**Proyecto Analítica de Textos**

**Caso: Elegibilidad de un paciente para ensayos clínicos.**

**Nicolas Orjuela, Camilo Salinas, Felipe Bedoya**

**Sección 1.**

Tabla de contenido

[**1.** **Comprensión del negocio y enfoque analítico.** 1](#_Toc99269400)

[**2.** **Comprensión de los datos y preparación de los datos** 2](#_Toc99269401)

[**3.** **Modelado y evaluación** 2](#_Toc99269402)

[**4.** **Resultados** 2](#_Toc99269403)

[**5.** **Trabajo en equipo** 2](#_Toc99269404)

# **Comprensión del negocio y enfoque analítico.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Oportunidad/problema Negocio** | | Se necesita determinar la elegibilidad de los pacientes para ensayos clínicos de cáncer a partir de textos descriptivos. | |
| **Descripción del requerimiento desde el punto de vista de aprendizaje de máquina** | | Escoger y entrenar un modelo que, a partir de la analítica de textos y un conjunto de textos descriptivos, pueda determinar la etiqueta de nuevos textos, siendo la etiqueta la elegibilidad para ensayos clínicos de cáncer. | |
| **Detalles de la actividad de minería de datos** | | | |
| **Tipo aprendizaje** | **Tarea de aprendizaje** | | **Algoritmo e hiper-parametros (con la justificación respectiva)** |
| Supervisado | Clasificación | | Naive Bayes ( |
| Supervisado | Clasificación | | Support Vector Machines ( |
| Supervisado | Regresión | | Logistic Regression ({'clf\_\_C': 1.0, 'clf\_\_penalty': 'l2', 'clf\_\_solver': 'newton-cg', 'tfidf\_\_ngram\_range': (1, 1), 'tfidf\_\_use\_idf': True}) |

Todos los hiperparámetros de los modelos fueron escogidos a través de un GridSearch que determinó cuales son los más apropiados para este caso en particular a partir de un conjunto de posibles hiperparámetros.

La escogencia de los hiperparámetros que debian ir en el GridSearch se baso en aquellos que mas influenciaran los resultados y el comportamiento del modelo. En LogisticRegression por ejemplo uno de los hiperparámetros es el solver, que puede tener implicaciones de velocidad y efectividad dependiendo del tamaño y tipo de los datos.

# **Comprensión de los datos y preparación de los datos**

Al cargar los datos y analizar sus características, nos dimos cuenta de que estos no tienen nulos, pero están presentados de maneras distintas. Algunos comienzan con comillas, otros son espacios. Era necesario lograr entradas similares. Adicionalmente la información importante del paciente estaba ubicada después del primer punto. Además, vimos que los labels son categóricos, por lo que era necesario volverlos numéricos. Aunque un label encoder haría esta tarea con facilidad, se quería conservar la categoría implícita que ya traen.

Otras transformaciones realizadas a los textos fueron la partición de estos en tokens y lematizándolos. Finalmente se utilizó un modelo de Bag Of Words y tf-idf para identificar las palabras importantes. También se utilizó el corpus de nltk para quitar las palabras conectoras que generaran ruido. Estas transformaciones fueron añadidas a un pipeline para poder ajustar el modelo escogido por cada uno.

# **Modelado y evaluación**

Se escogió Naive Bayes debido a

Se escogió Support Vector Machines debido a que estos modelos son considerados unos de los más apropiados para la analítica de textos gracias a su manejo de hiperplanos en la clasificación de datos. Se realizó un GridSearch para encontrar los mejores hiperparámetros para el modelo. El mejor modelo obtenido tuvo una exactitud de 0.82 frente a los datos de prueba.

Se escogió Logistic Regression debido a que presentaba un punto de vista distinto utilizando una tarea de aprendizaje contrastante con las anteriores. Aunque finalmente se utilice como un clasificador dada la naturaleza de las etiquetas, sirve para hacer interpretaciones y comprender mas a profundidad el comportamiento de los datos y que patrones siguen. Viendo un métricas buenas, se podría decir que los datos siguen una regresión logística, aunque sea muy difícil poner en contexto numérico al Bag of Words generado y la amplia matriz con palabras con columnas.

Para este algoritmo también se realizó un GridSearch con un conjunto de hiperparámetros que incluyen las dos penalidades posibles y un rango de coeficientes de regularización. También se incluye en el grid si se usa o no el IDF para ver esto como varía con el comportamiento frecuencial de los tokens. Esto se probó con dos solvers. El modelo más optimo tuvo como coeficiente de determinación R2 un valor de 0.84, junto con un valor de 0-39 en su Root-Mean-Square Error. La exactitud del modelo con los datos de prueba fue de 0.806.

# **Resultados**

Como se puede observar en el tablero de control,

# **Trabajo en equipo**

1. Roles:
   * Líder de proyecto: Felipe Bedoya
   * Líder de negocio: Camilo Salinas
   * Líder de datos: Felipe Bedoya
   * Líder de analítica: Nicolás Orjuela
2. Algoritmo trabajado:
   * Naive Bayes: Camilo Salinas
   * Support Vector Machines: Nicolás Orjuela
   * Logistic Regression: Felipe Bedoya
3. Retos enfrentados:
4. Soluciones planteadas:
5. Tareas realizadas:
   * Preprocesamiento de los datos: Felipe Bedoya (3 horas)
   * Naive Bayes: Camilo Salinas (X horas)
   * Support Vector Machines: Nicolás Orjuela (x horas)
   * Logistic Regression: (3 horas)
   * Análisis de resultados y creación del tablero de control:
   * Preparación del video y presentación:
6. Repartición de los puntos:

A

* + Felipe Bedoya:
  + Nicolás Orjuela:
  + Camilo Salinas:

1. Puntos por mejorar en el siguiente proyecto: