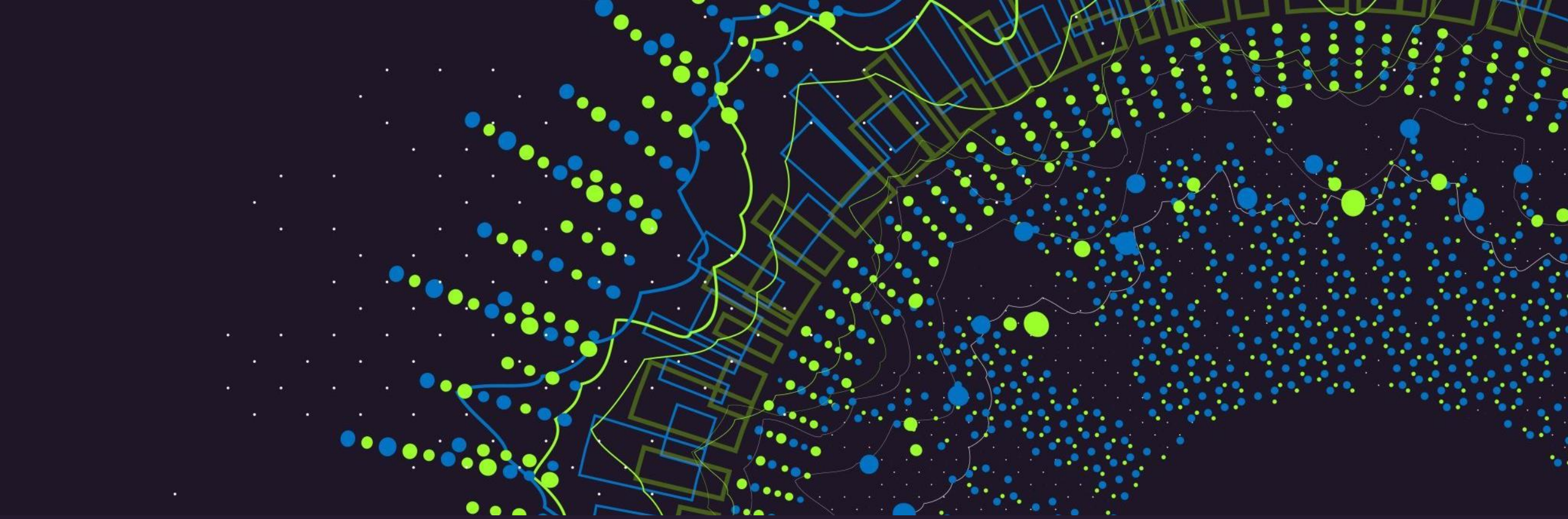


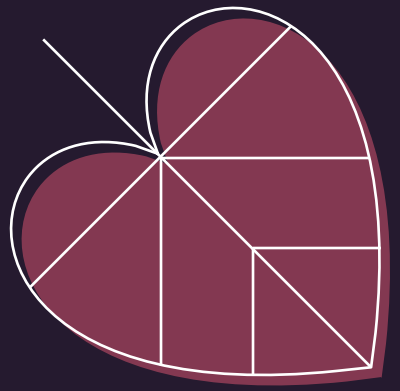
PROYECTO 1 – CLASIFICACIÓN DE PACIENTES ELEGIBLES PARA PRUEBAS DE CÁNCER

Juan Sebastián Ramírez 201923800

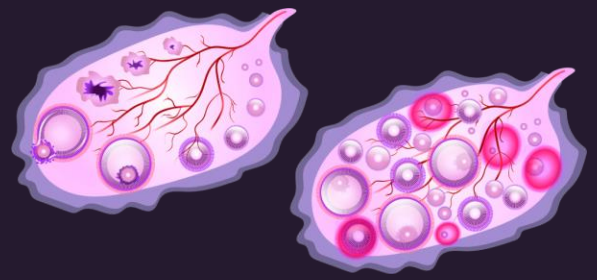
Andrés Santiago Triana 201923265

Gabriela García 201912531

OPORTUNIDAD/ PROBLEMA NEGOCIO



- El problema principal es que muchas veces los doctores no tienen mucho tiempo disponible para dedicar a cada paciente,
- La cantidad de expertos es mucho menor que en los países desarrollados.
- Los ensayos clínicos de cáncer pueden ser bastante costosos.
- Muchas veces los pacientes corren riesgos al hacer dichos exámenes, sobre todo los que requieren cualquier tipo de radiación como lo son los rayos x.
- Sería de gran utilidad que los doctores tuviesen un diagnóstico inicial de si un paciente pudiese requerir ensayos clínicos para cáncer



DESCRIPCIÓN DEL REQUERIMIENTO DESDE EL PUNTO DE VISTA DE APRENDIZAJE DE MÁQUINA

- Análisis de lenguaje, dado que la información que se tiene únicamente son dos columnas:

1. *Una columna con un label que dice 1 si el paciente no requiere ningún tipo de ensayo clínico*
2. *Una columna con dos frases separadas con un punto, la que va antes del punto indica el estudio realizado, mientras que la que va después tiene una descripción del tipo de cáncer que podría tener.*

- *La máquina debe hacer análisis del lenguaje con las dos frases que se le dan.*
- Entender si el estudio puede implicar que el paciente pueda requerir de los ensayos clínicos.

COMPRENSIÓN DEL NEGOCIO

- Modelo de salud para enfermedades complejas.
- Predicciones acerca de si un paciente requiere un ensayo clínico o no.
- Ahorro de dinero en la realización de exámenes.
- Priorizar pacientes sobre otros, para la detección del cáncer de manera más temprana.

OBJETIVOS

PRINCIPAL

- Apoyar las decisiones de los médicos acerca de si un paciente va a requerir ensayos clínicos para cáncer.

SECUNDARIOS

- Mejorar la comprensión acerca de qué palabras en el lenguaje son las que puede implicar que se requieran hacer ensayos clínicos o no.
- Proveer un modelo relativamente veloz, dado que lo ideal sería permitir reducciones en los tiempos que un doctor debe tomar para poder saber si un paciente va a requerir un ensayo clínico.
- Tener porcentajes de acierto relativamente altos, dado que es un tema de salud, los errores implican afectación en muchas personas.

TIPOS Y TAREAS DE APRENDIZAJE

Tipo aprendizaje	Tarea de aprendizaje	Algoritmo e hiperparámetros utilizados (con la justificación respectiva)
Supervisada	Árboles de decisión (clasificación)	DecisionTreeClassifier, entropía (Explicación y detalles en notebook)
Supervisada	Redes neuronales (clasificación)	MLPClassifier(Explicación y detalles en el notebook)
Supervisada	KNN (clasificación)	KNeighborsClassifier, n_neighbors=49(Explicación y detalles en notebook)

TRANSFORMACIÓN DE DATOS

- Tokenizar las frases en palabras, de manera que se pueda tener una lista de palabras que se usan en cada uno de los registros.
- Corregir contracciones que son usadas en el ingles, para que estas sean separadas en las palabras que significan "realmente". De esta forma una palabra como "it's" se va a reemplazar por "it is".
- Eliminación en el ruido de las palabras (mayúsculas, puntuaciones, caracteres no ASCII, números, *stopwords*, etc.)
- Finalmente se realizará una estandarización de las palabras, eliminado los prefijos y sufijos (stems) y la conjugación de los verbos (lemmatize)

ANÁLISIS DE RESULTADOS

- Resultados árbol de decisión

	precision	recall	f1-score	support
0	0.80	0.76	0.78	1245
1	0.75	0.80	0.78	1153
accuracy			0.78	2398
macro avg	0.78	0.78	0.78	2398
weighted avg	0.78	0.78	0.78	2398

ANÁLISIS DE RESULTADOS

- Resultado de redes neuronales

	precision	recall	f1-score	support
0	0.83	0.80	0.82	1245
1	0.80	0.82	0.81	1153
accuracy			0.81	2398
macro avg	0.81	0.81	0.81	2398
weighted avg	0.81	0.81	0.81	2398

ANÁLISIS DE RESULTADOS

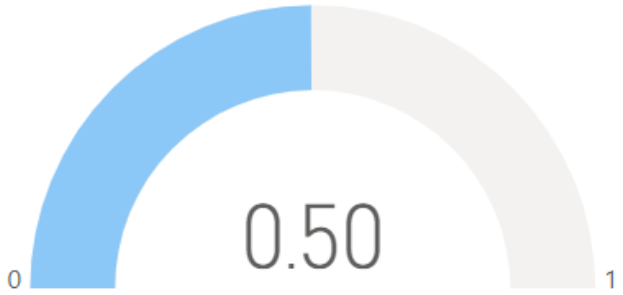
- Resultado de KNN

	precision	recall	f1-score	support
0	0.86	0.09	0.17	1245
1	0.50	0.98	0.66	1153
accuracy			0.52	2398
macro avg	0.68	0.54	0.42	2398
weighted avg	0.69	0.52	0.41	2398

ANÁLISIS DE RESULTADOS

ANÁLISIS DE RESULTADOS

Average of Real value, Min of Real value and Max of Real value



0.25

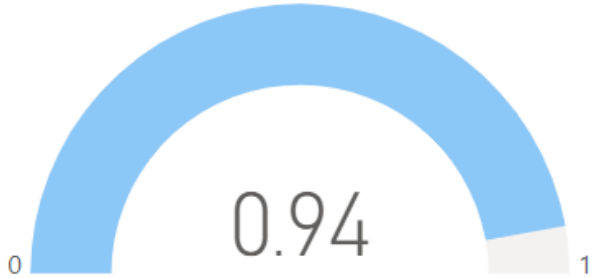
Variance of Real value

Average of Predicted value arbol, Min of Real value and Max of Real value

0.25

Variance of Predicted value arbol

Average of Predicted value knn, Min of Real value and Max of Real value



0.05

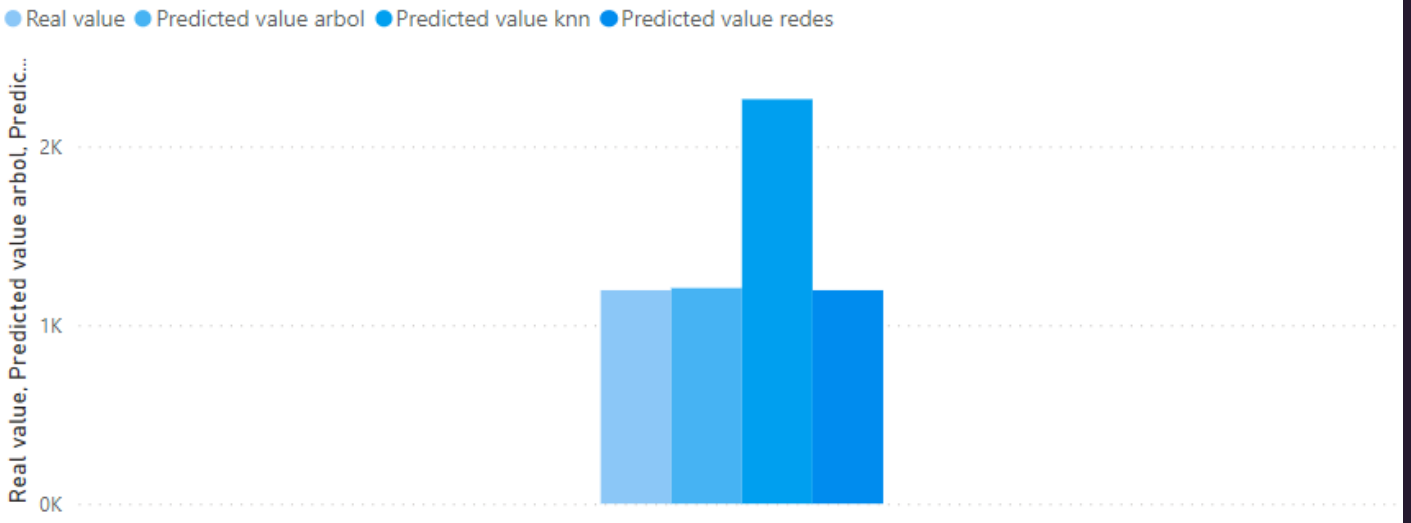
Variance of Predicted value knn

Average of Predicted value redes, Min of Real value and Max of Real value

0.25

Variance of Predicted value redes

Real value, Predicted value arbol, Predicted value knn and Predicted value redes



CONCLUSIONES

El algoritmo que obtuvo mejores resultados es el algoritmo de redes neuronales, a partir de los valores obtenidos en precisión, recall y f1-score se pudo evidenciar que realizo una. buena clasificación tanto para los pacientes elegibles como no elegibles. Como se puede ver en la matriz de confusión hay una mayor cantidad de aciertos (VP y FN). Los errores de predicción son menores