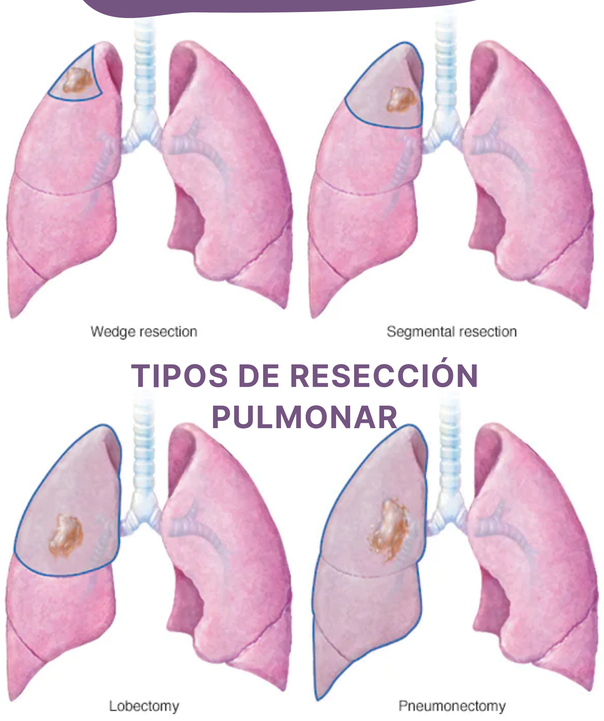
CLASIFICACIÓN DEL ÉXITO DE LAS CIRUGÍAS DE PULMÓN UTILIZANDO MODELOS DE APRENDIZAJE AUTOMÁTICO

VARGAS- DE LA ROSA Paola Itzel
paola.vargasdlr@uanl.edu.mx
Universidad Autónoma de Nuevo León

INTRODUCCIÓN



TIPOS DE RESECCIÓN PULMONAR

Fuente: Mayo Foundation for medical education and research

En los últimos años se ha visto un incremento de casos de enfermedades de las vías respiratorias, ya sea por la calidad del aire de las zonas urbanas a consecuencia de la industrialización, los hábitos personales y el tabaquismo.

El cáncer de pulmón tuvo una incidencia, en el 2020, de más de **2 millones de casos** en el mundo y alrededor de **1.8 millones de muertes** por esta causa. En **México**, se registraron **7 mil 588** casos nuevos y **7 mil 100** muertes por cáncer de pulmón (International Agency for Research on Cancer, 2020).

La resección pulmonar es la extirpación quirúrgica de todo o parte del pulmón debido a un cáncer de pulmón u otra enfermedad pulmonar. Por tal motivo, es importante monitorear el éxito de estas cirugías y clasificar a los pacientes con mayor riesgo de fallo, con el fin de servir de apoyo al momento de elegir el tratamiento adecuado para cada paciente según sus características.

CONJUNTO DE DATOS

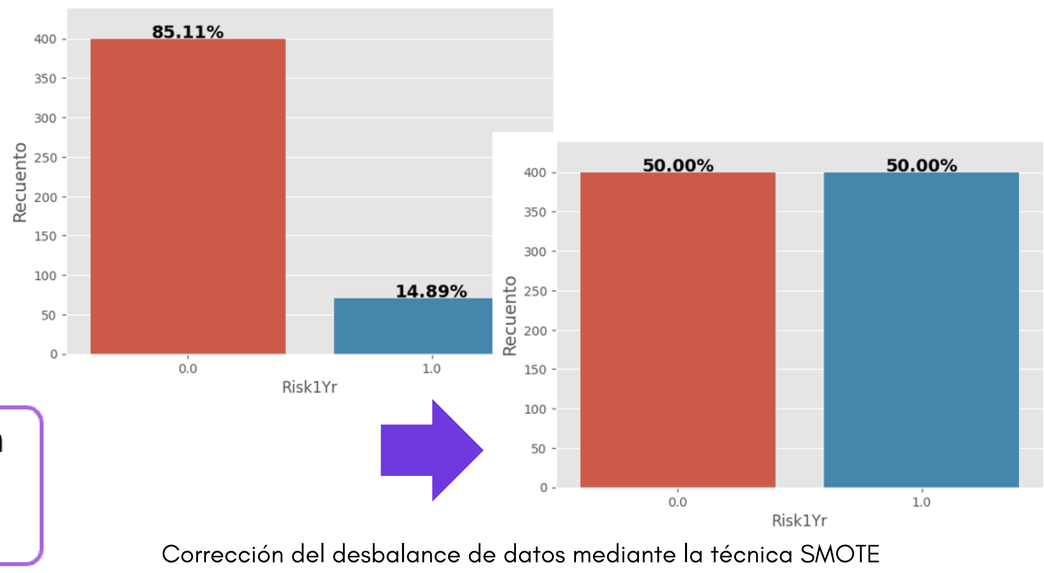
La base de datos está disponible en el sitio de UCI Machine Learning Repository (Lubicz, 2013), y contiene datos que se recopilaron en el Centro de Cirugía Torácica de Wroclaw, Polonia, para pacientes que se sometieron a resecciones pulmonares en los años 2007-2011.
Consta de 470 observaciones y 17 variables.

METODOLOGÍA

Preprocesamiento
- Normalización de variables numéricas.
- Transformación de variables categóricas.
- Sobremuestreo usando técnica SMOTE.

Modelos de aprendizaje automático
- Árboles de decisión
- Bosque aleatorio
- Máquinas de Vectores de Soporte

Evaluación de los modelos
- Accuracy
- F1 Score
- Precision
- Recall



Corrección del desbalance de datos mediante la técnica SMOTE

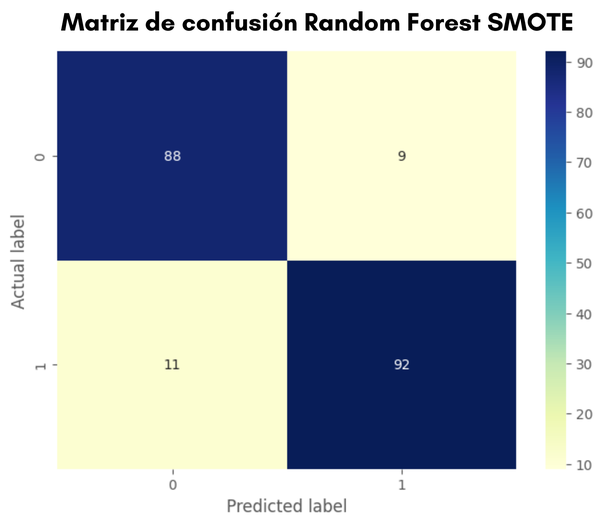
RESULTADOS PRELIMINARES

Se evaluaron los 3 modelos utilizando los datos originales y posteriormente los balanceados con la técnica SMOTE.

El modelo con mejor resultado de la métrica F1- score fue en de Random Forest, obteniendo un 90%.

Tabla 1. Resultados F1- Score

Modelo	Sin SMOTE	Con SMOTE
Decision Trees	24%	81%
Random Forest	0%	90%
SVM	39%	78%



CONCLUSIONES

- Se puede observar una mejora en el desempeño del modelo después de utilizar la técnica SMOTE. Esta mejora aplica para ambas clases, lo cual ayuda para el problema que se está estudiando.
- Al trabajar con datos desbalanceados, es importante conocer las distintas opciones de muestreo y métricas de evaluación, para elegir las que mejor se adapten a nuestro problema y necesidad. En este caso se decidió por una técnica de sobremuestreo debido a la cantidad de observaciones y la métrica F1-score, al tratarse de datos desbalanceados.
- Como trabajo a futuro, se tiene contemplado seguir explorando los hiperparámetros de los modelos sensibles a los costos, para tener la opción de utilizar los datos originales.

REFERENCIAS

- International Agency for Research on Cancer. (2020). Cancer Today. Obtenido de Cancer Today: https://gco.iarc.fr/today
- Lubicz, M. (2013). Thoracic Surgery Data. Obtenido de UC Irvine Machine Learning Repository: https://archive.ics.uci.edu/dataset/277/thoracic+surgery+data
- Chawla, N. V. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321-357.
- Brownlee, J. (2021). Imbalanced Classification with Python.