizar inconvenientes que podrían resultar al aplicar la técnica de sobre-muestreo “SMOTE”,

se han propuestos varios algoritmos de limpieza, tales como la edición de Wilson “ENN” y “Tomek Links”.

3.2.3 ALGORITMO DE EDICIÓN DE WILSON (ENN)

El algoritmo de Wilson también conocido como la regla del vecino más cercano editado (Edited Nearest Neighbor), surge en 1972 como una de las primeras propuestas que considera la reducción del tamaño del conjunto de datos, mediante la eliminación de instancias mal clasificadas empleando el método de los vecinos cercanos K-NN. Por tanto elimina instancias tanto de la clase mayoritaria como de la clase minoritaria.

La mecánica que propone este algoritmo consiste en el análisis de cada patrón contenido en el conjunto de datos, se observan las características de sus k vecinos y, si la etiqueta de la mayoría de estos vecinos no coincide con la del patrón se concluye que éste es atípico y por consiguiente es marcado para ser eliminado. Una vez marcados los patrones que cumplen con la condición, el procedimiento continúa con la eliminación de todos aquellos ejemplos que hayan sido marcados para ser borrados dejando una nueva muestra integrada por el conjunto de datos original sin los ejemplos marcados como atípicos [28].

Esta técnica es ampliamente recomendada en aquellos casos donde se encuentren clases solapadas, ya que se ha demostrado que la mayor parte de los patrones o ejemplos que son eliminados (atípicos) corresponden a aquellos que se encuentran en la región de solape entre dos o más clases.

El algoritmo de ENN realiza los siguientes pasos [16]:

1. Recibe como parámetro el número de vecinos a buscar para cada clase.
2. Calcula los k vecinos más cercanos de cada ejemplo.
3. Si la mayoría de los vecinos son de clase distinta, marca el ejemplo o instancia.
4. Elimina los ejemplos marcados.

A continuación se presenta el algoritmo ENN de Wilson [24].

Algoritmo 2. "ENN" WILSON

Entradas:
M: Conjunto de m patrones de la clase minoritaria $\{ x_i \ / \ i = 1, \dots, m\}$
k: Número de vecinos a considerar
Salidas:
CEE: Conjunto de datos editado.
Método
Para todo $i = 1 \ hasta \ i = m$ **hacer**
Buscar los k vecinos de x_i en $M - \{x_i\}$
Determinar la clase mayoritaria de los k vecinos
Si clase mayoritaria $\neq$ etiqueta x_i **entonces**
Marcar x_i como atípico
Fin si
Fin para todo
Fin método

3.2.4 ALGORITMO TOMEK LINK

Este algoritmo elimina las instancias de la clase mayoritaria y minoritaria que estén muy próximas entre sí.

Dadas dos instancias $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$ las cuales pertenecen a dos clases diferentes y sea $d(x_i, x_j)$ la distancia entre $\boldsymbol{x_i}$ y $\boldsymbol{x_j}$. Se dice que $(\boldsymbol{x_i}, \boldsymbol{x_j})$ es un Tomek Links si no existe una instancia x_k tal que $d(x_k, x_j) < d(\boldsymbol{x_i}, \boldsymbol{x_j})$ ó $d(x_k, x_i) < d(\boldsymbol{x_i}, \boldsymbol{x_j})$. Si dos instancias (x_i, x_j) forman un Tomek Links, entonces una de ellas es ruido o ambas están en el borde. Tomek Links puede ser usado como método de sub-muestreo o como método de limpieza. Como método de sub-muestreo, slo la instancia que pertenece a la clase mayoritaria es eliminada y como método de limpieza, las instancias de ambas clases son removidas o eliminadas [30].

Algoritmo 3. TOMEK LINKS

Entradas:
X: Conjunto de m patrones de la clase A { $x_i \ / i = 1, \dots, m$}
Y: Conjunto de n patrones de la clase B { $y_j \ / j = 1, \dots, n$}

Salidas:
CEE: Conjunto de datos editado.
Método
Para todo $i = 1$ *hasta* $i = m$
Para todo $j = 1$ *hasta* $i = n$ **hacer**
Hallar $d(x_i, y_j)$
Identificar los x_k tal que $d(x_k, x_j) < d(x_i, x_j)$ ó $d(x_k, x_i) < d(x_i, x_j)$
Si $d(x_k, x_j) < d(x_i, x_j)$ ó $d(x_k, x_i) < d(x_i, x_j)$ **entonces**
(x_k, x_j) *o* (x_k, x_i) son Tomek Links.
Y se prosigue a eliminar los x_k, x_j de su conjunto original
Fin si
Fin para todo
Fin para todo
Fin método

4 METODOLOGÍA

En el presente trabajo se realizó una comparación entre los métodos a nivel de muestreo que permiten solucionar el problema de clases desbalanceadas tales como: sobre-muestreo con reposición al azar, sobre-muestreo SMOTE, y combinación de sobre-muestreo "SMOTE" con los métodos de limpieza ENN y Tomek Links, con el fin de analizar los resultados que se obtienen al aplicar los clasificadores de regresión logística, árboles de decisión, discriminante lineal y vecinos más cercanos. Para establecer cuál o cuáles de los clasificadores nos brinda un mejor rendimiento se aplicaron las técnicas de evaluación obtenidas a través de la matriz de confusión de cada algoritmo de clasificación.

En este capítulo se describe la forma en que se realizó el estudio, se explicarán las técnicas de evaluación de algoritmos de clasificación utilizadas y se hará una breve descripción de los datos utilizados.

4.1 CRITERIOS PARA LA EVALUACIÓN DE ALGORITMOS DE CLASIFICACIÓN

Existen diversos criterios para evaluar el rendimiento de un clasificador. Estimar la bondad de un clasificador, se conoce como proceso de validación, el cual nos permite efectuar una medición sobre la capacidad de predicción del modelo generado a partir de un clasificador.

Existen diversas alternativas para verificar o medir la bondad del clasificador través de la matriz de confusión [31].

4.1.1 MATRIZ DE CONFUSIÓN PARA DOS CLASES

La matriz de confusión o matriz de contingencia es una herramienta de visualización que se emplea en aprendizaje supervisado para la presentación y el análisis de resultados de una clasificación debido a su capacidad de plasmar los conflictos entre las clases. Así no solo se conoce el porcentaje correcto de clasificación, sino también la fiabilidad para cada una de las clases y las principales confusiones entre ellas [13]

Se puede considerar una matriz de confusión para dos clases como una matriz con dos filas y dos columnas auxiliares para contabilizar diversos parámetros estadísticos, tales como: totales, la asignación real de instancias por clases, etc. Las columnas de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real (Figura 6).

Figura 6. MATRIZ DE CONFUSIÓN PARA DOS CLASES

Clase real	Clase pronosticada	
	P	N
P	Verdaderos Positivos "VP"	Falsos Negativos "FN"
N	Falsos positivos "FP"	Verdaderos Negativos "VN"

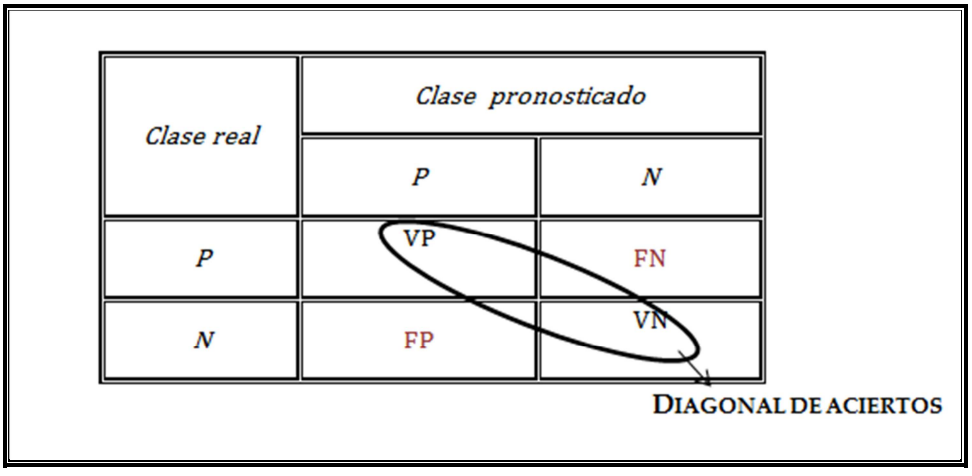
- **VP** (**Verdaderos positivos**): Representa el número de instancias clasificadas correctamente como positivas.
- **FN (Falsos negativos):** Representa el número de instancias que son de clase real positiva pero fueron clasificadas como negativas.
- **FP (Falsos positivos):** Representa el número de instancias que son de clase real negativa pero fueron clasificadas como positivas.
- **VN (Verdaderos negativos**): Representa el número de instancias clasificadas correctamente como negativas.

Sea N es el número total de instancias del conjunto de datos de entrenamiento, entonces:

$$N = VP + FN + FP + VN.$$

El número de instancias clasificadas correctamente es la suma de la diagonal de la matriz (Figura 7) y el resto están clasificadas de forma incorrecta.

Figura 7. MATRIZ DE CONFUSIÓN –DIAGONAL DE ACIERTOS



4.1.2 EVALUACIÓN CLÁSICA

- **Tasa de Error:** Es la proporción de instancias que han sido clasificadas incorrectamente, es decir:

$$Tasa\ de\ error = \frac{FP + FN}{N} \tag{36}$$

- **Tasa de Precisión o Acierto:** Es la proporción de instancias que han sido clasificadas correctamente, es decir:

$$Tasa\ de\ precisión\ o\ acierto = \frac{VP + VN}{N} \tag{37}$$

4.1.3 COBERTURA Y PRECISIÓN

- **Cobertura (recall):** La cobertura mide la proporción de instancias correctamente clasificadas como positivas respecto al total de instancias de clase real positiva.

$$Cobertura = \frac{VP}{VP + FN} \tag{38}$$

- **Precisión:** Mide la proporción de instancias clasificadas correctamente como positivas, respecto al total de instancias predichas positivas.

$$Precisión = \frac{VP}{VP + FP} \tag{39}$$

La cobertura y la precisión mantienen una relación inversa, es decir cuando aumenta la cobertura del modelo generado disminuye la precisión y viceversa cuando disminuye la cobertura aumenta la precisión. De manera tal que se obtiene una cobertura total, a costo de una precisión nula.

4.1.4 MEDIDA F

Es una medida que se utiliza para caracterizar con un único valor la bondad de un clasificador o algoritmo y corresponde a la media armónica ponderada de precisión y cobertura.

$$F = \frac{(\beta^2 + 1) * Precisión * Cobertura}{(\beta^2)Precisión + Cobertura} \tag{40}$$

donde β es la importancia relativa atribuida a la cobertura sobre la precisión.

Ahora bien, para que la precisión y cobertura tengan el mismo peso es recomendable usar $\beta = 1$. Por lo que la ecuación 39 se expresa:

$$F = \frac{2 * Precisión * Cobertura}{Precisión + Cobertura} \tag{41}$$

La medida F con porcentaje mayor (80% o más) se considera indicador de buen rendimiento del método de clasificación [33].

4.1.5 SENSITIVIDAD Y ESPECIFICIDAD

Especificidad y sensibilidad son dos medidas tomadas de la Teoría de la Detección de la Señal. Dicha teoría fue desarrollada durante la Segunda Guerra Mundial a causa de los fallos de detección de los equipos de radar en el ataque a Pearl Harbor. A partir de los 70 se siguió desarrollando para aplicaciones de diagnóstico médico y en los 90 se empezó a popularizar en Aprendizaje Automático. Fueron originadas en la Teoría de la Detección de la Señal y usados en diagnóstico médicos en 1970 y en Aprendizaje Automático desde 1990.

- **Sensitividad o tasa de verdaderos positivos (Cobertura)**: Mide la proporción de instancias que fueron clasificados como clase positiva, de entre todos los ejemplos que de verdad tienen clase positiva y su formulación es equivalente a la cobertura.

$$Sensitividad \text{ ó } TVP = \frac{VP}{VP + FN} \tag{42}$$

- **Especificidad o tasa de verdaderos negativos:** Mide la proporción de ejemplos que fueron clasificados como clase negativa de entre todas las instancias que de verdad pertenecen a clase negativa.

$$Especificidad \text{ ó } TVN = \frac{VN}{VN + FP} = 1 - \frac{FP}{FP + VN} \tag{43}$$

Entre mayores sean los resultados de estas medidas mejor será el desempeño del método de clasificación.

4.1.6 EVALUACIÓN BASADA EN LA MEDIA GEOMÉTRICA:

La media geométrica entre la precisión de las instancias de la clase positiva (sensitividad) y la precisión de las instancias de la clase negativa (especificidad) es un indicador o medida que permite analizar desempeño del clasificador en este dominio. Esta medida se relaciona con un punto de la curva ROC y la idea es maximizar la precisión en cada una de las dos clases, manteniendo estas precisiones equilibradas.

$$G = \sqrt{Sensitividad * Especificidad} \tag{44}$$

Cuando los resultados de sensitividad y especificidad son altos (mayor 80%), la medida G indica que el rendimiento del clasificador es bueno. Ahora bien cuando los resultados para sensitividad son altos y para especificidad bajos, resultaría un valor de G muy bajo que indica un desempeño pobre del clasificador [32]

4.1.7 SILENCIO Y RUIDO

- **Ruido:** Mide la proporción de instancias de clase real negativa que fueron clasificadas incorrectamente, con respecto al total las instancias de clase predichas positivas. Su formulación y su relación con la precisión es:

$$Ruido = \frac{FP}{FP + VP} = 1 - Precisión \tag{45}$$

- **Silencio:** Mide la proporción de instancias de clase real positiva que fueron clasificadas incorrectamente, con respecto al total las instancias de clase real positivas y está relacionado de manera complementaria con la cobertura:

$$Silencio = \frac{FN}{FN + VP} = 1 - Cobertura \tag{46}$$

El silencio y el ruido están relacionados de manera inversa, disminuir el silencio provoca ruido y disminuir el ruido aumenta el silencio. Entre más pequeño sea el porcentaje de estas medidas el desempeño del clasificador evaluado será mejor [31].

4.2 DATOS

Cada una de las medidas anteriores se calculó a través de las matrices de confusión de los métodos de clasificación aplicados a los datos en los siguientes conjuntos CRX, GERMAN, EST1 y EST2. Nótese que como muestra de entrenamiento se utilizó en cada caso el conjunto de datos respectivo. Los dos primeros conjuntos de datos se descargaron de la página del Doctor Edgar Acuña, y se encuentran en la UCI (http://archive.ics.uci.edu/ml/datasets.html). A continuación se hace una breve descripción de ellos.

4.2.1 CRX

Es un conjunto de datos con dos clases desbalanceadas de dimensión 393×16, que contiene 393 instancias con 15 columnas que corresponden a las variables predictoras y una columna que corresponde a la variable de clasificación. Las clases identificadas en este conjunto de datos son: "0" y "1". La clase minoritaria es la clase "0" con una dimensión de 36×16 y la clase mayoritaria es la clase "1" con una dimensión de 357×16. Note que la cantidad de instancias de la clase minoritaria corresponde aproximadamente al 9% de la cantidad de instancias del conjunto de datos, mientras que la clase mayoritaria corresponde al 91%.

4.2.2 GERMAN

Es un conjunto de datos con clases desbalanceadas de dimensión es 1,000×21, contiene 1,000 instancias con 20 columnas que corresponden a las variables predictoras y una columna que corresponde a la variable de clasificación. Las clases identificadas en este conjunto de datos son: "0" y "1". La clase minoritaria es la clase "0" con una dimensión de 300×21 y la clase mayoritaria es la clase "1" con una dimensión de 700×21. Note que la cantidad de instancias de la clase minoritaria corresponde al 30% de los datos mientras que la clase mayoritaria corresponde al 70% (ver http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/).

4.2.3 EST1

Este conjunto de datos corresponde al rendimiento académico de 6,924 estudiantes del Recinto Universitario de la Universidad de Puerto Rico. EST1 corresponde a un conjunto de dos clases desbalanceadas cuya dimensión es 6,924×7, con 6 columnas que corresponde a las variables predictoras y una variable de clasificación. Las clases identificadas en este conjunto de datos son: "0" y "1". La clase minoritaria es la clase "0" con una dimensión de 562×7 y la clase mayoritaria es la clase "1" con una dimensión de 6362×7. Note que la cantidad de instancias de la clase minoritaria corresponde aproximadamente al 8% de los datos mientras que la clase mayoritaria corresponde al 92%.

4.2.4 EST2

Este conjunto de datos corresponde al rendimiento académico de 6,924 estudiantes del Recinto Universitario de la Universidad de Puerto Rico. EST2 corresponde a un conjunto de dos clases desbalanceadas cuya dimensión es 6,924×7, con 6 columnas que corresponde a las variables predictoras y una variable de clasificación. Las clases identificadas en este conjunto de datos son: “0” y “1”. La clase minoritaria es la clase “0” con una dimensión de 2,718×7 y la clase mayoritaria es la clase “1” con una dimensión de 4,206×7. Note que la cantidad de instancias de la clase minoritaria corresponde aproximadamente al 39% de los datos mientras que la clase mayoritaria corresponde al 61%.

5 RESULTADOS

5.1 RESULTADOS PARA LOS DATOS "CRX"

5.1.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO CRX

Para llevar a cabo un análisis del rendimiento de los clasificadores que se aplicaron al conjunto con clases desbalanceadas "CRX", iniciaremos por interpretar las matrices de confusión para cada clasificador (Tabla 1).

Tabla 1. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A CRX

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	31	**5**	36
1	**43**	314	357
Total	74	319	393

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	19	**17**	36
1	**6**	351	357
Total	25	368	393

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	8	**28**	36
1	**5**	352	357
Total	13	370	393

ÁRBOLES DE DECISIÓN

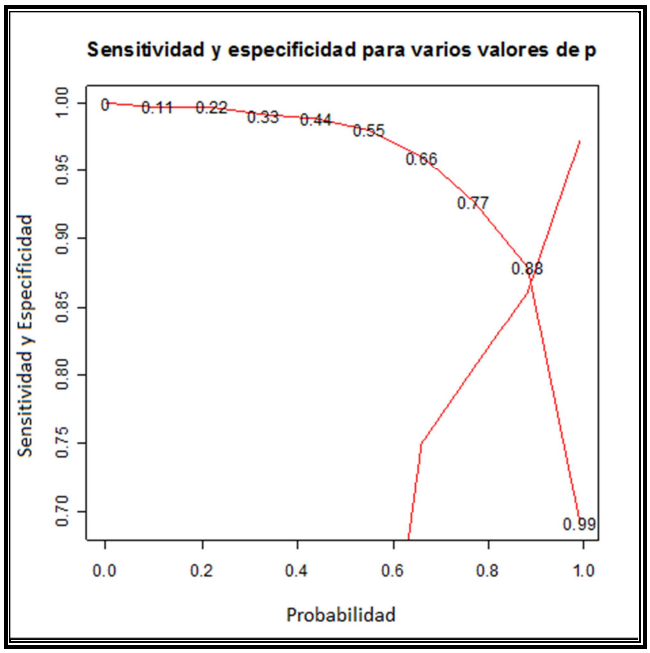
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	14	**22**	36
1	**7**	350	357
Total	21	372	393

En la Tabla 1 se muestran las matrices de confusión para los clasificadores de regresión logística, árboles de decisión, vecinos más cercanos y discriminante lineal aplicados a

los datos “CRX”. De acuerdo con estas matrices de confusión se puede visualizar como cada uno de los clasificadores, excepto regresión logística tienen una tendencia de clasificación a clase 1, es decir a la clase que tiene mayor número de instancias. Por ejemplo, el método discriminante lineal clasificó 368 instancias en la “clase 1”, de las cuales 351 se encuentran bien clasificadas y 17 mal clasificadas; y 25 instancias fueron predichas como “clase 0”, de las cuales 19 están bien clasificadas y 6 mal clasificadas.

Para hallar la matriz de confusión en el caso de regresión logística se usó el valor de $p = 0.88$ sugerido por la gráfica de sensibilidad y especificidad para varios valores de p (Gráfico 1). Los resultados de la matriz de confusión sobre los datos “CRX” que produjo este método no muestran tendencia a la clase mayoritaria.

Gráfico 1. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p
CRX



Al evaluar cada uno de los clasificadores aplicados al conjunto de datos CRX se obtuvieron los siguientes resultados (Tabla 2).

Tabla 2. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A CRX

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.12	0.87	0.58	0.14	0.56
Discriminante lineal	0.06	0.72	0.24	0.47	0.62
Vecinos más cercanos	0.08	0.47	0.38	0.78	0.33
Árboles de decision	0.07	0.62	0.33	0.61	0.49

A pesar de que la tasa de error de cada clasificador en el conjunto datos de "CRX" representa un porcentaje de error adecuado (porque es pequeño), no significa necesariamente que el rendimiento de los clasificadores sea el mejor. Note que los clasificadores tienen una tendencia a la clase mayoritaria excepto en el caso de la regresión logística.

Al analizar cada uno de los criterios tales como: "Medida F" y "el Ruido", los clasificadores muestran resultados que no son muy recomendables a la hora de escoger un clasificador con buen rendimiento, pues un buen clasificador debe arrojar 80% o más en las medidas F y G (basado en la media geométrica) [32][33]. Además el porcentaje de silencio tampoco revela buenos resultados en este conjunto de datos excepto en regresión logística. Mantener un porcentaje de silencio y de ruido en los datos superiores a 20% no es deseable, porque entre menos porcentaje de silencio se tenga, los datos proveerán más información acerca del comportamiento de las instancias[31].

5.1.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN

Al aplicar la técnica "sobre-muestreo aleatorio con reposición" se obtuvo un nuevo conjunto de datos con dimensión de 714×16, con 357 instancias de clase "0" y 357 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 3).

Tabla 3. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	324	**33**	357
1	**41**	316	357
Total	365	349	714

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	324	**33**	357
1	**64**	293	357
Total	388	326	714

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	350	**7**	357
1	**128**	229	357
Total	478	336	714

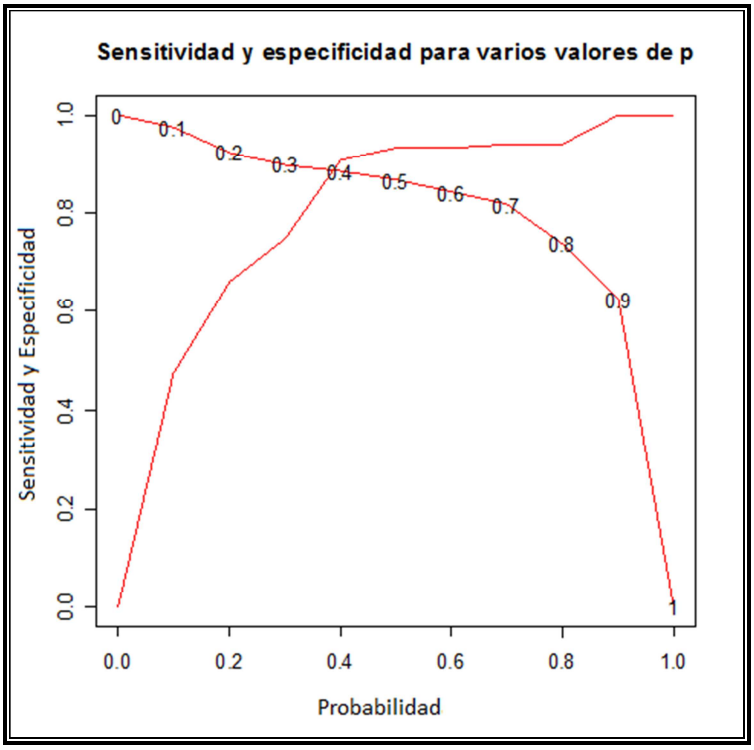
ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	353	**4**	357
1	**45**	312	357
Total	398	316	714

El método de "regresión logística" clasificó 349 instancias en la "clase 1", de las cuales 316 se encuentran bien clasificadas y 33 mal clasificadas (falso1); y 365 en "clase 0", de las cuales 324 se clasificaron correctamente y 41 instancias incorrectamente (falso 0).

En el Gráfico 2, se observa un valor de $p = 0.4$ cuando se intersectan las curvas de sensitividad y especificidad. Este valor se utilizó para calcular la matriz de confusión a través del lenguaje de programación "R".

Gráfico 2. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE P "CRX" -SOBRE-MUESTREO ALEATORIO



La Tabla 4 muestra la tasa de error, medida G, medida de ruido, medida de silencio y medida F, que son los criterios de evaluación que se calcularon a través de las matrices de confusión de los clasificadores. De los resultados de éstas métricas de evaluación, se puede concluir que el método de clasificación con mejor desempeño en el conjunto de datos "CRX" con clases balanceadas a través de la técnica "sobre-muestreo aleatorio con

reposición" es "árboles de decisión", con una tasa de error de 7% y medidas: G, de silencio, ruido y F de 93%, 11%, 1% y 94% respectivamente.

Tabla 4. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.10	0.90	0.11	0.09	0.90
Discriminante Lineal	0.14	0.86	0.16	0.09	0.87
Vecinos más cercanos	0.19	0.79	0.27	0.02	0.84
Árboles de Decisión	0.07	0.93	0.11	0.01	0.94

5.1.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE"

Al aplicar la técnica de sobre-muestreo "SMOTE" se obtuvo un nuevo conjunto de datos con dimensión de 717×16, con 360 instancias de clase "0" y 357 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 5).

Tabla 5. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	320	**40**	360
1	**37**	320	357
Total	357	360	717

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	337	**23**	360
1	**66**	291	357
Total	403	314	717

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	317	**43**	360
1	**93**	264	357
Total	410	307	717

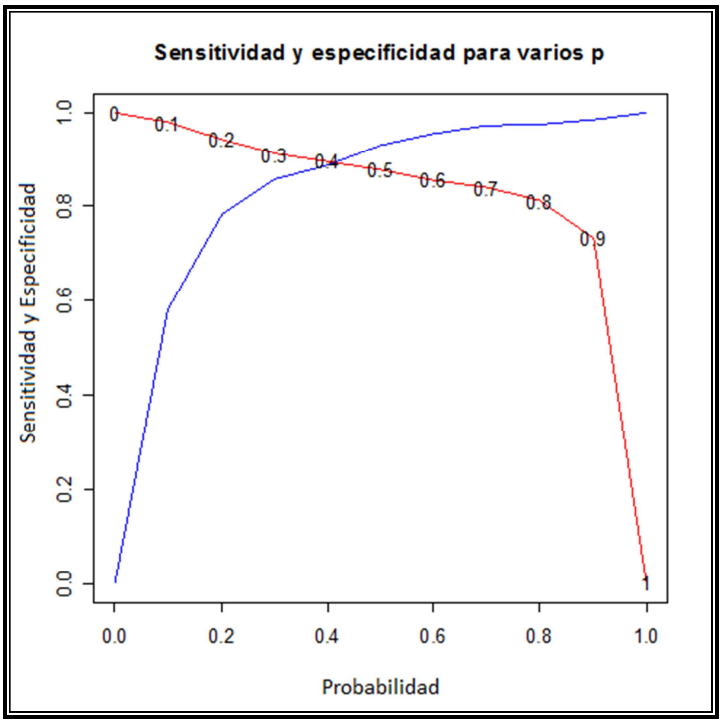
ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	341	**19**	360
1	**31**	326	357
Total	372	345	717

El gráfico de sensitividad y especificidad para diferentes valores de p de los datos con clases balanceadas a través de la técnica "SMOTE", muestra un valor óptimo $p = 0.4$, el

cual se utilizó para calcular la matriz de confusión del método de clasificación por regresión logística (Gráfico 3).

Gráfico 3. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE P "CRX-SMOTE"



Los clasificadores presentan de manera general buen desempeño. Sin embargo cabe anotar que "árboles de decisión" es el método de clasificación con mejor desempeño en los datos con clases balanceadas a través de SMOTE. Este clasificador muestra una tasa de error de 7%, y medidas: G, de silencio, ruido y F de 93%, 8%, 0.05% y 93% respectivamente (Tabla 6)

.

Tabla 6. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.11	0.89	0.10	0.11	0.89
Discriminante lineal	0.12	0.87	0.16	0.06	0.88
Vecinos más cercanos	0.19	0.81	0.23	0.12	0.82
Árboles de decision	**0.07**	**0.93**	**0.08**	**0.05**	**0.93**

5.1.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN)

Al aplicar la combinación de las técnicas "SMOTE y ENN" se obtuvo un nuevo conjunto de datos con dimensión de 600×16, con 318 instancias de clase "0" y 282 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 7).

Tabla 7. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	292	**26**	318
1	**25**	257	282
Total	317	283	600

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0		
0	300	**18**	318
1	**55**	227	282
Total	355	245	600

VECINOS CERCANOS

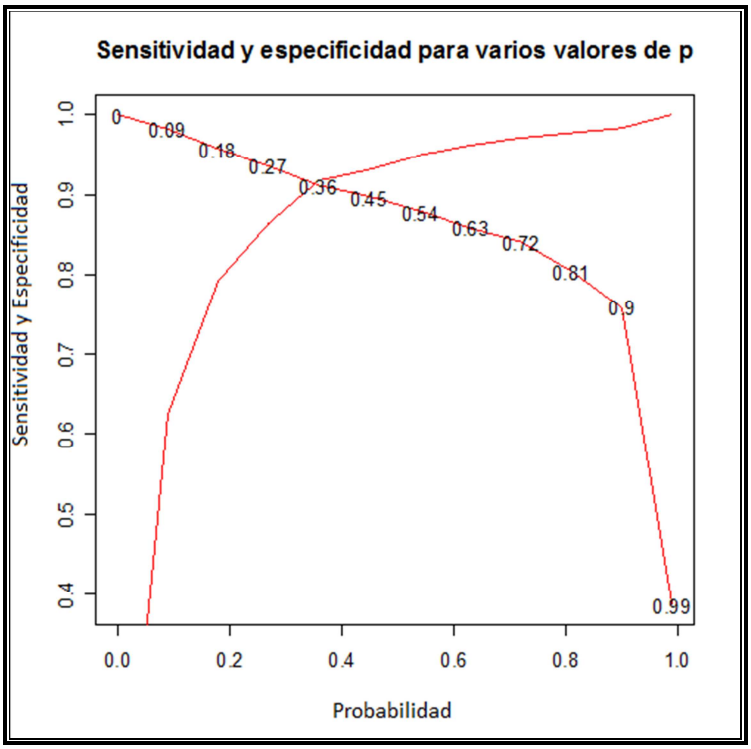
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	280	**38**	318
1	**59**	223	282
Total	339	261	600

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	300	**18**	318
1	**23**	259	282
Total	323	277	600

El valor óptimo de p para realizar la clasificación por el método de regresión logística es $p = 0.36$ (Gráfico 4).

Gráfico 4. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "CRX- SMOTE+ENN"



En la Tabla 8, se observan buenos resultados de las métricas de evaluación para todos los métodos de clasificación trabajados en esta tesis. Sin embargo se destacan entre éstos "Árboles de decisión y regresión logística". El primero de éstos tiene una tasa de error de 7%, una tasa de precisión de 93% y medidas: G, de silencio, ruido y F 93%, 7%, 6% y 94% respectivamente. El segundo presenta una tasa de error de 9%, medidas: G, de silencio, ruido y F 91%, 8%, 8% y 92% respectivamente.

Tabla 8. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.09	0.91	0.08	0.08	0.92
Discriminante lineal	0.12	0.87	0.15	0.06	0.89
Vecinos más cercanos	0.16	0.83	0.17	0.12	0.85
Árboles de decision	**0.07**	**0.93**	**0.07**	**0.06**	**0.94**

5.1.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (CRX) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "Tomek Links"

Al aplicar la combinación de las técnicas "SMOTE y Tomek Links" se obtuvo un nuevo conjunto de datos de dimensión 631×16, el cual contiene 314 instancias de clase "0" y 317 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 9).

Tabla 9. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	285	**29**	314
1	**28**	289	317
Total	313	318	631

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	298	**57**	314
1	**16**	260	317
Total	314	317	631

VECINOS CERCANOS

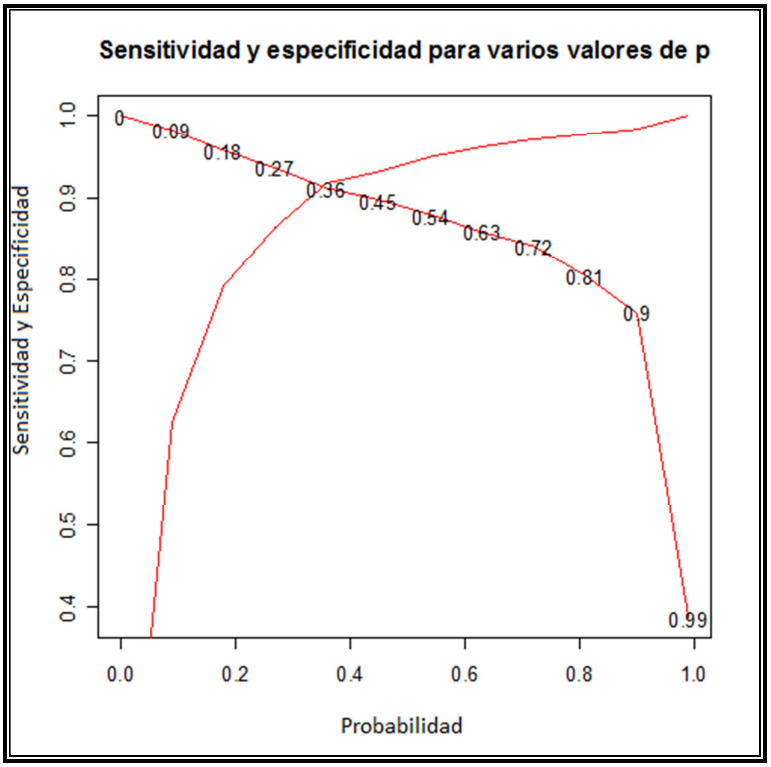
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	248	**66**	314
1	**81**	236	317
Total	329	302	631

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	286	**28**	314
1	**35**	282	317
Total	321	310	631

La matriz de confusión del método de clasificación regresión logística aplicado a los datos con clases balanceadas a través de la combinación "SMOTE y Tomek Links" se calculó usando el valor $p = 0.36$ (Gráfico 5).

Gráfico 5. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "CRX- SMOTE+TOMEK LINKS"



En la Tabla 10, se observan buenos resultados de las métricas de evaluación para todos los métodos de clasificación con los que trabajó en esta tesis excepto el método k-nn. "Regresión logística y "arboles de decisión" fueron los clasificadores que presentan mejor con mejor desempeño. El primero presenta una tasa de error de 9%, tasa de precisión de 91%, y medidas: G, de silencio, ruido y F de 91%, 9%, 9% y 91% respectivamente; y el segundo clasificador ilustra una tasa de error de 10%, medidas: G, de silencio, ruido y F de 90%, 11%, 9%, y 90% respectivamente.

Tabla 10. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS CRX BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.09	0.91	0.09	0.09	0.91
Discriminante lineal	0.12	0.89	0.05	0.16	0.89
Vecinos más cercanos	0.23	0.77	0.25	0.21	0.77
Árboles de decision	**0.10**	**0.90**	**0.11**	**0.09**	**0.90**

Ahora se analizarán detalladamente los resultados de las técnicas de muestreo que se aplicaron para solucionar el problema de distribución de clases desbalanceadas de los datos CRX, teniendo en cuenta los clasificadores que nos brindaron mejor desempeño (regresión logística y árboles de decisión).

Tabla 11. RESULTADOS DE LAS MÉTRICAS DE EVALUACIÓN GENERALES APLICADAS REGRESIÓN LOGÍSTICA SOBRE LOS DATOS "CRX" BALANCEADOS

MÉTRICAS	SIN APLICAR TECNICAS	SOBREMUESTREO ALEATORIO	SMOTE	SMOTE+ENN	SMOTE+ TOMEK LINKS
Tasa de error	0.12	0.10	0.11	**0.09**	0.09
Medida G	0.87	0.90	0.89	**0.91**	0.91
Ruido	0.58	0.11	0.10	**0.08**	0.09
Silencio	0.14	0.09	0.11	**0.08**	0.09
Medida F	0.56	0.90	0.89	**0.92**	0.91

La técnica que además de resolver el problema de desbalance de dos clases nos proporciona mejor desempeño en el clasificador de regresión logística es la combinación entre SMOTE y el método de limpieza de Wilson (ENN), dado que la tasa

de error y las medidas G, ruido, silencio y F fueron de 9%, 91%, 8%, 8% y 92% respectivamente (Tabla 11).

Tabla 12. RESULTADOS DE LAS MÉTRICAS DE EVALUACIÓN GENERALES APLICADAS A ÁRBOLES DE DECISIÓN SOBRE LOS DATOS "CRX" BALANCEADOS

MÉTRICAS	SIN APLICAR TECNICAS	SOBRE-MUESTREO ALEATORIO	SMOTE	SMOTE+ENN	SMOTE+ TOMEK LINKS
Tasa de error	0.07	0.07	0.07	**0.07**	0.10
Medida G	0.62	0.90	0.93	**0.93**	0.90
Ruido	0.33	0.11	0.08	**0.07**	0.11
Silencio	0.61	**0.01**	0.05	0.06	0.09
Medida F	0.49	0.94	0.93	**0.94**	0.90

La técnica que además de resolver el problema de desbalance de dos clases, nos proporciona mejor desempeño en el clasificador "árboles de decisión", es la combinación entre SMOTE y el método de limpieza de Wilson (ENN), la tasa de error y las medidas G, ruido, silencio y F fueron de 7%, 93%, 7%, 6% y 94% respectivamente (Tabla 12).

5.2 RESULTADOS PARA LOS DATOS "GERMAN"

5.2.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO GERMAN

Para llevar a cabo un análisis del rendimiento de los clasificadores que se aplicaron al conjunto con clases desbalanceadas "GERMAN", iniciaremos por interpretar las matrices de confusión para cada clasificado (Tabla 13).

Tabla 13. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A GERMAN

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	209	**91**	300
1	**219**	481	700
Total	428	572	1000

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	**1**	
0	98	**202**	300
1	**57**	643	700
Total	155	845	1000

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	39	**261**	300
1	**26**	674	700
Total	65	935	1000

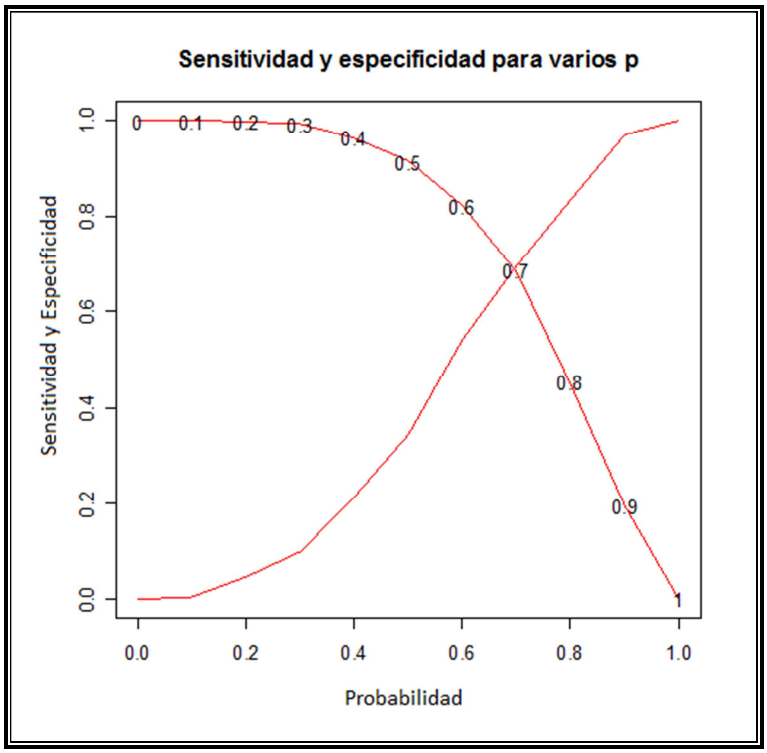
ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	71	**229**	300
1	**64**	636	700
Total	135	865	1000

Para hallar la matriz de confusión en el caso de regresión logística se usó el valor de $p = 0.7$ sugerido por el gráfico de sensibilidad y especificidad para varios valores de p (Gráfico 6). Los resultados de la matriz de confusión sobre los datos "GERMAN" que

produjo este método de igual manera que con los datos “CRX” no muestran tendencia a la clase mayoritaria, en este caso a la clase 1.

Gráfico 6. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE P GERMAN



Al evaluar cada uno de los clasificadores aplicados al conjunto de datos GERMAN se obtuvieron los resultados que se presentan en la Tabla 14.

Tabla 14. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A GERMAN

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.31	0.69	0.51	0.30	0.57
Discriminante lineal	0.26	0.55	0.37	0.67	0.43
Vecinos más cercanos	0.29	0.35	0.40	0.87	0.21
Árboles de decision	0.29	0.46	0.47	0.76	0.33

Al analizar la tasa de error, medida G, el ruido, el silencio y medida F, los clasificadores muestran resultados que no son muy recomendables a la hora de escoger un clasificador con buen rendimiento. Además el porcentaje de silencio y silencio tampoco revelan buenos indicadores.

5.2.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN

Al aplicar la técnica "sobre-muestreo aleatorio con reposición" se obtuvo un nuevo conjunto de datos con dimensión de 1,400×21, con 700 instancias de clase "0" y 700 de clase "1". Las matrices de confusión de los clasificadores aplicados a este nuevo conjunto de datos fueron obtenidas del lenguaje de programación "R" (Tabla 15).

Tabla 15. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	520	**180**	700
1	**214**	486	700
Total	734	666	1400

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	521	**179**	700
1	**217**	483	700
Total	738	662	1400

VECINOS CERCANOS

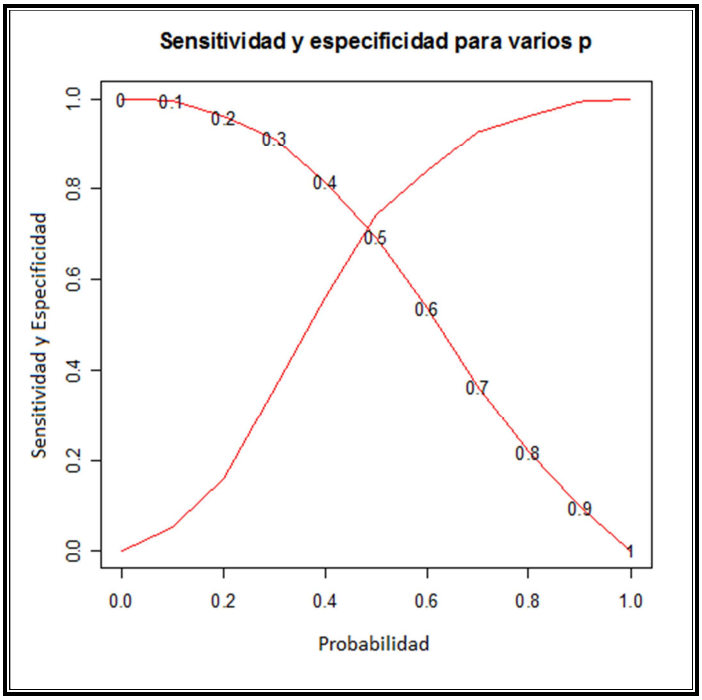
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	511	**189**	700
1	**301**	399	700
Total	812	588	1400

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	526	**174**	700
1	**233**	467	700
Total	759	641	1400

Se observa que el valor de p donde se intersectan las curvas de sensitividad y especificidad corresponde a $p = 0.5$ (Gráfico 7). Este valor se utilizó para calcular la matriz de confusión de método de regresión logística a través del lenguaje de programación "R".

Gráfico 7. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "GERMAN-SOBREMUESTREO"



La Tabla 16, ilustra la tasa de error, medida G, medida de ruido, medida de silencio y medida F, los criterios de evaluación que se calcularon a través de las matrices de confusión de los clasificadores. Los métodos de clasificación "regresión logística y discriminante lineal" arrojaron mejores resultados en el conjunto de datos "GERMAN" con clases balanceadas a través de la técnica "sobre-muestreo aleatorio con reposición". Sin embargo analizando los porcentajes de cada medida se observa que los resultados de los clasificadores muestran un desempeño no tan deseable.

Tabla 16. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	**0.28**	**0.72**	**0.29**	**0.26**	**0.73**
Discriminante Lineal	**0.28**	**0.72**	**0.29**	**0.26**	**0.72**
Vecinos más cercanos	0.35	0.65	0.37	0.27	0.68
Árboles de Decisión	0.29	0.71	0.31	0.25	0.72

5.2.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE"

Al aplicar la técnica de sobre-muestreo "SMOTE" se obtuvo un nuevo conjunto de datos con dimensión de 1,300 ×21, con 600 instancias de clase "0" y 700 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R". (Véase Tabla 17.)

Tabla 17. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	406	**194**	600
1	**206**	494	700
Total	612	688	1300

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	392	**208**	600
1	**189**	511	700
Total	581	719	1300

VECINOS CERCANOS

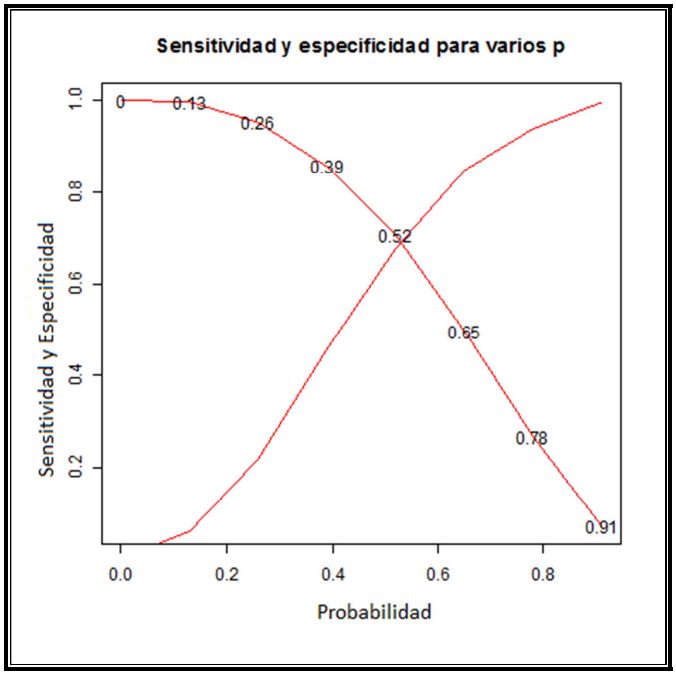
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	352	**248**	600
1	**242**	458	700
Total	594	706	1300

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	329	**271**	600
1	**201**	499	700
Total	530	770	1300

Para calcular la matriz de confusión del método de clasificación por regresión logística aplicado a los datos "GERMAN" balanceado a través de la técnica "SMOTE" se usó el valor de $p = 0.52$ (Gráfico 8).

Gráfico 8. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "GERMAN-SMOTE"



Los clasificadores "regresión logística y discriminante lineal" son los clasificadores que presentan mejores resultados. Note que la tasa de error de cada clasificador presenta un porcentaje alto, las medidas G y F son menores que 80%, y el ruido y el silencio están sobre 30% (Tabla 18). Estos resultados indican que los desempeños o rendimientos de los clasificadores no deseables.

Tabla 18. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	**0.31**	**0.69**	**0.34**	**0.32**	**0.67**
Discriminante lineal	**0.31**	**0.69**	**0.33**	**0.35**	**0.66**
Vecinos más cercanos	0.38	0.62	0.41	0.41	0.59
Árboles de decision	0.36	0.63	0.38	0.45	0.58

5.2.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN)

Al aplicar la combinación de las técnicas "SMOTE y ENN" se obtuvo un nuevo conjunto de datos con dimensión de 917 ×21, con 441 instancias de clase "0" y 476 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 19).

Tabla 19. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	287	**154**	441
1	**146**	330	476
Total	433	484	917

DISCRIMINANTE LINEAL

Valor real	*Valor pronosticado*		*Total*
	0	*1*	
0	283	**158**	441
1	**140**	336	476
Total	423	494	917

VECINOS CERCANOS

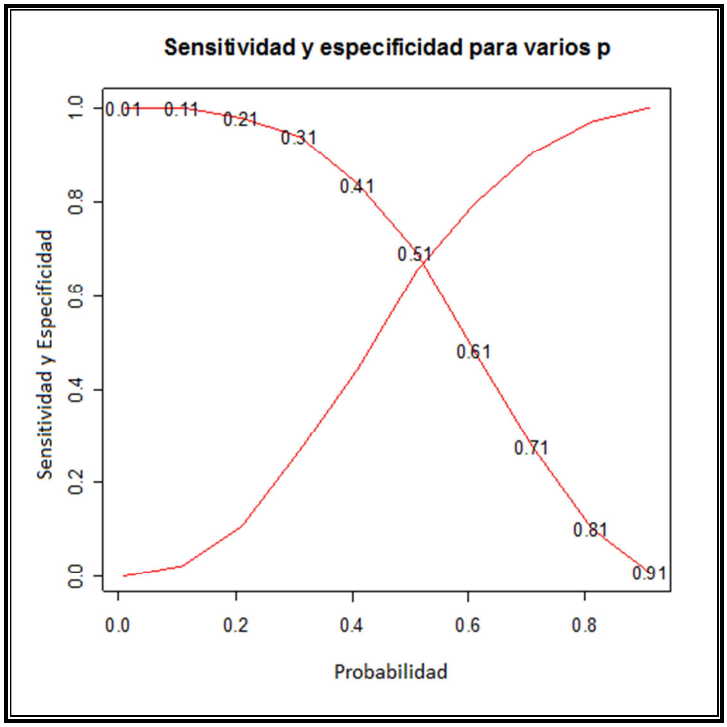
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	247	**194**	441
1	**172**	304	476
Total	419	498	917

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	225	**216**	441
1	**167**	309	476
Total	392	525	917

El Gráfico 9 muestra el valor de la probabilidad en el que se intersectan las curvas de sensitividad y especificidad. Este valor de $p = 0.51$, se utilizó para calcular la matriz de confusión del método de clasificación regresión logística.

Gráfico 9. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "GERMAN- SMOTE+ENN"



Los clasificadores "regresión logística y discriminante lineal" fueron los que presentaron mejor desempeño en comparación con los otros clasificadores (Tabla 20).

Tabla 20. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	**0.33**	**0.67**	**0.34**	**0.35**	**0.66**
Discriminante lineal	**0.32**	**0.67**	**0.33**	**0.36**	**0.66**
Vecinos más cercanos	0.40	0.60	0.41	0.44	0.57
Árboles de decision	0.42	0.58	0.43	0.49	0.54

5.2.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (GERMAN) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "Tomek Links"

Al aplicar la combinación de las técnicas "SMOTE y Tomek Links" se obtuvo un nuevo conjunto de datos de dimensión 1,232×16, el cual contiene 563 instancias de clase "0" y 669 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 21).

Tabla 21. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	390	**173**	563
1	**209**	460	669
Total	599	633	1232

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	353	**210**	563
1	**175**	494	669
Total	528	704	1232

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	361	**202**	563
1	**230**	439	669
Total	591	641	1232

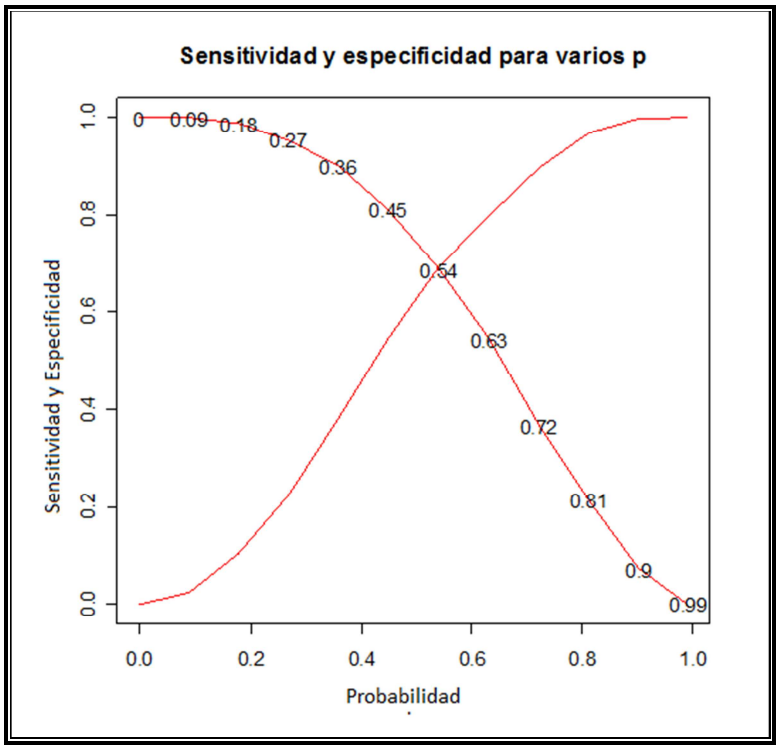
ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	331	**232**	563
1	**187**	482	669
Total	518	714	1232

El punto de intersección entre la curva de sensitividad y especificidad de los datos con clases balanceadas a través de la combinación "SMOTE y Tomek Links", corresponde al

valor de $p = 0.54$ (Gráfico 10). Este valor se utilizó para calcular la matriz de confusión del método de clasificación regresión logística.

Gráfico 10. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE P "GERMAN- SMOTE+TOMEK LINKS"



Los clasificadores "regresión logística y discriminante lineal" son los clasificadores que presentan mejores resultados (Tabla 22). Aunque la tasa de error de cada clasificador presenta un porcentaje alto, las medidas G y F son menores que 80%, el ruido y el silencio superan el 30%, los cuales indican que el desempeño o rendimiento de los clasificadores es moderado.

Tabla 22. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS GERMAN BALANCEADOS A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	**0.31**	**0.69**	**0.35**	**0.31**	**0.67**
Discriminante lineal	**0.31**	**0.68**	**0.33**	**0.37**	**0.65**
Vecinos más cercanos	0.35	0.65	0.39	0.36	0.63
Árboles de decisión	0.34	0.65	0.36	0.41	0.61

Analizando detalladamente cada uno de los resultados obtenidos al aplicar los métodos basados en muestreo en el conjunto de German, se observó que los métodos de clasificación "regresión logística y discriminante lineal" mostraron en comparación con los demás mejores resultados. Sin embargo, esto no indica que presentan buen desempeño o rendimiento, puesto que las métricas de evaluación así lo indican. Además si comparamos los resultados de las métricas de evaluación de todas las técnicas de muestreo aplicadas con las obtenidas en el conjunto original se puede notar que el mejoramiento en los desempeños de los clasificadores no es tan marcado. Esto tal vez ocurre a que se al desbalance de las clases no es muy pronunciado.

En la Tabla 23 se puede observar un resumen de los resultados de estas métricas aplicadas a regresión logística.

Tabla 23. RESULTADOS DE LAS MÉTRICAS DE EVALUACIÓN GENERALES APLICADAS REGRESIÓN LOGÍSTICA SOBRE LOS DATOS "GERMAN"

MÉTRICAS	SIN APLICAR TECNICAS	SOBREMUESTREO ALEATORIO	SMOTE	SMOTE+ENN	SMOTE+ TOMEK LINKS
Tasa de error	0.31	**0.28**	0.31	0.33	0.31
Medida G	0.69	**0.72**	0.69	0.67	0.69
Ruido	0.51	**0.29**	0.34	0.34	0.35
Silencio	0.30	**0.26**	0.32	0.35	0.31
Medida F	0.57	**0.73**	0.67	0.66	0.67

5.3 RESULTADOS PARA LOS DATOS "EST2"

Para reafirmar lo que se estableció anteriormente observe los resultados que muestra la Tabla 24, la cual contiene las métricas calculadas sobre las matrices de confusión de los clasificadores aplicados al conjunto de datos de dos clases desbalanceados EST2, cuyas clases contienen 2,718 instancias clasificadas en la clase "0" y 4,206 instancias en la clase "1". Esta Tabla recoge información del clasificador que obtuvo mejor desempeño y cada una de las métricas calculadas.

Tabla 24. RESULTADOS DE LAS MÉTRICAS DE EVALUACIÓN GENERALES APLICADAS ÁRBOLES DE DECISIÓN SOBRE LOS DATOS "EST2"

MÉTRICAS	SIN APLICAR TECNICAS	SOBREMUESTREO ALEATORIO	SMOTE	SMOTE+ENN	SMOTE+TOMEK LINKS
Tasa de error	0.34	0.38	0.36	0.35	0.35
Medida G	0.59	0.62	0.62	0.60	0.60
Ruido	0.40	0.39	0.33	0.35	0.34
Silencio	0.58	0.38	0.27	0.18	0.19
Medida F	0.50	0.62	0.70	0.73	0.73

5.4 RESULTADOS PARA LOS DATOS "EST1"

5.4.1 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO DESBALANCEADO EST1

Para llevar a cabo un análisis del rendimiento de los clasificadores que se aplicaron al conjunto con clases desbalanceadas "EST1", iniciaremos por interpretar las matrices de confusión para cada clasificador.

Tabla 25. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A EST1

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		Total
	0	*1*	
0	414	**148**	562
1	**1759**	4603	6362
Total	2173	4751	6924

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		Total
	0	*1*	
0	0	**562**	562
1	**0**	6362	6362
Total	0	6924	6924

VECINOS CERCANOS

Clase real	*Clase pronosticada*		Total
	0	*1*	
0	6	**556**	562
1	**5**	6358	6362
Total	13	6914	6924

ÁRBOLES DE DECISIÓN

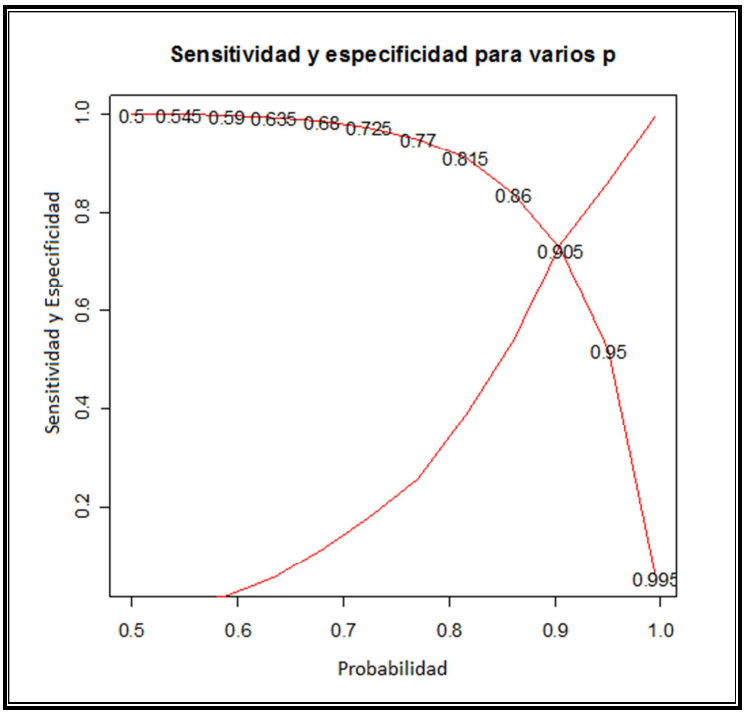
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	0	**562**	562
1	**0**	6362	6362
Total	0	6924	6924

En la Tabla 25, observamos las matrices de confusión para los clasificadores de regresión logística, árboles de decisión, vecinos más cercanos y discriminante lineal aplicados a los datos "EST1". De acuerdo con estas matrices se puede observar como

cada uno de los clasificadores tienen una tendencia de clasificación a clase “1”, excepto en este caso en el método de clasificación por regresión logística. Por ejemplo el método discriminante lineal clasificó 6,924 instancias en la “clase 1”, de las cuales 6,362 se encuentran bien clasificadas y 562 mal clasificadas; y no clasificó instancias en “clase 0”.

Para hallar la matriz de confusión en el caso de regresión logística se usó el valor de *p* óptimo sugerido por la gráfica de sensibilidad y especificidad para varios valores de *p*. Este valor es “0.905” (Gráfico 11). Los resultados de la matriz de confusión sobre los datos “EST” que produjo este método no muestran tendencia a la clase mayoritaria, en este caso a la clase 1.

Gráfico 11. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE *p* “EST1”



Al evaluar cada uno de los clasificadores aplicados al conjunto de datos EST1 se obtuvieron los siguientes resultados (Tabla 26).

Tabla 26. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A EST1

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.28	0.73	0.81	0.26	0.30
Discriminante lineal	0.08	0.00	NA	1.00	NA
Vecinos más cercanos	0.08	0.10	0.45	0.99	NA
Árboles de decisión	0.08	0.00	NA	1.00	NA

A pesar de que la tasa de error de cada clasificador en el conjunto datos de "EST1" representa un porcentaje de error adecuado (porque es pequeño) esto no significa que necesariamente el rendimiento de los clasificadores sea el mejor, pues como se observó anteriormente los clasificadores tienen una tendencia a la clase mayoritaria excepto en el caso de la regresión logística. Al aplicar y analizar la medida F, el ruido y silencio los clasificadores muestran resultados no muy recomendables a la hora de escoger un clasificador con buen rendimiento. Los valores que aparecen con NA representan las medidas cuyos cálculos resultaron de la forma indeterminada $\frac{0}{0}$. Además el porcentaje de silencio fue bastante alto excepto en regresión logística.

5.4.2 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA DE TÉCNICA DE SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN

Al aplicar la técnica "sobre-muestreo aleatorio con reposición" se obtuvo un nuevo conjunto de datos con dimensión de 12,724×7, con 6,362 instancias de clase "0" y 6,362 de clase "1". Usando el lenguaje de programación "R" se calcularon las matrices de confusión para cada clasificador (Tabla 27).

Tabla 27. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS EST1 BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	4605	**1757**	6362
1	**1695**	4667	6362
Total	6300	6424	12724

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5056	**1306**	6362
1	**2200**	4162	6362
Total	7256	5468	12724

VECINOS CERCANOS

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5130	**1232**	6362
1	**2285**	4077	6362
Total	7615	5109	12724

ÁRBOLES DE DECISIÓN

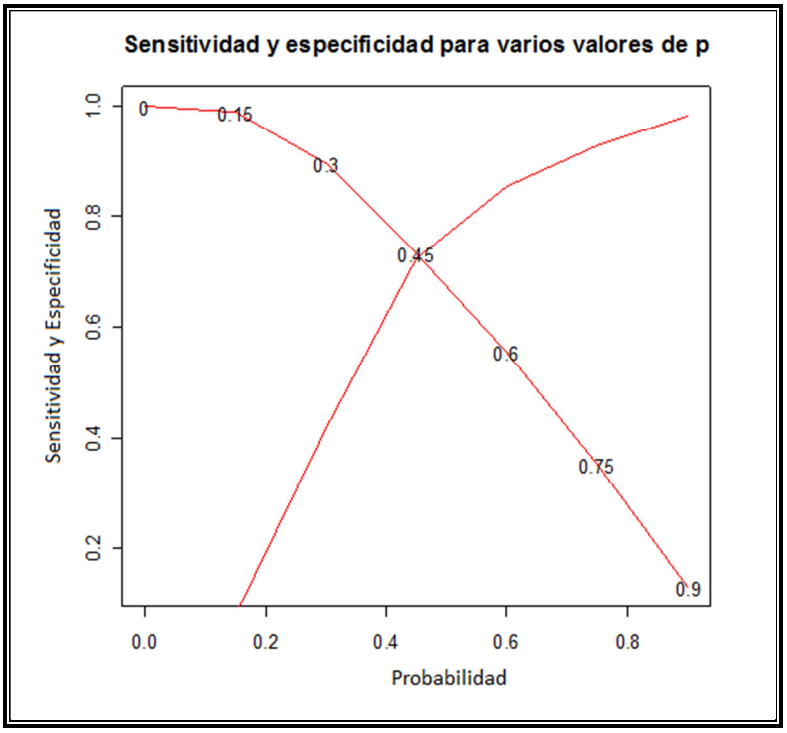
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	4659	**1703**	6362
1	**1780**	4582	6362
Total	6439	6285	12724

Al observar los resultados que arroja la Tabla 27 se puede concluir que el método de "regresión logística" clasificó 6424 instancias en la "clase 1", de las cuales 4667 se

encuentran bien clasificadas y 1757 mal clasificadas (falso1); y 6300 en “clase 0” , de las cuales 4605 se clasificaron correctamente y 1695 instancias incorrectamente (falso 0).

El valor de p optimo utilizado para calcular la matriz de confusión de regresión logística fue $p = 0.45$ que corresponde al valor sugerido por el gráfico de sensitividad y especificidad para varios valores de probabilidad (Gráfico 12).

Gráfico 12. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p “EST1” (SOBRE-MUESTREO)



De los resultados de las métricas que muestra la Tabla 28 se puede notar que los métodos arrojan resultados muy parecidos entre sí.

Tabla 28. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS A TRAVÉS DE LA TÉCNICA "SOBRE-MUESTREO ALEATORIO CON REPOSICIÓN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.27	0.73	0.27	0.28	0.73
Discriminante Lineal	0.28	0.72	0.30	0.21	0.74
Vecinos más cercanos	0.28	0.72	0.31	0.19	0.74
Árboles de Decisión	0.27	0.73	0.28	0.27	0.73

5.4.3 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA TÉCNICA DE SOBREMUESTRO "SMOTE"

Al aplicar la técnica de sobre-muestreo "SMOTE" se obtuvo un nuevo conjunto de datos con dimensión de 13,106×7, con 6,744 instancias de clase "0" y 6,362 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 29).

Tabla 29. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

REGRESION LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	4048	**2696**	6744
1	**2762**	3600	6362
Total	6810	6296	13106

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	4036	**2708**	6744
1	**2751**	3611	6362
Total	6787	6319	13106

VECINOS CERCANOS

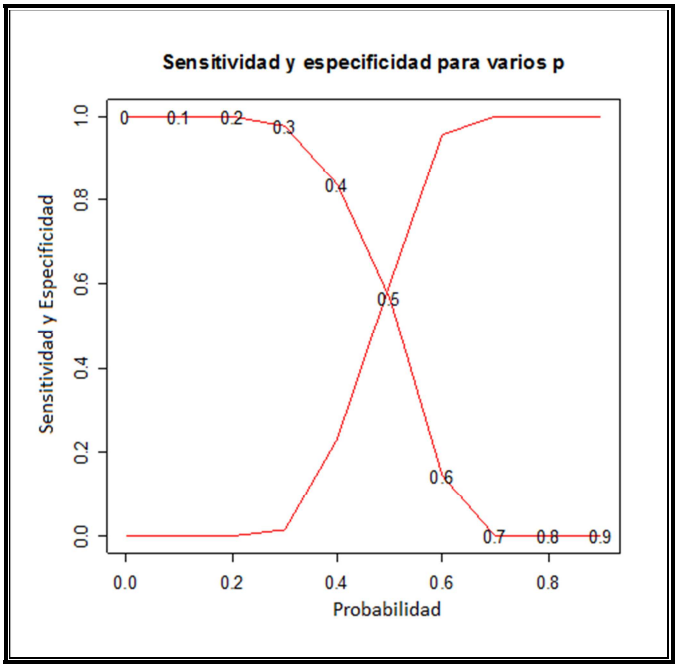
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5937	**807**	6744
1	**3562**	2800	6362
Total	9499	3607	13106

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5694	**1050**	6744
1	**4159**	2203	6362
Total	9853	3253	13106

El punto de intersección entre la curva de sensitividad y especificidad de los datos con clases balanceadas a través de la técnica "SMOTE", corresponde al valor de $p = 0.5$, el cual se utilizó para calcular la matriz de confusión del método de clasificación por regresión logística (Gráfico 13).

Gráfico 13. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "EST1 - SMOTE"



Según las métricas usadas para evaluar el desempeño o rendimiento de los clasificadores mostrados en la Tabla 30 se puede observar que el clasificador "vecinos más cercanos" presenta mejores resultados en comparación con los demás. Muestra una tasa de error de 33%, una medida G de 62%, un ruido de 37%, un silencio de 12% y una medida F de 73% .

Tabla 30. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS EST1 BALANCEADOS A TRAVÉS DE LA TÉCNICA "SMOTE"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.42	0.58	0.41	0.40	0.60
Discriminante lineal	0.42	0.58	0.41	0.40	0.60
Vecinos más cercanos	**0.33**	**0.62**	**0.37**	**0.12**	**0.73**
Árboles de decisión	**0.40**	**0.54**	**0.42**	**0.16**	**0.69**

5.4.4 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y SUBMUESTREO "LIMPIEZA" DE WILSON (ENN)

Al aplicar la combinación de las técnicas "SMOTE y ENN" se obtuvo un nuevo conjunto de datos con dimensión de 11,036×7, con 4,456 instancias de clase "0" y 6,580 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R" (Tabla 31).

Tabla 31. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS EST1 A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	4043	**2537**	4456
1	**1844**	2612	6580
Total	5887	5149	11036

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5623	**957**	4456
1	**3027**	1429	6580
Total	8750	2485	11036

VECINOS CERCANOS

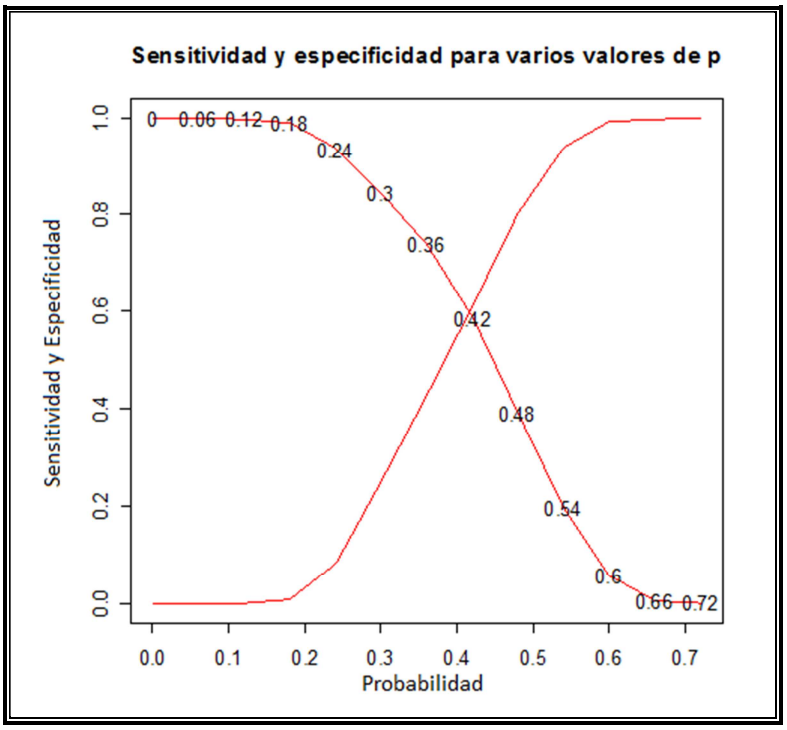
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	6199	**381**	4456
1	**2611**	1845	6580
Total	8810	2226	11036

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	6070	**510**	4456
1	**3123**	1333	6580
Total	9193	1843	11036

El valor óptimo p usado para calcular la matriz de confusión del método de clasificación regresión logística fue de $p = 0.42$ sugerido por el Gráfico 14.

Gráfico 14. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "EST1 SMOTE+ENN"



Según los resultados de las métricas de evaluación que muestra la Tabla 32 se destaca el clasificador "vecinos más cercanos" por obtener mejores resultados en comparación con los demás clasificadores. Este clasificador tiene una tasa de error de 27%, una medida G de 62%, un ruido de 30%, un silencio de 0.06% y una medida F de 81%.

Tabla 32. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS EST1 BALANCEADOS EST1 A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y ENN"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.40	0.60	0.31	0.39	0.65
Discriminante lineal	0.36	0.52	0.35	0.15	0.74
Vecinos más cercanos	**0.27**	**0.62**	**0.30**	**0.06**	**0.81**
Árboles de decisión	**0.33**	**0.53**	**0.34**	**0.08**	**0.77**

5.4.5 ANÁLISIS DEL RENDIMIENTO DE LOS CLASIFICADORES APLICADOS AL CONJUNTO BALANCEADO (EST1) A TRAVÉS DE LA COMBINACION DE TÉCNICAS DE SMOTE Y MÉTODO DE LIMPIEZA "Tomek Links"

Al aplicar la combinación de las técnicas "SMOTE y Tomek Links" se obtuvo un nuevo conjunto de datos de dimensión 11,977×7, el cual contiene 6,180 instancias de clase "0" y 5,797 de clase "1". Las matrices de confusión de clasificadores aplicados a este nuevo conjunto de datos se calcularon en el lenguaje de programación "R". (Véase Tabla 33.)

Tabla 33. MATRIZ DE CONFUSIÓN DE LOS CLASIFICADORES APLICADOS A LOS DATOS EST1 BALANCEADOS EST1 A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

REGRESIÓN LOGÍSTICA

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	3512	**2668**	6180
1	**2277**	3520	5797
Total	5789	6188	11977

DISCRIMINANTE LINEAL

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	3748	**2432**	6180
1	**2453**	3344	5797
Total	6201	5776	11977

VECINOS CERCANOS

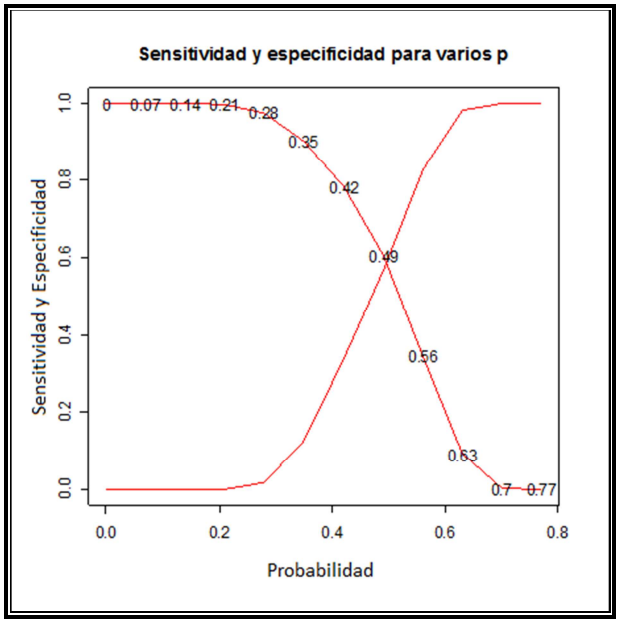
Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5505	**675**	6180
1	**3046**	2751	5797
Total	8551	3426	11977

ÁRBOLES DE DECISIÓN

Clase real	*Clase pronosticada*		*Total*
	0	*1*	
0	5663	**517**	6180
1	**4129**	1668	5797
Total	9792	2185	11977

El Gráfico 15 muestra un valor de probabilidad óptimo de $p = 0.49$, el cual se utilizó calcular la matriz de confusión del método de clasificación regresión logística.

Gráfico 15. SENSITIVIDAD Y ESPECIFICIDAD PARA VARIOS VALORES DE p "EST1- SMOTE+TOMEK LINKS"



En la Tabla 34, se observan los resultados de las métricas de evaluación para todos los métodos de clasificación con los que trabajó en esta tesis. El algoritmo de clasificación "vecinos más cercanos" presenta mejores resultados en comparación con los demás

algoritmos. Este clasificador tiene una tasa de error de 31%, y medida G, ruido, silencio y F de 65%, 36%, 11% y 75% respectivamente.

Tabla 34. MÉTRICAS DE EVALUACIÓN DE CLASIFICADORES APLICADOS A LOS DATOS BALANCEADOS EST1 A TRAVÉS DE LA COMBINACIÓN DE "SMOTE Y TOMEK LINKS"

CLASIFICADORES	Tasa Error	Medida G	Ruido	Silencio	Medida F
Regresión logística	0.41	0.59	0.39	0.43	0.59
Discriminante lineal	0.41	0.59	0.40	0.39	0.61
Vecinos más cercanos	**0.29**	**0.65**	**0.36**	**0.11**	**0.75**
Árboles de decisión	**0.39**	**0.51**	**0.42**	**0.08**	**0.71**

Ahora se analizarán detalladamente los resultados de las técnicas de muestreo que se aplicaron para solucionar el problema de distribución de clases desbalanceadas de los datos EST1 teniendo en cuenta los clasificadores que nos brindaron mejor desempeño (vecinos más cercanos). Se tomaran en cuenta las métricas de evaluación usadas para realizar este análisis general.

Tabla 35. RESULTADOS DE LAS MÉTRICAS DE EVALUACIÓN GENERALES APLICADAS A VECINOS MÁS CERCANOS SOBRE LOS DATOS "EST1"

MÉTRICAS	SIN APLICAR TECNICAS	SOBREMUESTREO ALEATORIO	SMOTE	SMOTE+ENN	SMOTE+ TOMEK LINKS
Tasa de error	0.08	0.28	0.33	**0.27**	0.29
Medida G	0.10	**0.72**	0.62	0.62	0.65
Ruido	0.45	0.31	0.37	**0.30**	0.36
Silencio	0.99	0.19	0.12	**0.06**	0.11
Medida F	NA	0.74	0.73	**0.81**	0.75

Según los resultados que arroja la Tabla 35, se puede concluir que la técnica que además de aliviar el problema de desbalance de dos clases, nos proporciona un mejor desempeño en comparación con los clasificadores en el conjunto de datos EST1 original (con clases desbalanceada) es la combinación de la técnica "SMOTE" y el método de limpieza de Wilson (ENN) con tasa de error, medidas G, ruido, silencio y F de 27%, 62%, 30%, 6% y 81% respectivamente. Cabe resaltar que los resultados cuando se aplica la técnica de sobre-muestreo aleatorio nos revelan un mejoramiento casi similar al de SMOTE+ENN, aunque ya sabemos de antemano que esta técnica tiene sus desventajas, que son superadas al implementar SMOTE+ENN. El desempeño de los clasificadores no fue el esperado, pero tal vez pudo haber factores en la clasificación inicial que incidan en estos resultados.

6 CONCLUSIONES

El propósito de esta tesis fue abordar el problema de dos clases desbalanceadas en los procesos de clasificación a través de la implementación de técnicas al nivel de datos, es decir basadas en muestreo, y aplicar clasificadores apropiados a este dominio con el fin de evaluar el rendimiento de éstos.

Los clasificadores aplicados a los datos CRX, GERMAN, EST1 y EST2 con clases desbalanceadas mostraron matrices de confusión con tendencia marcada a la clase que tiene mayor cantidad de datos, excepto en el caso de "regresión logística". Al aplicar cada una de las medidas de evaluación a los clasificadores, se observó de manera general que aunque las tasas de mala clasificación en algunos casos son pequeñas, las otras medidas tales como el ruido, el silencio y medidas F y G arrojaron resultados que nos indican que el desempeño de los clasificadores: árboles de decisión, regresión logística, vecinos más cercanos y discriminante lineal fue poco adecuado.

Cuando se aplican las técnicas de sobre-muestreo aleatorio, sobre-muestreo "SMOTE", combinación de la sobre-muestreo "SMOTE" con los métodos de limpieza "ENN" y "Tomek Links" sobre los datos antes mencionados se pudo notar que los resultados de las métricas usadas para evaluar el desempeño de los clasificadores muestran mejores resultados cuando se aplican las combinaciones de sobre-muestreo "SMOTE" con los

métodos de limpieza “ENN” y “Tomek Links” , siendo la combinación de “SMOTE y ENN” la que indicó mejor desempeño en los clasificadores.

De los clasificadores aplicados a los conjuntos de datos balanceados a través de las técnicas de muestreo, predominó el algoritmo de clasificación de árboles de decisión, aunque todos los clasificadores usados son adecuados en este tipo de dominios.

Por último es importante destacar que a través de este estudio se logró visualizar que el efecto de las técnicas aplicadas a los datos fue más impactante en los conjuntos donde el desbalance entre las clases fue muy pronunciado (“CRX” y “EST1”), que con conjuntos con desbalance no muy pronunciado (“GERMAN” y “EST2”). Esto podría mostrar que el uso de estas técnicas es más apropiado cuando el desbalance de las dos clases es muy marcado.

REFERENCIAS

[1] N. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer. SMOTE Boost: Improving prediction of the minority class in Boosting. Journal: Principles of data mining and knowledge discovery 2003.pp. 1-12

[2] N. Chawla, N. Japkowicz, A. Kotcz. Editorial: Special issue on learning from imbalanced data sets sigkdd Explorations. Volume 6, Issue 1. 2004 pp.1-6

[3] E. Acuña. Discriminación usando regresión logística. Departamento de Matemáticas, Universidad de Puerto Rico-Recinto Universitario de Mayagüez. 2009.

[4] P. Juez, J. Diez. "Probabilidad y estadística en medicina". Ediciones Díaz de santos S.A, Madrid España, 1997. pp. 177

[5] J. Marín. Análisis discriminante. Departamento de Estadística, Universidad Carlos III de Madrid. 2010

[6] L. Vieira, L. Ortiz, S. Ramírez. Introducción a la minería de datos. Editorial ltda, Rio de Janeiro Brasil. 2009

[7] M. Pena, Análisis de datos multivariantes. Editorial McGRAW. Espana.2002

[8] G. Choque. Árboles de decisión. Editorial diario.net. 2009.

[9] E. Acuña. Árboles de decisión. Departamento de Matemáticas, Universidad de Puerto Rico- Recinto Universitario de Mayagüez. 2009.

[10] J. Marín. Árboles de decisión. Departamento de Estadística, Universidad Carlos III de Madrid.

[11] L. Breiman, J. Friedman, R. Ohlsen & C. Stone (1984). Classification and regression trees. Wadsworth, Belmont, CA.

[12] E. Fix, J. Hodges. "Discriminatory analysis nonparametric discrimination: consistency properties", Technical Report 4, Project No. 21-49-004, USAF School of Aviation Medicine, Randolph field, Texas 1951.

[13] R.M. Valdovinos. Técnicas de submuestreo, toma de decisiones y análisis de diversidad en aprendizaje supervisado con sistemas múltiples de clasificación, pp.34, 2006

[14] E. Acuña. Clasificación basada en los vecinos más cercanos. Departamento de Matemáticas, Universidad de Puerto Rico- Recinto Universitario de Mayagüez.

[15] http://es.wikipedia.org/wiki/Knn#Vecinos_m.C3.A1s_cercanos_con_distancia_ponderada

[16] J. Moreno, D. Rodríguez, M. Sicilia, J. Riquelme, R. Ruiz. SMOTE -I: Mejora del algoritmo SMOTE para balanceo de clases minoritarias. Departamento de ciencias de la Computación. Escuela Técnica Superior de Ingeniería Informática. Escuela Politécnica Superior, 2009.

[17] S. Suresh, N. Sundararajan, y P. Saratchandran. Risksensitive loss functions for sparse multicategory classification problems. Information Sciences, 178(12):2621-2638, 2008.

[18] Y. M. Huang, C. M. Hung, y H. C. Jiau. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 7(4):720-747, 2006.

[19] L. Bruzzone y S.B. Serpico. Classification of imbalanced remote-sensing data by neural networks. Patterns Recognition. Letters, 18(11-13):1323-1328, 1997.

[20] T. E. Fawcett y F. Provost. Adaptive Fraud Detection. Data mining and knowledge Discovery 3 (1997) 291-316

[21] M. Kubat, R. Holte, S. Matwin. Machine Learning for the oil spills in satellite radar images. Machine Learning 30 (1998) 195-215

[22] D. Lewis, J. Catlett. Heterogeneous uncertainty sampling for supervised learning. Eleventh International Conferences of Machine Learning (1994) 148-156

[23] N. Japkowicz, C. Myers y M. Gluck. A novelty detection appoach to classification. Fourteenth Joint Conferences on Artificial Intelligence (1995) 518-523

[24] X. Peng y I. King. Robust BMPM training based on second-order cone programming and its application in medical diagnosis. Neural Networks, 21(2-3):450-457, 2008.

[25] F. Herrera. Curso: M6: Nuevas tendencias en minería de datos. Departamento Ciencias de la Computación. Universidad de Granada. España, 2008.

[26] C. Ling y Li. Data mining for direct marketing problems and solutions. Fourth International Conference on Knowledge Discovery and Data Mining, 1998.

[27] P. Domingo. Metacost: A General method for making classifiers cost-sensitive.Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155-164, 1999.

[28] R.M. Valdovinos. Técnicas de submuestreo, toma de decisiones y análisis de diversidad en aprendizaje supervisado con sistemas múltiples de clasificación, pp. 54, 2006

[29] N. Chawla, K. Bowyer, L. Hall and P. Kegelmeyer. SMOTE: Synthectic Minority Over-Sampling Technique. Journal of Artificial Intelligence. pp. 321-357, 2002.

[30] I. King, J. Wang, L. Chan y D. Wang. Neural information processing. 13th International conference, ICONIP 2006, Hong Kong, China, October 2006. Proceedings, part III. pp. 22-23

[31] J. Porta, "Técnicas cuantitativas para la extracción de términos en un corpus", Universidad autónoma de Madrid, 2006.

[32] M. Kubat y S. Matwin. Addresing the curse of imbalanced training set: One sided selection. University Otawa, Canada.

[33] N. Chinchor. MUC-4 evaluation metrics. MUC'1992. pp.22-29.