



21°

**CONGRESO
INTERNACIONAL CNB
COLEGIO NACIONAL DE BACTERIOLOGÍA**

INTERCULTURALIDAD • INCLUSIÓN • DESARROLLO SOSTENIBLE

M E D E L L Í N

2023

www.congresocnb.com



Aplicación de la inteligencia artificial y el aprendizaje automático en los datos clínicos, para extraer y aprovechar su información, logrando mejorar diagnósticos y resultados

Guido Marcelo Rincón Arévalo
Bact UIS-Esp-MSc-PhD



Contenido

- Introducción
 - Estado del Arte
 - Objetivos
 - Metodología
 - Técnicas de ML
 - Uso de datos*
- Entorno de desarrollo integrado
 - Resultados de los modelos *
 - Cálculos y métricas
 - Resultados
 - Conclusiones y Recomendaciones



Introducción

- **Machine Learning**: Disciplina de la IA que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones.
- **Algoritmo no supervisado**: Trata los objetos de entrada como un conjunto de variables aleatorias o conjunto no etiquetado.
- **Algoritmo supervisado**: Intenta expresar matemáticamente la dependencia entre entradas y salidas, deduciendo una función a partir de datos.
- **Variable dependiente** no está restringida a dos categorías y busca predecir el riesgo.
- **Regresión Logística**: Técnica de análisis de datos que utiliza las matemáticas para encontrar relaciones con los datos en categorías: A, M, B / binario.
- **Python**: Lenguaje de programación utilizado en la ciencia de datos.



Estado del arte

Tabla 1: Estudios recientes en sector salud

Autor	Año	Título	Conclusión
Kohli M, Prevedello LM, Filice RW, Geis JR	2018	Implementing Machine Learning in Radiology Practice and Research	El éxito de los algoritmos requerirá la participación de los profesionales en los próximos años, lo que generará participación en lugar de reemplazo
Uribe CF, Mathotaarachchi S, Gaudet V, Smith KC, Rosa-Neto P, Bénard F, Black SE, Zukotynski K	2019	Machine Learning in Nuclear Medicine	Los algoritmos en medicina pueden ser útiles y están empezando a ayudar en los diagnósticos
Zukotynski K, Gaudet V, Uribe CF, Mathotaarachchi S, Smith KC, Rosa-Neto P, Bénard F, Black SE	2021	Machine Learning in Medicine	Las nuevas familias de algoritmos llevan a evidente evolución en el campo. BD con imágenes compartidas son útiles, "evaluar": Ético, Privado, Propiedad de datos
Caballé I, Buño A, Bernabeu FA, Canalias F, Moreno A, Ibarz M, Puzo J, González C, González Á	2023	Estado actual y retos futuros de la medicina de laboratorio en España: un análisis de la Sociedad Española de Medicina de Laboratorio	Formación de profesionales clínicos en innovación tecnológica y aplicación del Big data,

<https://pubmed.ncbi.nlm.nih.gov/32978286/>
<https://pubmed.ncbi.nlm.nih.gov/30733322/>
<https://pubmed.ncbi.nlm.nih.gov/28125274/>



Objetivos

Realizar tareas mediante el análisis y el procesamiento de datos para predecir riesgos de enfermedad utilizando técnicas de *Machine Learning*.

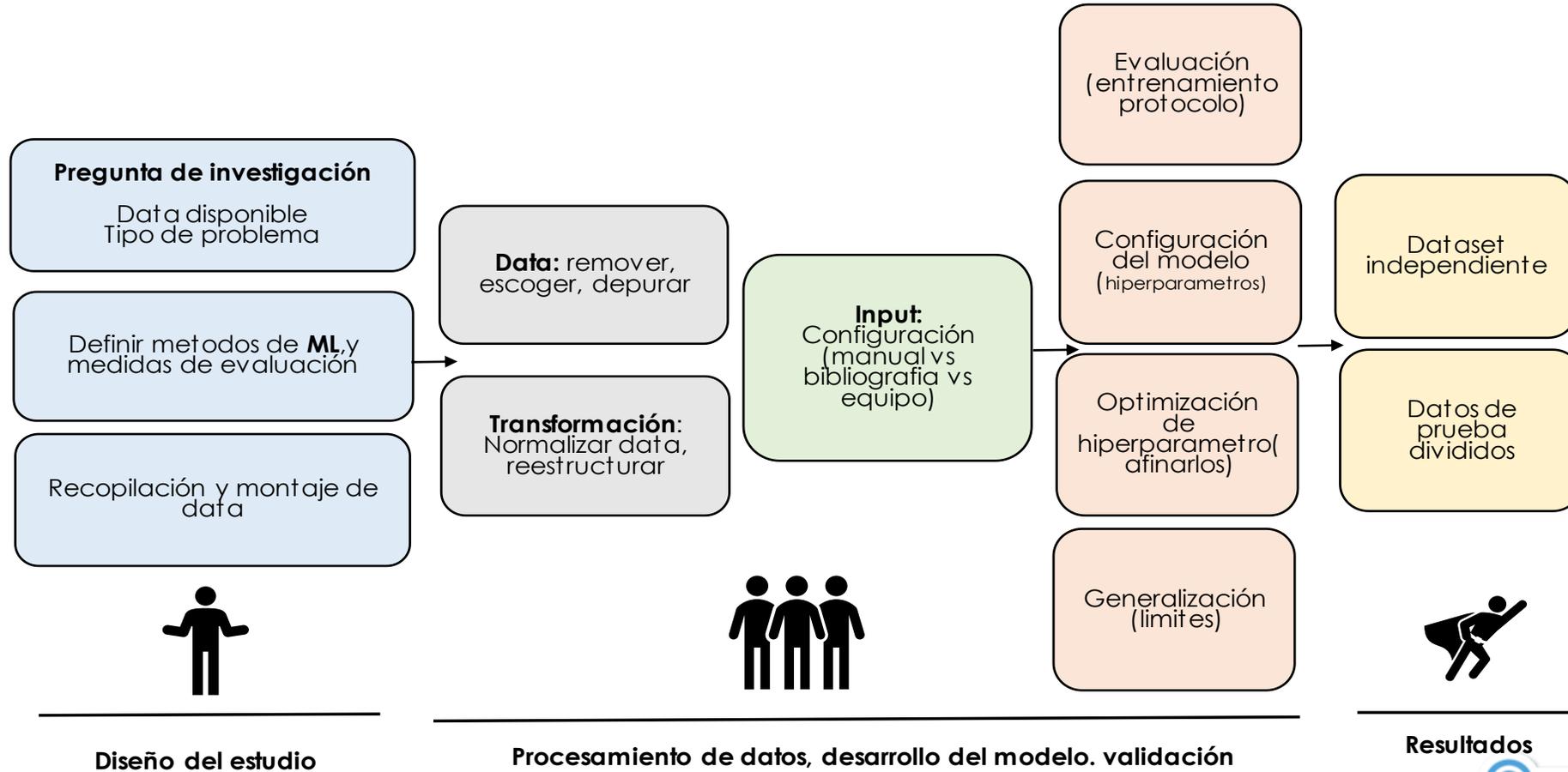
Determinar las técnicas más adecuadas para el análisis predictivo diagnóstico a partir de los inputs definidos en la clínica y de un problema de investigación.

Emplear resultados históricos de laboratorio hechos a pacientes, previo diligenciamiento del consentimiento informado para realizar pruebas y entrenamiento teniendo en cuenta los principios éticos en investigación.



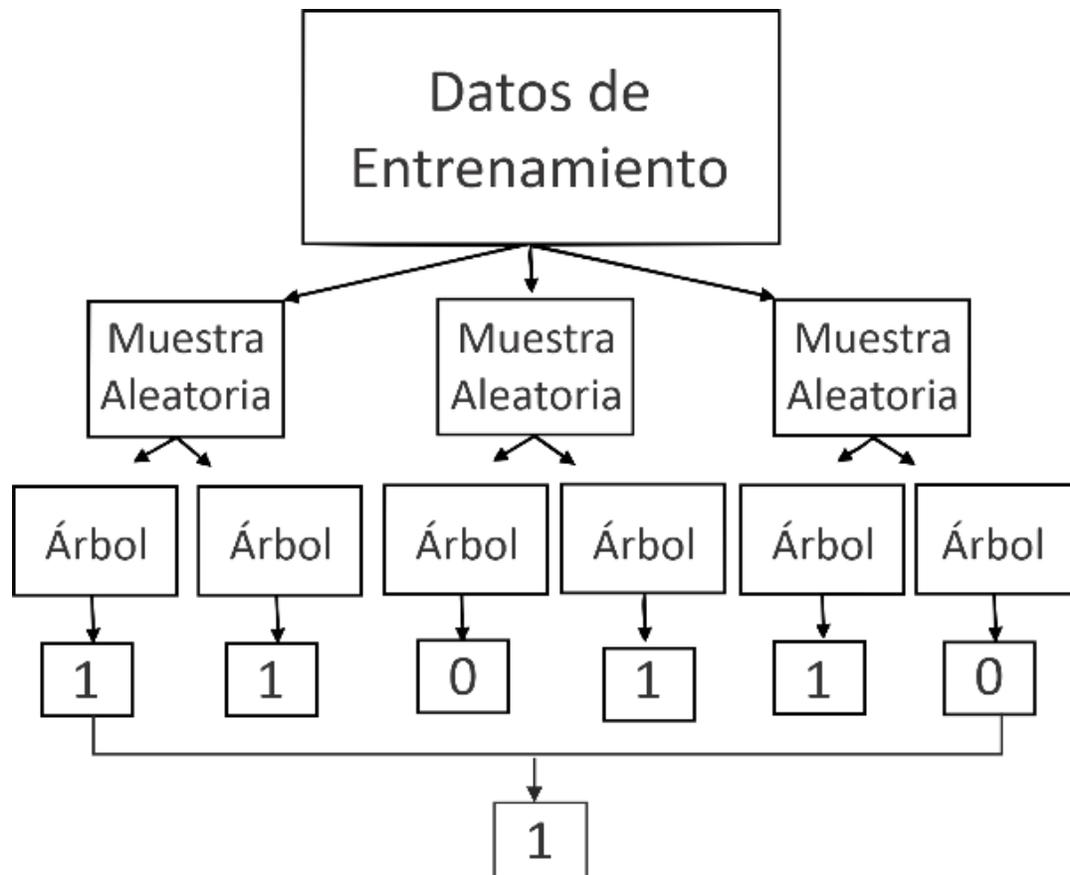
Metodología ML

Figura 1: Descripción general del flujo de trabajo



Técnicas de Machine Learning

Figura 4: Random Forest



Medidas de selección:

Criterios para calcular valores para cada atributo. Los valores se ordenan y los atributos se colocan en el árbol, el atributo con un valor alto se coloca en la raíz.

•Ganancia de información

Medir la aleatoriedad o la incertidumbre de una variable aleatoria, X se define mediante Entropía

•Índice de Gini

Métrica para medir la frecuencia con la que un elemento elegido al azar se identificaría incorrectamente

Uso de datos históricos*

Tabla 2: Datos iniciales

CODIGO	APELLIDOS Y NOMBRES	GENERO	OCUPACIÓN / CARGO	HOSPITAL	FECHA DE INGRESO AL SERVICIO	UBICACIÓN DEL DOSIMETRO	ENERO					FEBRERO					PERIODO DE USO DEL DOSIMETRO		
							PERIODO DE USO DEL DOSIMETRO	DOSIS DEL PERIODO (msv) Hp(10)	DOSIS DEL PERIODO (mSV) Hp(0.07)	DOSIS DEL PERIODO (mSV) Hp(3)	DOSIS ACUMULAD A DESDE EL INGRESO (mSV) Hp(10)	DOSIS ACUMULAD A 31/12/2018 (mSV) Hp(10)	PERIODO DE USO DEL DOSIMETRO	DOSIS DEL PERIODO (msv) Hp(10)	DOSIS DEL PERIODO (mSV) Hp(0.07)	DOSIS DEL PERIODO (mSV) Hp(3)		DOSIS ACUMULAD A DESDE EL INGRESO (mSV) Hp(10)	DOSIS ACUMULAD A 31/12/2018 (mSV) Hp(10)
16316		M	TECNOLOGO EN IMÁGENES DX	HLP	jun-16	TORAX	1 AL 31	0,13	0,16	0,16	0,53	0,4	1 AL 28	0,12	0,12	0,12	0,65	0,4	1 AL 31
16466		F	HIGIENISTA ORAL	HLP	ago-16	TORAX	1 AL 31	0,22	0,51	0,22	0,38	0,16	1 AL 28	0,03	0,07	0,05	0,41	0,16	1 AL 31
16470		M	ODONTOLOGO	HLP	ago-16	TORAX	1 AL 31	0,38	0,4	0,39	0,38	0	1 AL 28	ND	ND	ND	0,38	0	1 AL 31
17062		M	TECNOLOGO EN IMÁGENES DX	HLP	ene-17	TORAX	1 AL 31	0,19	0,32	0,19	0,4	0,21	1 AL 28	0,14	0,38	0,14	0,54	0,21	1 AL 31
17186		M	TECNOLOGO EN IMÁGENES DX	HLP	feb-17	TORAX	1 AL 31	0,31	0,31	0,31	0,41	0,1	1 AL 28	0,17	0,17	0,17	0,58	0,1	1 AL 31
17357		M	ODONTOLOGO	HLP	mar-17	TORAX	1 AL 31	ND	ND	ND	0,16	0,16	1 AL 28	ND	ND	ND	0,16	0,16	1 AL 31
17441		F	MEDICO	HLP	may-17	TORAX	1 AL 31	0,17	0,22	0,22	0,32	0,15	1 AL 28	0,26	0,28	0,27	0,58	0,15	1 AL 31
17714		M	TECNOLOGO EN IMÁGENES DX	HLP	jul-17	TORAX	1 AL 31	0,32	0,32	0,32	0,81	0,49	1 AL 28	0,06	0,21	0,06	0,87	0,49	1 AL 31
18329		F	ODONTOLOGO	HLP	oct-17	TORAX	1 AL 31	0,11	0,11	0,11	0,11	0	1 AL 28	ND	ND	ND	0,11	0	1 AL 31
20153		F	MEDICO	HLP	dic-18	TORAX	1 AL 31	0,36	0,36	0,36	0,36	0	1 AL 28	0,08	0,08	0,08	0,44	0	1 AL 31
20183		F	HIGIENISTA ORAL	HLP	dic-18	TORAX	1 AL 31	0,09	0,1	0,1	0,09	0	1 AL 28	0,24	0,24	0,24	0,33	0	1 AL 31



Uso de datos históricos

Tabla 3: Formato CSV

- Eliminación de columnas
- Transformar
- Ingreso al IDE (Integrated Development Environment)
- Spyder
- Uso de bibliotecas

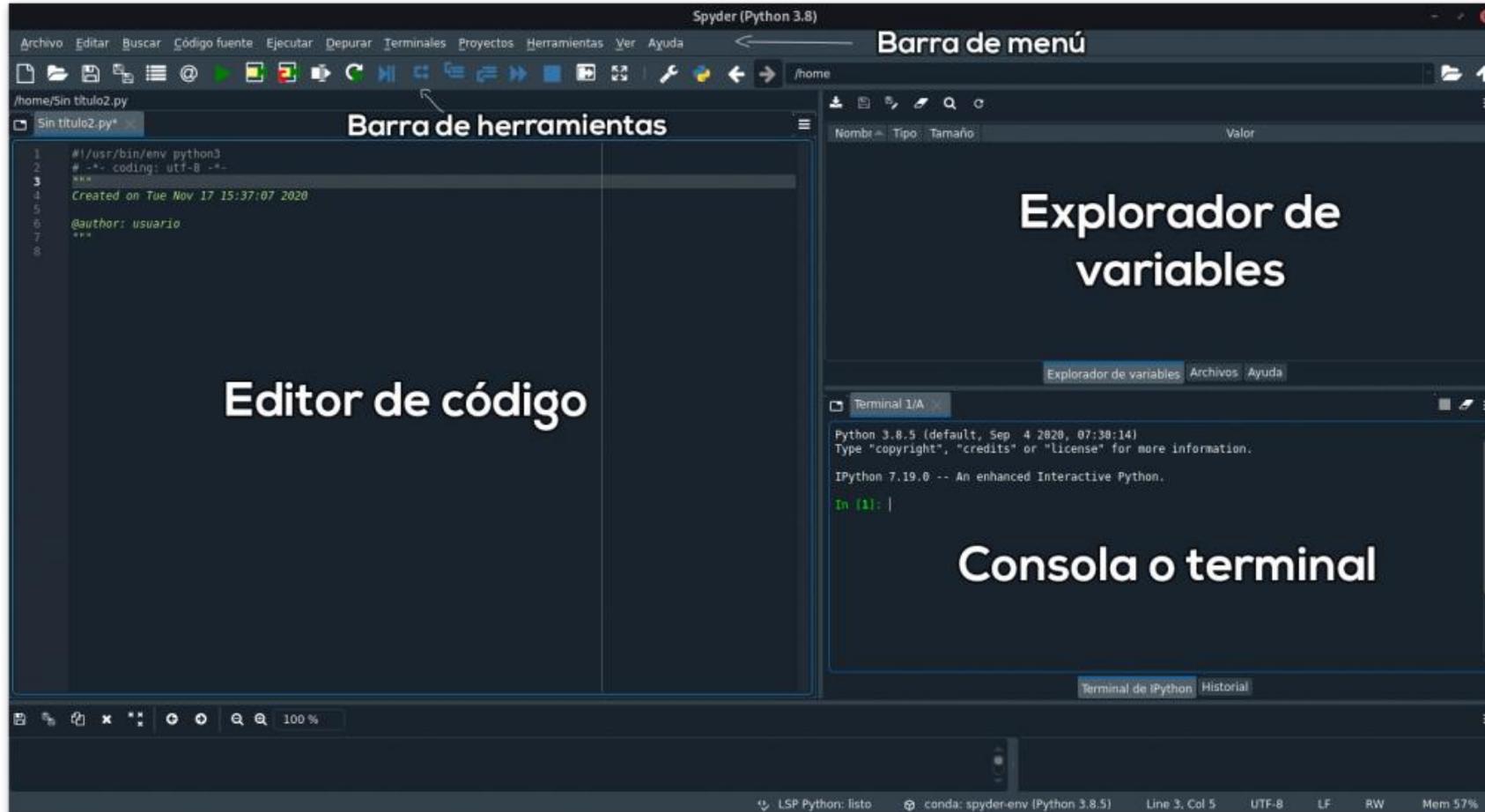
1	codigo,year_month,class,06_data_dosis,05_data_dosis,04_data_dosis,03_
2	16316,201907,0,0.13,0.12,0.26,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
3	16466,201907,0,0.22,0.03,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
4	16470,201907,0,0.38,0.1,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
5	17062,201907,0,0.19,0.14,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
6	17186,201907,0,0.31,0.17,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
7	17357,201907,0,0.1,0.1,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01
8	17441,201907,0,0.17,0.26,0.25,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
9	17714,201907,0,0.32,0.06,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
10	18329,201907,0,0.11,0.1,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
11	20153,201907,0,0.36,0.08,0.08,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
12	20183,201907,0,0.09,0.24,0.17,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.0
13	16316,201908,0,0.12,0.26,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
14	16466,201908,0,0.03,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
15	16470,201908,0,0.1,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
16	17062,201908,0,0.14,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
17	17186,201908,0,0.17,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
18	17357,201908,0,0.1,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0
19	17441,201908,0,0.26,0.25,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.17,0.01,0.0



Entorno de desarrollo integrado

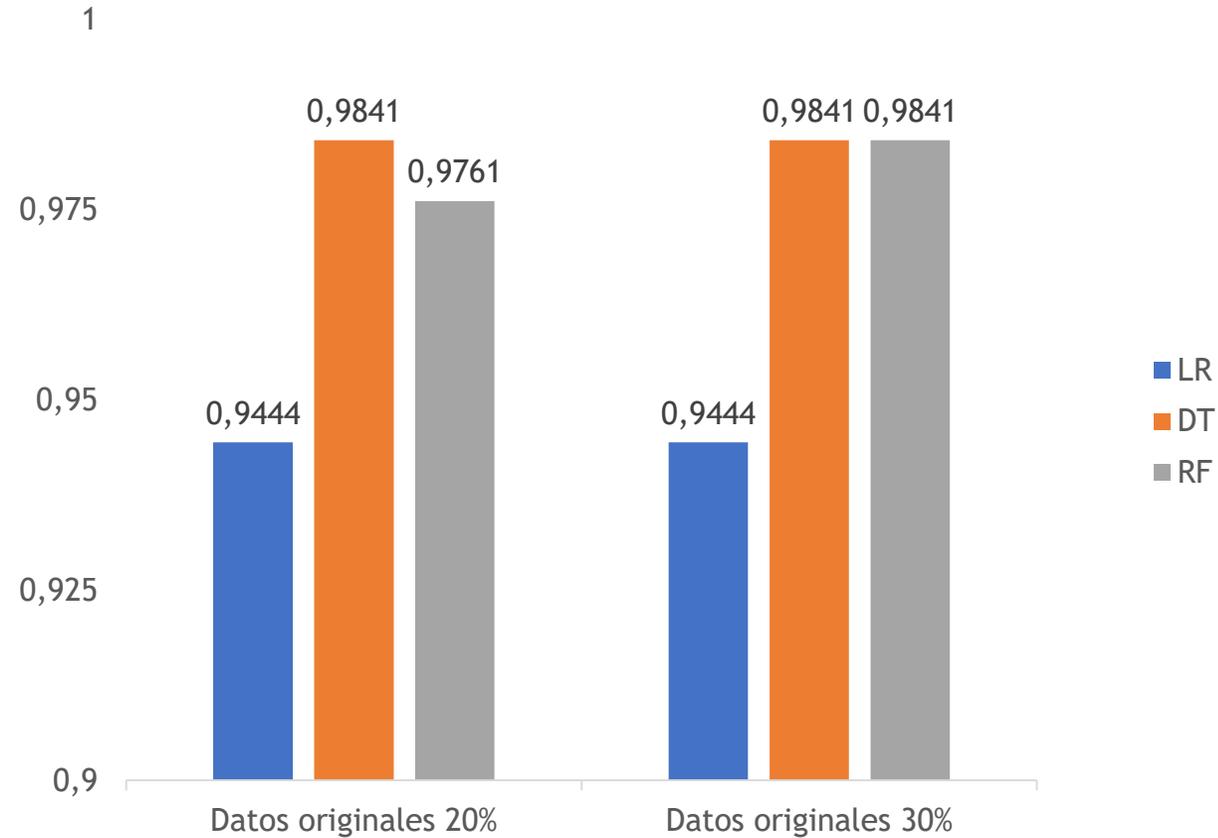


Gráfico 1: Spyder entorno de desarrollo



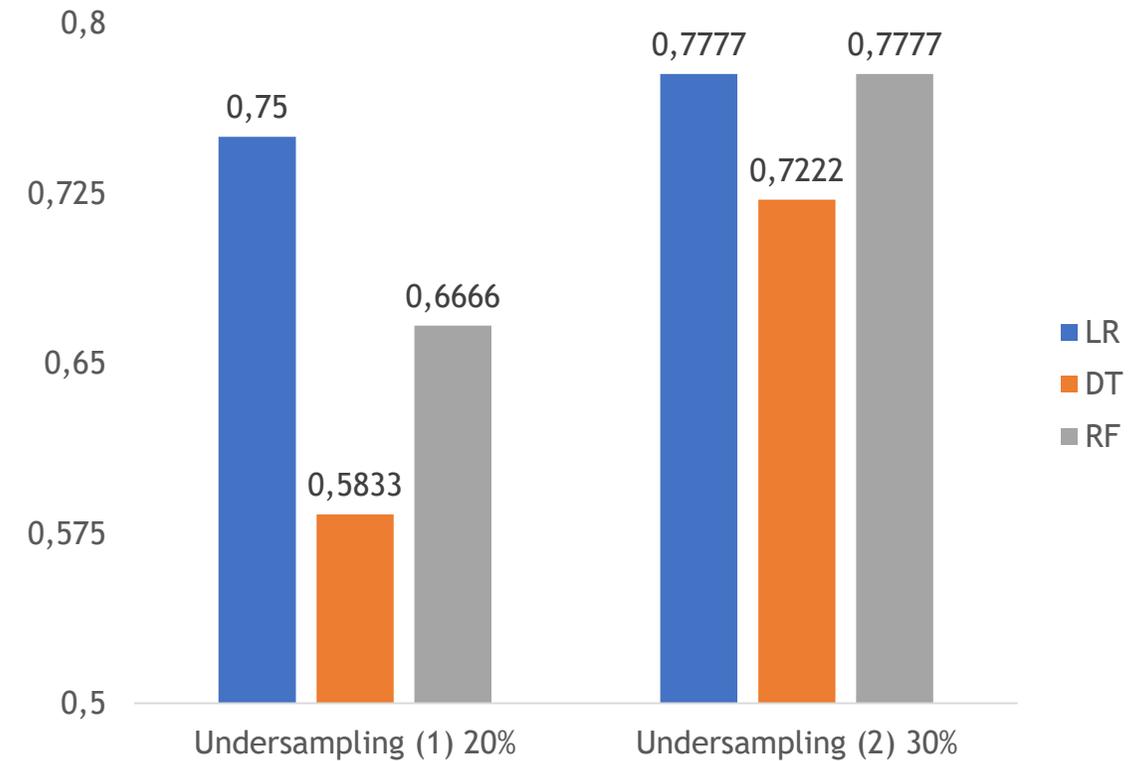
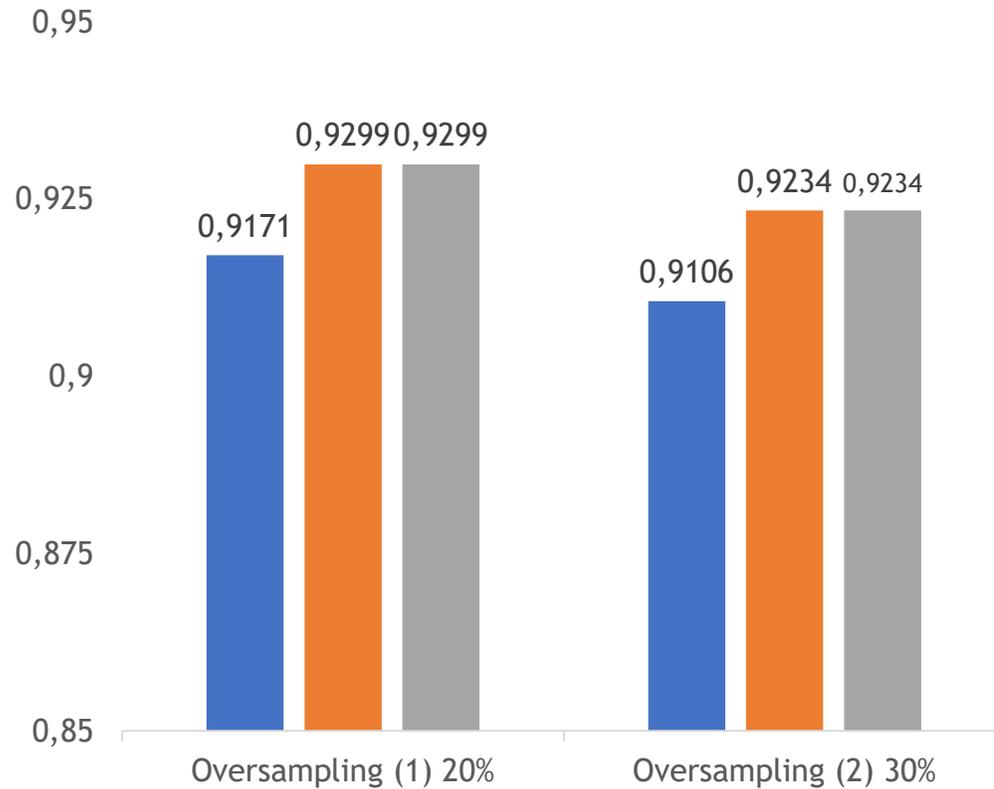
Resultados de los modelos empleados

Gráfico 2: Resultados con datos originales 20% y 30% de test



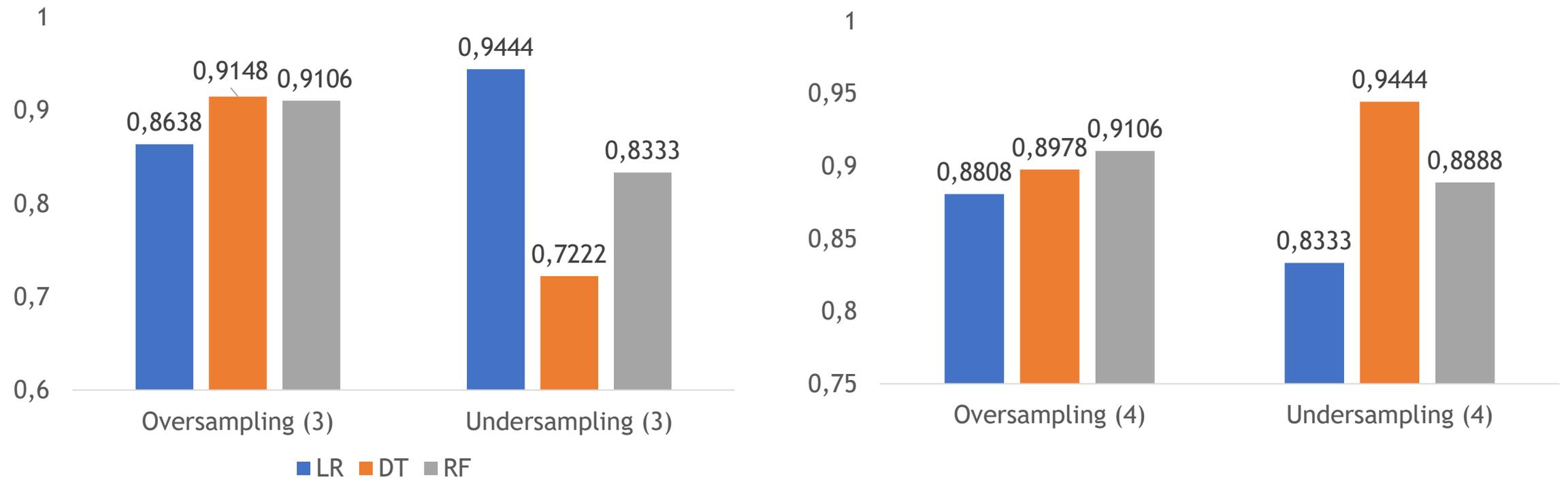
Resultados de los modelos empleados

Gráfico 3: Resultados con Oversampling y Undersampling



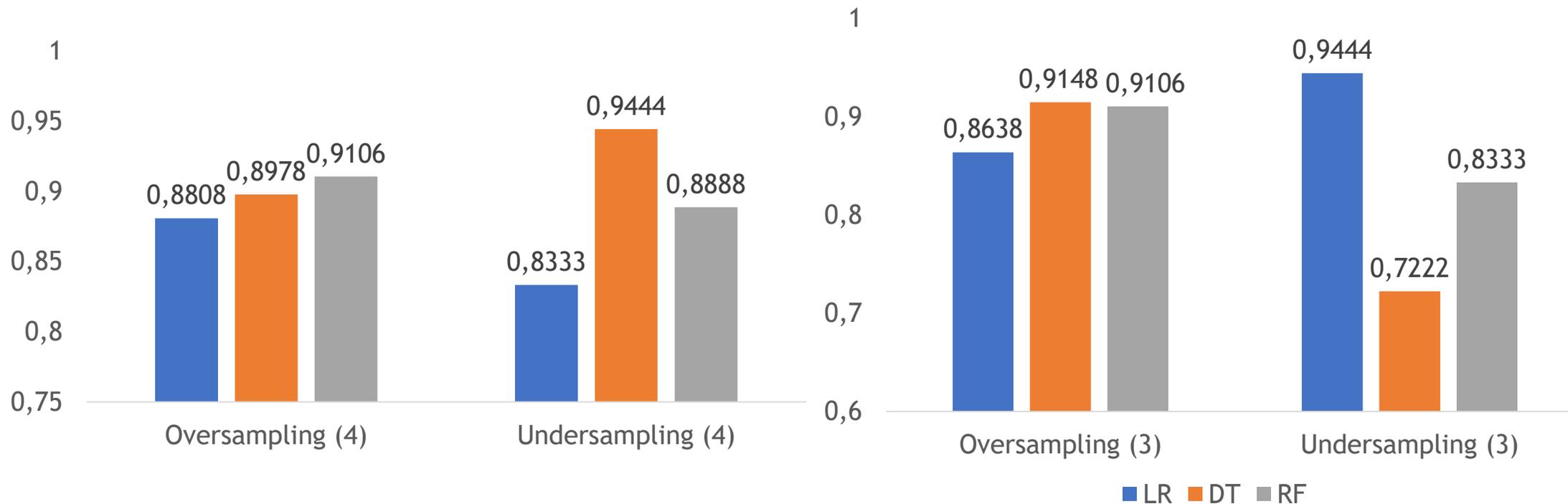
Resultados de los modelos empleados

Gráfico 4: Resultados con Oversampling 3 y Undersampling 3 (30 % Random_state 1000)



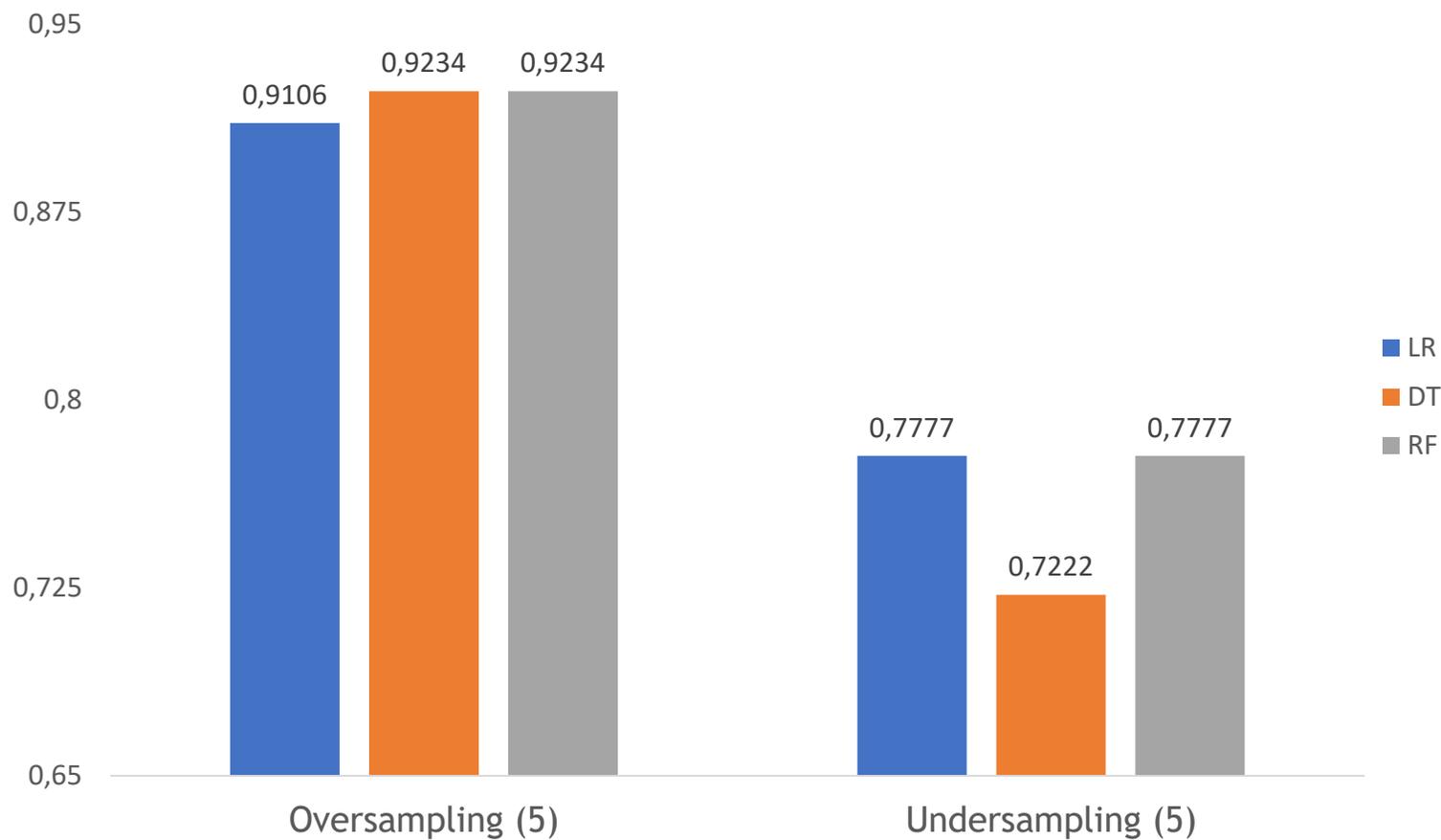
Resultados de los modelos empleados

Gráfico 5: Resultados con Oversampling 4 y Undersampling 4 (30 % Random_state 100)



Resultados de los modelos empleados

Gráfico 6: Resultados con Oversampling5 y Undersampling5 (30 % Random_state 10)



Cálculo de métricas de desempeño

Gráfico 7: Matriz de confusión



Métricas calculadas

PRECISION:
$$\frac{VP}{VP + FP}$$

RECALL:
$$\frac{VP}{VP + FN}$$

F1:
$$\frac{2 VP}{2 VP + FP + FN}$$



Métricas de rendimiento

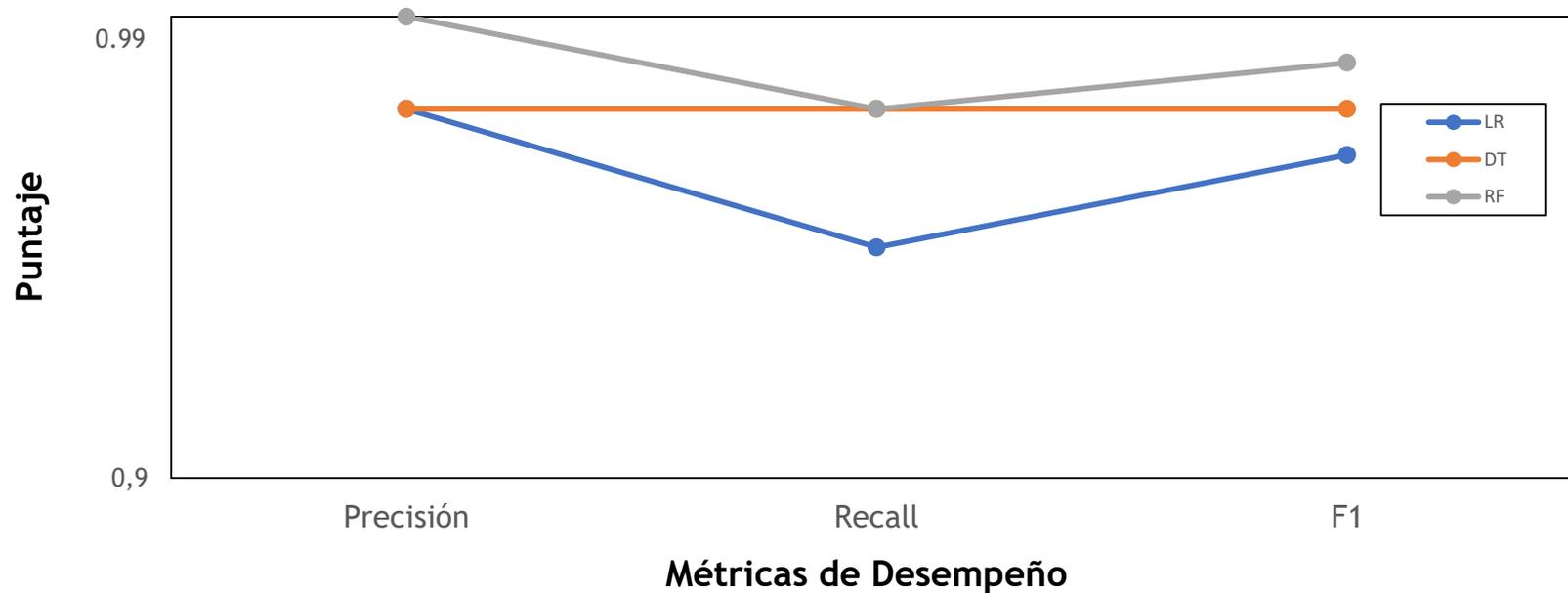
Tabla 4: Modelos y métricas de desempeño

Modelo	Precision	Recall	F1
Regresión logística	0.98	0.95	0.97
Árbol de decisión	0.98	0.98	0.98
<i>Random forest</i>	1.0	0.98	0.99



Métricas de rendimiento

Gráfico 8: Modelos y métricas de desempeño



Regresión logística (RL):

Fue el algoritmo que **menor** rendimiento mostro según los resultados obtenidos con el 20 y 30 % de test.

Árbol de decisión (DT):

Se obtuvo el 98% de clasificación correcta con su uso según las métricas empleadas. El comportamiento óptimo del modelo fue con el 30% de test y un entrenamiento 70% de los datos originales.

Random forest (RF):

Dio los mejores resultados según las métricas empleadas

Precisión: tuvo una precisión perfecta.

Recall: tasa de verdaderos positivos (98%). La capacidad del algoritmo para detectar el riesgo.

F1: Compara (rendimiento combinado) la precisión respecto al recall, en el modelo de RF fue de 0.99
 $2 * (\text{Precision} * \text{Recall}) / (\text{Precisión} + \text{Recall})$



Conclusiones

Las predicciones realizadas con los datos históricos generan utilidad diagnóstica en el sector de la salud, siendo alineados con normas legales y confidenciales, que se logran empleando información individual del paciente, sin embargo, su uso requiere mantener criterios asociados con los principios básicos de la bioética y manejo confidencial de los datos.

Los ejercicios académicos se pueden llevar a lo práctico: ARL y se lleva una vigilancia real de la exposición con el uso del dato



Recomendaciones

ML presenta ventajas importantes en términos de rendimiento predictivo y de identificación de patrones no descubiertos en pacientes con diagnósticos específicos, entonces, se recomienda conocer la estructura del dato...

A pesar de esta popularidad, y de encontrar técnicas adecuadas muchos clínicos e investigadores aún no están familiarizados con la evaluación e interpretación de los análisis de ML obteniendo los inputs y resultados de los mismos procesos clínicos

El uso de históricos de mediciones de laboratorio hechas a pacientes es un buen inicio para normalizar los datos, ya sea como custodia y/o como manejo predictor de riesgos aún más cercano



Referencias

- Recommendations for Reporting Machine Learning Analyses in Clinical Research. Available in: <https://pubmed.ncbi.nlm.nih.gov/33079589/>
- The Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. Available in: <https://www.equator-network.org/reporting-guidelines/record/>
- Prediction of Diabetic Patient Readmission Using Machine Learning. Available in: https://link.springer.com/chapter/10.1007/978-3-030-36211-9_7
- Estado actual y retos futuros de la medicina de laboratorio en España: un análisis de la Sociedad Española de Medicina de Laboratorio. Available in: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10197193/>
- Universidad Antonio Nariño. Herramienta para predecir el riesgo de exposición a radiaciones ionizantes en trabajadores del sector salud utilizando técnicas de Machine Learning. Disponible en: http://repositorio.uan.edu.co:8080/bitstream/123456789/6616/1/2021_GuidoMarceloRinc%C3%B3nAr%C3%A9valo.pdf
- Ley 226 de 1996: Principios bioéticos. Disponible en: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/IGUB/ley-266-de-1996.pdf>
- Ley 23 de 1981 Ética médica. Disponible en: https://www.redjurista.com/Documents/ley_23_de_1981_congreso_de_la_republica.aspx#/
- Resolución 8430 de 1993 Riesgo mínimo. Disponible en: https://www.redjurista.com/Documents/resolucion_8430_de_1993.aspx#/
- Linear regression implementation in Python. Available in: <https://dataaspirant.com/linear-regression-implementation-in-python/>
- RandomOverSampler. Available in: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html
- Sklearn.utils.resamples. Available in: <https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html>
- Oversampling methods. Available in: https://imbalanced-learn.org/stable/references/over_sampling.html
- Rincon G. Available in: <https://www.hindawi.com/journals/bmri/2021/2322257/>
- Extension of the Hepar II. Model to multiple disorder diagnosis. Available in: <https://www.pitt.edu/~druzdzal/psfiles/springer00.pdf>





21°

**CONGRESO
INTERNACIONAL CNB**
COLEGIO NACIONAL DE BACTERIOLOGÍA

INTERCULTURALIDAD • INCLUSIÓN • DESARROLLO SOSTENIBLE

MEDELLÍN
2023

¡Gracias!

guido.rincon@correo.uis.edu.co
grincon46@uan.edu.co

