Proyecto #1

Apoyo al diagnóstico de pacientes: Identificación del problema del paciente a partir de una descripción general dada por un médico

Universidad de Los Andes

ISIS 3425 – Sistemas Empresariales

Integrantes:

Nicolás Segura Castro – Sección 2

Jaime Andrés Torres – Sección 2

David Ruiz Villota – Sección 2

Bogotá, 2021

1. Comprensión del negocio y enfoque analítico

- **Definición de los objetivos y criterios de éxito desde el punto de vista del negocio**: El objetivo principal de este proyecto está centrado en analizar el contenido de una serie de problemas médicos descritos en el idioma inglés para que, de acuerdo con las categorías que se describen en la fuente de datos, se pueda deducir con cierta probabilidad a qué problema corresponde un determinado texto. De esta manera, resulta pertinente: filtrar las descripciones de manera correcta para optimizar el análisis de los datos, adecuar la información de tal manera que se pueda utilizar para la construcción de modelos de entendimiento de lenguaje natural y utilizar dichos modelos para determinar a qué problema corresponde un determinado texto. Para este caso, teniendo en cuenta que el volumen de datos no resultaba considerablemente abundante; decidimos conservar todos los registros que consideramos relevantes, de esta manera, contamos con una muestra de 12000. En este caso resulta como criterio de éxito que el modelo pueda identificar a qué problema clínico corresponde el problema dado. Si el Fscore fue mayor a 0.5, el modelo se considerará exitoso.

- **Determinación de las tareas de analítica de textos que se consideran adecuadas**: Una vez tomada la muestra representativa se consideran pertinentes las siguientes tres etapas o tareas para poder realizar el pre-procesamiento de los datos:

- **Eliminación del ruido:** Se utiliza para dejar el archivo el texto plano y para eliminar caracteres especiales y dejar todo en minúsculas.
- **Preselección:** Como en el conjunto de datos propuesto para el ejercicio del proyecto se encontraban palabras no relacionadas con el lenguaje clínico, se utilizó la librería y los modelos de scispacy para detectar fácilmente qué palabras podían resultar útiles para el proyecto.
- **Tokenización:** Permite dividir frases u oraciones en palabras con el fin de desglosar las palabras correctamente para su posterior análisis.

- **Normalización:** En esta tarea se realiza la eliminación de los prefijos y sufijos además de la lematización de los verbos para únicamente conservar los términos relevantes para este caso.

- Descripción de relaciones entre los requerimientos de negocio y los requerimientos de aprendizaje de máquina propuestos:

Oportunidad/ problema Negocio	Identificar los sentimientos asociados
Descripción del requerimiento desde el punto de vista de aprendizaje de máquina	A partir de la implementación de modelos de análisis de lenguaje natural determinar el problema médico descrito en un texto dado a través de algoritmos de clasificación.

Detalles de la actividad de minería de datos

Tarea	Técnica	Algoritmo e hiper-parámetros
Clasificación	Árboles de decisión	`clf3 = DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None,max_features=None,max_leaf_nodes=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2 min_weight_fraction_leaf=0.0, random_state=None, splitter='best')`
Clasificación	Regresión logística	`clf = LogisticRegression(penalty= 'elasticnet', solver= 'saga' l1_ratio=0.5, random_state=1)`
Clasificación	Multi-layer Perceptron classifier	`clf2 = MLPClassifier(random_state=1, max_iter=300)`

2. Comprensión de los datos y preparación de los datos

Antes de empezar es necesario entender que dentro del conjunto de datos únicamente se encontraron dos columnas, una que representa los problemas descritos por el texto (problems_described) y otra con el texto (medical_abstracts).

Para empezar con la preparación de los datos se dividieron los datos de entrenamiento y los datos de prueba. Sobre dichos datos de entrenamiento se realizaron las tareas de tokenización, normalización, eliminación de ruido y registros vacíos, preselección y transformación de campos que fueron explicadas anteriormente para únicamente contar con información significativa para el modelo, todo con el fin de garantizar su calidad y no contar con datos poco útiles. Además de lo anterior, cabe recalcar que se hizo use de la técnica **SMOTE** debido a que podían ocurrer problemas de over y sub sampling que sesgaran el modelo.

3. Modelado y evaluación

Con el fin de validar la tarea de clasificación realizada por los modelos con base al conjunto de datos preseleccionado seleccionamos los siguientes algoritmos con el objetivo de identificar cuál de ellos nos podía presentar mejores resultados de acuerdo con el requerimiento principal de clasificar problemas de textos médicos:

- Regresión logística: Algoritmo de análisis de regresión utilizado para predecir el resultado de una variable categórica de acuerdo con un conjunto de variables independientes.

	precision	recall	f1-score	support
Neoplasms	0.68	0.67	0.67	657
General pathological conditions	0.45	0.57	0.50	998
Nervous system diseases	0.48	0.36	0.41	400
Cardiovascular diseases	0.67	0.61	0.64	634
Digestive system diseases	0.53	0.35	0.42	311
accuracy			0.55	3000
macro avg	0.56	0.51	0.53	3000
weighted avg	0.56	0.55	0.55	3000

- Árboles de decisión: Este algoritmo estadístico fue implementado con el hiperparámetro de *Gini* con el fin de lograr la clasificación de los problemas médicos. En este caso, la intención del modelo estaba centrada en la predicción de las categorías encontradas en los textos que, entrenados en el modelo arrojaron los siguientes resultados:

Predicted labels

	precision	recall	f1-score	support
Neoplasms	0.36	0.32	0.34	657
General pathological conditions	0.21	0.21	0.21	999
Nervous system diseases	0.16	0.15	0.16	400
Cardiovascular diseases	0.35	0.36	0.35	634
Digestive system diseases	0.57	0.63	0.60	998
accuracy			0.36	3688
macro avg	0.33	0.33	0.33	3688
weighted avg	0.35	0.36	0.35	3688

- MLPClassifier: El clasificador Perceptron multicapa se conecta a una red neuronal y se basa en la misma para poder realizar la clasificación. Este es uno de los algoritmos más populares dentro de la tarea anteriormente mencionada debido a la eficiencia que suelen ofrecer las redes neuronales:

	precision	recall	f1-score	support
Neoplasms	0.56	0.51	0.53	657
General pathological conditions	0.34	0.35	0.34	999
Nervous system diseases	0.34	0.30	0.32	400
Cardiovascular diseases	0.52	0.53	0.53	634
Digestive system diseases	0.79	0.85	0.82	998
accuracy			0.54	3688
macro avg	0.51	0.51	0.51	3688
weighted avg	0.53	0.54	0.53	3688

Una vez analizados los resultados decidimos utilizar como criterio de evaluación de desempeño del modelo la métrica de precisión y el f1-score, ya que, en el caso médico, resulta de **vital importancia** el contar con un modelo confiable y que tenga pocos falsos negativos. Partiendo de lo anterior encontramos que el mejor modelo resultaría ser el implementado a partir de **Regresión Logística**, ya que además de contar con el mejor puntaje de precisión, contaba con una buena calificación de recall. Antes de continuar, es importante tener en cuenta que esta decisión fue basada en que el modelo de **Regresión logística** resultó ser el más “equilibrado” entre los 3 modelos, esto quiere decir que para la clasificación de la mayoría de las categorías obtuvo un mayor puntaje tanto de precisión como un mejor f1-score. Lo anterior resulta relevante debido a que en el caso del modelo de **MLPClassifier**, aunque se obtuviera una precisión bastante alta para clasificar problemas relacionados con el sistema digestivo, los otros puntajes no fueron muy buenos en comparación con los obtenidos en el modelo de **Regresión Logística**. Respecto al modelo de **Árboles de Clasificación** no se puede decir mucho, ya que fue el que obtuvo peores resultados a pesar de ejecutarse tanto con **Gini** como con **Entropy**.

Analizando entonces el puntaje obtenido a través del mejor modelo (**Regresión logística**) llegamos a la conclusión de que, si bien era capaz de clasificar algunas categorías con cierta probabilidad de éxito interesante, esta clasificación resultó ser un poco ambigua e insuficiente para los objetivos propuestos. Es importante tener en cuenta que para este caso en particular resulta de suma importancia el hacer una clasificación lo más precisa posible con el fin de obtener una relación clara entre la categoría y el texto obtenido, por lo que podemos concluir que **no es del todo viable** confiar en el modelo para su tarea de clasificación de problemas que no estén relacionados con los Neoplasmas o los problemas cardiovasculares. Sin embargo, de

los resultados del proyecto podemos rescatar que podríamos utilizar el modelo de **MLPClassifier** de manera confiable para determinar si un problema médico dado se refiere o no a un problema relacionado con el sistema digestivo. Además de lo anterior, identificamos también que es extremadamente probable que con un conjunto de datos más grande y mejor estructurado puede ser posible mejorar significativamente los resultados de los modelos. En resumen, nuestro modelo de **Regresión logística**, puede resultar de cierta utilidad para ayudar a los equipos médicos a la clasificación de ciertos problemas médicos en las categorías de **Neoplasmas** y **Problemas cardiovasculares** con aproximadamente un 67% de exactitud.

4. Enlaces a entregables

Los siguientes enlaces redirigen a los entregables de presentación, video y repositorio con el notebook y Dashboard del proyecto.

Video y Presentación: https://www.canva.com/presentacion/

Repositorio: https://github.com/drvillota/Proyecto-1-ML