# Proyecto 1

Grupo 7:
Arturo Rubio Caballero 201815176, Juan Camilo González 201620257, Sofía Abadía Bermeo 201612390
Inteligencia de Negocios
Profesora: Haydemar María Nuñez Castro
Departamento de Ingeniería de Sistemas y Computación
Universidad de Los Andes, Bogotá
{a.rubio, jc.gonzalezv, s.abadia} @uniandes.edu.co
Marzo 28, 2022

## Tabla de contenido

**1. INTRODUCCIÓN** ..... **2**
**2. COMPRESIÓN DEL NEGOCIO Y ENFOQUE ANALÍTICO** ..... **2**
**3. COMPRESIÓN Y PREPARACIÓN DE LOS DATOS** ..... **3**
3.1. EXPLORACIÓN DE LOS DATOS ..... 3
3.2. PREPROCESAMIENTO Y LIMPIEZA ..... 3
**4. MODELADO Y EVALUACIÓN** ..... **4**
4.1. NAIVE BAYES MULTINOMIAL ..... 4
4.2. ÁRBOLES DE DECISIÓN ..... 5
4.3. RANDOM FOREST ..... 5
**5. RESULTADOS** ..... **6**
**6. TRABAJO EN EQUIPO** ..... **7**
**BIBLIOGRAFÍA** ..... **7**

# 1. Introducción

El objetivo de este documento en presentar y explicar el paso a paso del desarrollo del Proyecto 1 de la clase de Inteligencia de negocios. El objetivo de este proyecto es realizar un análisis de texto con el fin de determinar la elegibilidad de un paciente para ensayos clínicos de cáncer (ISIS 3301, 2022). Para realizar la tarea de análisis de texto se van a aplicar 3 algoritmos a los datos, MultinomialNB, Árboles de Decisión y Random Forest. Con el fin de que los algoritmos tengan sus mejores resultados, se comenzará realizando una exploración de los datos de forma que se puedan determinar las acciones de limpieza y poder tomar, de manera informada, todas las decisiones al respecto. Una vez los datos estén limpios, se aplicará cada algoritmo y se realizará una comparación para llegar a seleccionar el mejor. Finalmente, se presentarán una serie de recomendaciones al negocio; esta parte se complementará con una presentación y un video.

# 2. Compresión del negocio y enfoque analítico

El objetivo principal del negocio es encontrar a aquellos pacientes que son aptos para que se les realicen ensayos clínicos.

El ensayo clínico de cáncer intervencionista es una prueba de eficacia de un tratamiento médico suministrado a una persona con cáncer, la cual ha sido seleccionada bajo determinados criterios, como edad o comorbilidades. Lo que se busca con el análisis de textos es determinar si una persona es elegible para un ensayo clínico de acuerdo con las declaraciones clínicas breves realizadas. Con esto, el tiempo necesario para iniciar un ensayo clínico se acorta significativamente, permitiendo que el medicamento testeado, en caso de obtener un resultado positivo, sea prontamente usado en las personas con un respectivo cáncer.

*Tabla 1. Entendimiento*

| | | |
|---|---|---|
| **Oportunidad / Problema de negocio** | Determinar los pacientes elegibles para ensayos clínicos de cáncer intervencionista | |
| **Descripción del requerimiento desde el punto de vista de aprendizaje de máquina** | Primero se separarán las palabras y realizará un conteo de cada palabra con el fin de determinar su importancia en la frase. Una vez hecho esto, se construirán los 3 modelos y se determinará la exactitud de cada, para poder entregar al negocio el mejor modelo. | |
| **Tipo de aprendizaje** | **Tarea de aprendizaje** | **Algoritmo e hiper-parámetros** |
| Supervisado | Clasificación | MultinomialNB:<br>Permite la predicción de la clase a la que pertenece cierto dato. |

| | | |
|---|---|---|
| | | Se dejó el hiper-parámetro default, alpha = 1.0 |
| Supervisado | Clasificación | Árboles de decisión:<br>Es un algoritmo de clasificación sencillo y rápido.<br>Hiper-parámetros: encontrados a través de GridSearch<br>o Criterio: gini<br>o Max_depth: 20<br>o Min_samples_split: 4 |
| Supervisado | Clasificación | Random Forest:<br>Útil para trabajar en clasificación de textos, provee una predicción.<br>Hiper-parámetros: encontrados a través de GridSearch<br>o Max_features: 4<br>o N_estimators: 9 |

# 3. Compresión y preparación de los datos

Los datos originales sobre la elegibilidad de pacientes se encuentran en la plataforma Kaggle, bajo un proyecto llamado “Clinical Trials on Cancer” (Auriml, 2017). Sin embargo, a nosotros se nos proporcionó un archivo de datos basado en los datos originales, pero con ciertas modificaciones.

## 3.1. Exploración de los datos

En el archivo de datos hay 2 columnas y 12000 filas. La primera columna – label – indica si el paciente es elegible para ensayo clínico o no, mientras que la segunda columna – “study_and_condition” – se encuentra la declaración clínica breve.

La variable “label” tiene 2 posibles valores, label 0 si el paciente es elegible, label 1 en caso contrario; ésta variable está perfectamente balanceada, es decir, se tienen 6000 datos que pertenecen a label 0 y 6000 que pertenecen a label 1.

No se tienen valores vacíos y hay 12 filas duplicadas.

## 3.2. Preprocesamiento y limpieza

1. Se remueven todos los caracteres que no pertenezcan a los caracteres ASCII, pues los datos están en inglés, por lo que se espera que todas las palabras hagan parte del alfabeto latino.

2. Se pasan todas las letras a minúsculas para que la comparación de palabras sea efectiva.
3. Se eliminan todos los signos de puntuación, ya que sólo nos interesan las palabras.
4. Se reemplazan todos los números que aparezcan en la variable "study_and_condition" por su representación textual.
5. Se eliminan todas las stopwords, es decir, todas las palabras que no son significativas para el análisis como the, I, are, entre otras.
6. La variable objetivo "label" se vuelve numérica:
    - Label 0 = 0
    - Label 1 = 1
7. Se corrigen las palabras con contracciones gramaticales, así, "they've" pasa a ser "they have" por ejemplo.

# 4. Modelado y evaluación

Antes de desarrollar los modelos, se utilizó CountVectorizer y TfidfTransformer para poder analizar el texto de la segunda variable "study_and_condition".

Con CountVectorizer lo que se busca es convertir todas las palabras a columnas (variables) y contar cuantas veces está repetida una palabra en la frase de forma que, se convierten todas las frases en una matriz de recuentos de palabras (Scikit Learn, 2022).

Posteriormente, se utilizó TfidfTransformer para eliminar todas aquellas palabras que se usan mucho en las frases pero que no tienen un aporte significativo (Scikit Learn, 2022). Algunas de estas palabras son study o interventions, pues todas las frases comienzan con estas dos palabras, lo que quiere decir que, en realidad, no son significativas.

Una vez realizados estos procesos, se obtienen 8209 columnas y 12000 filas (el número de filas no cambia) y pasa a dividir los datos en los conjuntos de entrenamiento y de prueba.

## 4.1. Naive Bayes Multinomial

Este algoritmo es eficaz a la hora de generar una clasificación con características discretas, como lo es el recuento de palabras para la clasificación de un texto (Scikit Learn, 2022).

A este algoritmo no se le asignaron hiper-parámetros, se le dejó el valor default (1.0) a alpha, por lo que se pasó a probar inmediatamente. Los resultados obtenidos al hacer la predicción con el conjunto de prueba fueron los siguientes:

*Tabla 2. Resultados de MultinomialNB*

| | Precisión | Recall | Exactitud |
|---|---|---|---|
| **Label 0: 0** | 0.8 | 0.76 | 0.78 |
| **Label 1: 1** | 0.76 | 0.8 | |

Estos datos muestran que es un buen modelo en general y, aunque, hay varios valores que se están clasificando en la categoría que no les corresponde, los valores de Presición y Recall son lo suficientemente altos para afirmar que los falsos positivos y los falsos negativos en la variable de decisión no son muy significativos.

## 4.2. Árboles de decisión

Para implementar el algoritmo de árboles de decisión no es necesario hacer una normalización de los datos por lo que se realizó, en primer lugar, una búsqueda de hiper-parámetros con GridSearch. El espacio de búsqueda que se determinó y el mejor modelo obtenido son los siguientes:

*Tabla 3. Hiper-parámetros Árboles*

| | Espacio de búsqueda | Mejor modelo |
|---|---|---|
| **Criterio** | Entropy, Gini | Gini |
| **Máxima profundidad** | [4, 6, 8, 10, 20] | 20 |
| **Particiones Mínimas** | [2, 3, 4, 5] | 4 |

Los resultados que se obtuvieron sobre el conjunto de prueba se presentan a continuación.

*Tabla 4. Resultados Árboles*

| | Precisión | Recall | Exactitud |
|---|---|---|---|
| **Label 0: 0** | 0.74 | 0.58 | 0.68 |
| **Label 1: 1** | 0.64 | 0.78 | |

Con este modelo no se obtuvieron tan buenos resultados, pues el valor de Recall del label 0 es bajo, lo que quiere decir que hay una cantidad significativa de falsos negativos en esta categoría. Adicionalmente, la exactitud tampoco es muy buena.

## 4.3. Random Forest

El algoritmo de Random Forest, al igual que el de árboles de decisión, no requiere normalización de los datos, por lo que el primer paso será buscar los hiper-parámetros que nos den el mejor modelo. El espacio de búsqueda y el mejor modelo obtenido son los siguientes.

*Tabla 5. Hiper-parámetros RF*

| | Espacio de búsqueda | Mejor modelo |
|---|---|---|
| **Rango de estimadores** | (1, 10) | 9 |
| **Máximo de características** | (1, 5) | 4 |

Los resultados que se obtuvieron sobre el conjunto de prueba se presentan a continuación.

*Tabla 6. Resultados RF*

| | Precisión | Recall | Exactitud |
|---|---|---|---|
| **Label 0: 0** | 0.75 | 0.74 | 0.74 |
| **Label 1: 1** | 0.73 | 0.75 | |

El modelo obtenido es un buen modelo, pues los valores de Precisión y Recall son los suficientemente altos para afirmar que no hay una gran cantidad de falsos negativos y falsos positivos. También se obtuvo un buen porcentaje de exactitud.

# 5. Resultados

Para seleccionar el mejor modelo se hará una comparación de las Tablas 2, 4 y 6, para esto se hará una recopilación de la información en la Tabla 7.

*Tabla 7. Comparación de algoritmos*

| | MultinomialNB | Árboles de decisión | Random Forest |
|---|---|---|---|
| **Precisión** | 0: 0.8 | 0: 0.74 | 0: 0.75 |
| | 1: 0.76 | 1: 0.64 | 1: 0.73 |
| **Recall** | 0: 0.76 | 0: 0.58 | 0: 0.74 |
| | 1: 0.8 | 1: 0.78 | 1: 0.75 |
| **Exactitud** | 0.78 = 78% | 0.68 = 68% | 0.74 = 74% |

Con base en la Tabla 7, el algoritmo que se le recomienda al negocio es el MultinomailNB ya que es el que provee una mejor exactitud y en donde los valores de Precisión y Recall son más altos, es decir, es el algoritmo que mejor realiza la tarea de clasificación, pues no genera una gran cantidad de falsos positivos y falsos negativos.

Este resultado implica que, al atender a un nuevo paciente con cáncer, se va a poder determinar con una exactitud del 78%, si es un candidato apto para ensayo clínico. Esta información será de gran utilidad para el negocio, pues se podrá disminuir el tiempo de inicio del ensayo y obtener un resultado final más rápidamente.

De esta forma, se realizan las siguientes recomendaciones al negocio:

1. Utilizar el algoritmo de MultinomialNB para la realización del análisis de textos.
2. Recolectar la mayor cantidad de datos de pacientes con cáncer, por medio de asociaciones con hospitales que les provean esta información.
3. Tratar de que los datos sean lo más limpios posible, de forma que el proceso de limpieza no sea largo.

# 6. Trabajo en equipo

Durante el desarrollo del proyecto nos encontramos con varias dificultades, primero nos demoramos un buen tiempo entendiendo qué era lo que nos pedían hacer y, una vez lo logramos definir, tuvimos percances tratando de encontrar los algoritmos óptimos para resolver el problema, pues habían algoritmos que solucionaban el requerimiento del negocio pero que no se podían utilizar por el formato de los datos; mientras que habían otros que recibían correctamente los datos de entrada, pero que no servían para resolver el requerimiento del negocio.

Uno de los algoritmos que intentamos implementar fue Word2Vec, sin embargo, este algoritmo no nos servía pues no podía categorizar al paciente, es decir, predecir. Este algoritmo lo dejamos al final del notebook como muestra de que se intentaron diversos algoritmos que, al final, no funcionaban para lo que se requería.

Al final, decidimos no dividir el trabajo para resolver el proyecto, sino que todos trabajamos a la par; mientras que unos estaban buscando algoritmos que funcionaran, el otro iba implementándolos en el notebook o, mientras se terminaba la implementación otros iban trabajando en el informe, la presentación y el video.

En total, el tiempo invertido en el desarrollo del proyecto fue de aproximadamente 15 horas, entre la realización del notebook, el informe, la presentación y el video.

A continuación, se expondrán los roles que desempeñó cada estudiante y su puntaje de participación:

*Tabla 8. Roles y participación*

| Estudiante | Rol(es) | Puntaje de participación |
|---|---|---|
| **Arturo Rubio Caballero** | Líder de datos | 33.3% |
| **Juan Camilo González** | Líder de analítica<br>Líder de proyecto | 33.3% |
| **Sofía Abadía Bermeo** | Líder de negocio | 33.3% |

# Bibliografía

ISIS 3301. (2022). *Enunciado Primer Proyecto.*

Auriml. (2017). *Clinical Trials on Cancer*. Obtenido de Kaggle: https://www.kaggle.com/datasets/auriml/eligibilityforcancerclinicaltrials

Scikit Learn. (2022). *Count Vectorizer*. Obtenido de Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Scikit Learn. (2022). *TfidfTransformer*. Obtenido de Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

Scikit Learn. (2022). *MultinomialNB*. Obtenido de Scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html