# Simulación de Sistemas Y Laboratorio: Entrega Parcial Proyecto

**Jaime Londoño Ciro**
**Universidad de Antioquia**

**Sebastián Jiménez Caro**
**Universidad de Antioquia**

**Resumen—** Se abordará un problema de clasificación donde se evalúa la esperanza de vida de un paciente de cirugía de tórax debido a cáncer de pulmón. Se aplicarán los métodos de Discriminantes Gaussianas, K-vecinos más cercanos, Random Forest y Redes Neuronales Artificiales, con el fin de encontrar el más eficiente para este problema particular. En cada método se usará Cross-Validation como técnica de validación y se aplicaran técnicas de submuestreo y sobremuestreo inteligente para tratar los problemas de desbalance de los datos.

**Palabras clave:** Clasificación, Cirugía de tórax, Discriminantes Gaussianas, Redes Neuronales Artificiales, Knn, Cross-Validation, submuestreo, sobremuestreo.

## 1. INTRODUCCIÓN

En este artículo se presentan modelos de predicción para clasificar la esperanza de vida de pacientes de cáncer de pulmón luego de ser sometidos a cirugía. Esto es, de acuerdo a una serie de características del paciente, diagnosticar si sobrevive más de un año luego de la cirugía. El objetivo es definir la mejor técnica entre las siguientes para abordar este problema en particular: Discriminantes Gaussianas, K-vecinos más cercanos, Random Forest y Redes Neuronales Artificiales.

El sistema que salga definido como el más eficiente, puede ser bastante útil en el área de la salud, específicamente en aquellos departamentos encargados de evaluar la condición de personas que han sido sometidas a cirugía debido a cáncer de pulmón. Área en la que en los datos que registra suele presentarse desproporción en la distribución de clases, lo que se conoce como problema de desbalanceo de datos.

Para este modelamiento, se utiliza un conjunto de datos brindado por el repositorio para aprendizaje de máquinas de la Universidad de Irvine. (UCI Machine Learning Repository) llamado Thoracic Surgery Data [1].

A continuación se enumeran las variables de entrada que posee el sistema:

1. DGN: Diagnóstico— Combinación específica del código ICD-10 para identificar el tipo de tumor. (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
2. PRE4: Capacidad Vital Forzada - FVC (numérico)
3. PRE5: Volumen que ha sido exhalado al final del primer segundo de espiración forzada - FEV1 (numérico)
4. PRE6: Calidad de vida del paciente - escala Zubrod (PRZ2,PRZ1,PRZ0)
5. PRE7: Dolor antes de la cirugía (T,F)

6. PRE8: Hemoptisis antes de la cirugía (T,F)
7. PRE9: Disnea antes de la cirugía (T,F)
8. PRE10: Tos antes de la cirugía (T,F)
9. PRE11: Estado débil antes de la cirugía (T,F)
10. PRE14: T en el sistema de estadificación TNM - tamaño del tumor primario, siendo OC11 el más pequeño y OC14 el más grande (OC11,OC14,OC12,OC13)
11. PRE17: El paciente presenta Diabetes Mellitus Tipo 2 (Diabetes del adulto) (T,F)
12. PRE19: El paciente ha presentado Infarto Agudo de Miocardio en los últimos 6 meses (T,F)
13. PRE25: El paciente presenta Enfermedad Arterial Periférica (EAP). (T,F)
14. PRE30: El paciente fuma (T,F)
15. PRE32: El paciente presenta Asma (T,F)
16. AGE: Edad del paciente al momento de la cirugía (numérico)

Mientras que la variable de salida o variable a predecir es la siguiente:
1. Risk1Y: Periodo de supervivencia de un año - Valor (T)rue si murió (T,F)

## 2. TRABAJOS RELACIONADOS

Existen, en la literatura científica, pocos aportes conocidos que aborden problemas con la misma base de datos.

En [2] se presenta la técnica de Máquina de Vectores de Soporte (SVM) para solucionar el problema del desbalance en la base de datos. Esta solución propuesta combina los beneficios de usar clasificadores de ensamble para datos desiguales, junto con Máquinas de Vectores de Soporte Sensibles al Costo (CSVM).
Para la evaluación del sistema, se utiliza la técnica Stratified Cross-Validation, que consiste en conseguir que en cada subconjunto en el que se divide el total de muestras, haya igual número de ejemplares de cada clase representada.
Entre los resultados obtenidos en este trabajo, destacan el uso de la técnica UnderBagging que obtuvo un 68.57% en True Positive, y un 100% en True Negative, usando la técnica JRip.

En [3] se presentan diferentes técnicas de aprendizaje con las que se aborda el problema de predicción de la esperanza de vida de pacientes de cirugía por cáncer de pulmón. Entre ellas están Naive Bayesian Classifier, PART (Partial Decision Trees), J48 Decision Tree, OneR Classifier, Random Forest y Decision Stump.
Como técnica de evaluación de los sistemas, se utiliza K-fold Cross-Validation. En el artículo se concluye que la mejor técnica para resolver este problema es el Random Forest, que devuelve un 95.65% de precisión.

En [4] se trabaja con el mismo problema y la misma base de datos, donde se reconoce también el problema de desbalance de los datos. Abordan el problema implementando la técnica Ensemble SVM (ESVM).
Además, se usa Cross-Validation de 10 folds con 10 repeticiones como técnica de evaluación del sistema. Y por último comparan los resultados obtenidos con los resultados de usar

SVM entrenado usando SMO, y con los resultados de Cost-Sensitive SVM entrenado con SMO modificado. De esto, se deja muestra que si bien la precisión fue más baja (60.37%), el valor de GMean y el porcentaje de las muestras de la clase positiva detectadas correctamente fueron mayores que en los otros dos métodos (47.33% de Gmean y 33.79% de TruePositive).

En [5], si bien no se trabaja con la misma base de datos, sí se trabaja un problema similar, y su enfoque está en una base de datos para detectar cáncer de próstata. Por otro lado, en [6] se enfocan en diferentes técnicas para afrontar el desbalanceo en diferentes conjuntos de datos que plantean un acercamiento diferente a este tipo de problemas.

## 3. EXPERIMENTOS

Como se estableció previamente, en este artículo se trabaja el problema de clasificación que plantea la base de datos Thoracic Surgery [1], en la que se cuenta con 470 muestras, donde 400 pertenecen a la clase etiquetada “negativa” y 70 de la “positiva”, lo cual deja en evidencia el problema de desbalance tan grande que posee la base de datos.

Este problema fue tratado aplicando técnicas de sobremuestreo inteligente sobre la clase minoritaria y submuestreo inteligente sobre la clase mayoritaria, logrando así a un mismo tiempo crear nuevas muestras de una clase y eliminar los datos ruidosos o atípicos de la otra.

Se utilizaron varios experimentos para la predicción de la esperanza de vida de los pacientes. En primer lugar está la técnica de Funciones Discriminantes Gaussianas como modelo paramétrico. Segundo, está la técnica de los k-vecinos más cercanos (KNN). Luego, el modelo Random Forest, y por último el uso de Redes Neuronales Artificiales.

En todos estos casos se usa Cross-Validation como método de validación, mientras que se tomarán como criterio de validación el error de clasificación, la especificidad, la sensibilidad y porcentaje de falsos negativos, ya que nos interesa más que el sistema tenga menos porcentaje de pacientes que se diga que mueren dentro de un año luego de la cirugía, pero el sistema los clasifica como que sobreviven, pues existe la posibilidad de que un paciente esté tranquilo debido a que no morirá en el primer año, y que sí muera en ese periodo, esta medida es igual a 1 - Sensibilidad.

## 4. RESULTADOS

- **Funciones Discriminantes Gaussianas**

Este modelo es paramétrico y crea una función discriminante para cada una de las clases de la base de datos. Y con cada nueva muestra se calcula la probabilidad de pertenecer a cada una de las clases mediante dichas funciones, y se clasifica de acuerdo al resultado cuya probabilidad sea mayor.
Los resultados de esta técnica son mostrados en la Tabla1.

| | Datos balanceados | Datos sin balancear |
|---|---|---|
| **Intervalo de Confianza** | +- 0.0623 | +-0.0302 |
| **Sensibilidad** | 0.2 | 0 |
| **Especificidad** | 0.8 | 0.9926 |
| **Eficiencia** | 0.5286 | 0.8447 |
| **Media Geométrica** | 0.4 | 0 |

**Tabla1. Resultados arrojados por Discriminantes Gaussianas.**

Este modelo se aplicó inicialmente sobre la base de datos original con el objetivo de evidenciar los problemas que el desbalance ocasiona. Se observa en la tabla que mientras los valores de eficiencia y especificidad arrojan muy buenos resultados, la media geométrica y la sensibilidad no lo hacen. Esto demuestra que el sistema aprendió muy bien como clasificar la clase mayoritaria, mientras para la minoritaria todas sus predicciones fueron erradas. De lo anterior, se concluye entonces que una mejor medida de ajuste del modelo al problema es la media geométrica, debido a que toma en cuenta las muestras bien clasificadas de cada clase para arrojar su resultado.

Una vez balanceadas las muestras, la aplicación de este modelo deja en evidencia su bajo desempeño para el conjunto de datos que se tiene, debido a su simplicidad (modelamiento de una clase con una sola función de probabilidad), en relación con la complejidad del problema que se trata.

- **K-vecinos más cercanos (KNN)**

En esta técnica, se hace la clasificación dependiendo de la moda de los k-vecinos más cercanos a la nueva muestra que se está evaluando. El número de vecinos se entrega como parámetro del modelo. Los resultados fueron registrados en la Tabla2.

| K-Vecinos | Intervalo de Confianza | Sensibilidad | Especificidad | Eficiencia | Media Geométrica |
|---|---|---|---|---|---|
| 3 | +- 0.0849 | 0.8384 | 0.6204 | 0.7262 | 0.7212 |
| 5 | +- 0.0753 | 0.7721 | 0.4983 | 0.6357 | 0.6202 |
| 7 | +- 0.0570 | 0.7256 | 0.5674 | 0.6381 | 0.6417 |
| 9 | +- 0.0947 | 0.7918 | 0.5806 | 0.6810 | 0.6780 |
| 11 | +- 0.0647 | 0.8049 | 0.4916 | 0.6524 | 0.6290 |

**Tabla2. Resultados obtenidos por Knn.**

El mejor resultado para este modelo se alcanzó con tres vecinos como parámetro, siendo mayor el porcentaje de muestras de la clase etiquetada “negativa” (pacientes que sobreviven luego de un año) bien clasificadas, que los de la clase “positiva” (pacientes que mueren antes de pasar un año de la cirugía).
En la tabla se evidencia además, que el modelo obtiene mejores resultados para una clase que para otra (mayor sensibilidad que especificidad para todos los valores de k), lo cual es un buen signo, dado que un error de clasificación en la clase “negativa” es más perjudicial en el contexto de este problema, que un error en la clase “positiva”.

- **Random Forest**

Este modelo utiliza árboles de decisión, reunidos por medio de la función TreeBagger, proporcionada por una librería de Matlab [7]. Este modelo es no paramétrico, y recibe el número de árboles y número de atributos que serán evaluados en cada nodo de decisión. Los resultados se detallan en la Tabla3.

| Número de árboles | Intervalo de Confianza | Sensibilidad | Especificidad | Eficiencia | Media Geométrica |
|---|---|---|---|---|---|
| 10 | +- 0.0728 | 0.7756 | 0.6902 | 0.7310 | 0.7317 |
| 20 | +-0.0795 | 0.7153 | 0.6815 | 0.7000 | 0.6982 |
| 30 | +-0.0552 | 0.7571 | 0.7056 | 0.7238 | 0.7309 |
| 40 | +-0.0968 | 0.7755 | 0.7439 | 0.7595 | 0.7595 |
| 50 | +-0.0635 | 0.7963 | 0.7302 | 0.7619 | 0.7625 |

**Tabla3. Resultados de la técnica Random Forest**

El mejor resultado para este modelo se obtuvo con un comite de 50 árboles, siendo no solo su media geométrica la mejor, sino también obteniendo buenos resultados para la sensibilidad y la especificidad, lo cual es una muestra del buen ajuste del modelo a ambas clases y con ello la alta confiabilidad en las predicciones realizadas por el sistema.

- **Redes Neuronales Artificiales**

Este modelo está formado por diversas unidades de procesamiento interconectadas, llamadas perceptrones. El número de perceptrones es uno de los parámetros de este modelo. Los resultados arrojados en este modelo se exponen en la Tabla4.

| Número de perceptrones | Intervalo de Confianza | Sensibilidad | Especificidad | Eficiencia | Media Geométrica |
|---|---|---|---|---|---|
| 5 | +- 0.0613 | 0.6692 | 0.4752 | 0.5762 | 0.5639 |

| | | | | | |
|---|---|---|---|---|---|
| 7 | +- 0.0972 | 0.6321 | 0.6181 | 0.6024 | 0.6250 |
| 9 | +- 0.0685 | 0.6108 | 0.6338 | 0.6095 | 0.6222 |
| 11 | +- 0.0781 | 0.5984 | 0.6077 | 0.5929 | 0.6031 |
| 13 | +- 0.0816 | 0.6224 | 0.6616 | 0.6381 | 0.6417 |
| 15 | +- 0.0891 | 0.6585 | 0.5404 | 0.6024 | 0.5965 |

**Tabla4. Resultados del modelo RNA.**

En la tabla no se incluyeron los resultados obtenidos con un número de perceptrones mayor debido a que a pesar del alto costo computacional que representan, no alcanzaron mejores resultados que los mostrados en la tabla para el problema.

El mejor resultado conseguido por esta técnica fue el dado con el número de perceptrones igual a 5. Si bien la eficiencia más alta se obtuvo con 13 perceptrones, la sensibilidad es más baja que con 5 perceptrones y esto en términos del contexto del problema representa un error mayor.

## 5. DISCUSIÓN

En general, en todos los modelos evaluados se obtuvo una eficiencia menor que en los casos de trabajos relacionados donde se trabaja con la misma base de datos. Como se decidió tomar la sensibilidad como el mejor criterio de validación, con un peso mayor incluso que el del error de clasificación, se le da más importancia a este resultado que a la eficiencia en sí. Esto es debido a que es más importante que el sistema evalúe mejor los pacientes que sobreviven en el período de un año después de la cirugía.

A pesar de que la eficiencia más alta fue obtenida por la técnica de Random Forest (76% con 50 árboles), no se escoge como el mejor modelo, ya que al verificar los valores de Sensibilidad, la técnica que mejor desempeño tuvo fue Knn (Sensibilidad 84%, Eficiencia 72% con k=3).

## 6. CONCLUSIONES

La selección e implementación de la técnica más adecuada para modelar los datos es tan importante como el tratamiento del desbalanceo en las bases de datos que cuenten con este problema para obtener un bajo error de clasificación en todas las clases.

Como consecuencia de lo anterior, se evidencia que no siempre una alta eficiencia es la mejor medida de desempeño de un sistema, ya que esta no toma en cuenta el ajuste del modelo a las diferentes clases y puede ocultar los problemas de desbalance presentes en un conjunto de datos.

En aplicaciones médicas es importante conocer que tipos de error son más nocivos, dado que estos son los que se deben minimizar.

## 7. REFERENCIAS


**[1]** K. Bache y M. Lichman, «UCI Learning Machine Repository,» University of California, School of Information and Conputer Sciences, [En línea] Disponible en: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

**[2]** Maciej Zięba, et al. "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients". Applied Soft Computing, Volume 14, Part A, January 2014, Pages 99-108, ISSN 1568-4946, Disponible en: http://dx.doi.org/10.1016/j.asoc.2013.07.016

**[3]** Sindhu. V, et al. "Thoracic Surgery Analysis Using Data Mining Techniques". Int.J.Computer Technology & Applications,Vol 5 (2),578-586. ISSN:2229-6093, Disponible en: http://www.ijcta.com/documents/volumes/vol5issue2/ijcta2014050253.pdf

**[4]** Zieba, M.; Swiatek, J., "Ensemble SVM for imbalanced data and missing values in postoperative risk management," e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on , vol., no., pp.95,99, 9-12 Oct. 2013, Disponible en: 10.1109/HealthCom.2013.6720646

**[5]** Artan, Y.; Haider, M.A.; Langer, D.L.; van der Kwast, T.H.; Evans, A.J.; Yongyi Yang; Wernick, M.N.; Trachtenberg, J.; Yetik, I.S., "Prostate Cancer Localization With Multispectral MRI Using Cost-Sensitive Support Vector Machines and Conditional Random Fields," Image Processing, IEEE Transactions on , vol.19, no.9, pp.2444,2455, Sept. 2010
doi: 10.1109/TIP.2010.2048612

**[6]** Haibo He; Garcia, E.A., "Learning from Imbalanced Data," Knowledge and Data Engineering, IEEE Transactions on , vol.21, no.9, pp.1263,1284, Sept. 2009
doi: 10.1109/TKDE.2008.239

**[7]** The MathWorks, Inc., «MathWorks,» MathWorks Inc., 2014. [En línea]. Disponible en: http://www.mathworks.com/help/stats/treebagger.html