**Proyecto 1**

**Contenido**

[**Comprensión del negocio y enfoque analítico.** 2](#_Toc99574614)

[**Comprensión y preparación de los datos** 2](#_Toc99574615)

[**Modelado y evaluación** 3](#_Toc99574616)

[**Resultados** 5](#_Toc99574617)

[**Trabajo en equipo** 6](#_Toc99574618)

[**Referencias** 7](#_Toc99574619)

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

|  |  |  |
| --- | --- | --- |
| **Oportunidad/problema Negocio** | Determinar la elegibilidad de un paciente para ensayos clínicos de cáncer a partir de texto descriptivo. | |
| **Descripción del requerimiento desde el punto de vista de aprendizaje de máquina** | De acuerdo con las declaraciones clínicas breves, predecir si estas se consideran criterios de inclusión o exclusión de un paciente en un ensayo clínico de cáncer. | |
| **Detalles de la actividad de minería de datos** | | |
| **Tipo de aprendizaje** | **Tarea de aprendizaje** | **Algoritmo e hiperparámetros utilizados (con la justificación respectiva)** |
| Análisis de palabras en las declaraciones | Procesamiento de lenguaje natural, análisis de texto | Tokenización,  Análisis de sentimientos |
| Supervisado | Clasificación | Se decide aplicar 3 diferentes algoritmos de clasificación con el fin de elegir el que mejor se ajuste a los datos.  -Naive Bayes  - Regresión Logística  - KNN  El algoritmo seleccionado luego de comparar los resultados fue: ***Regresión logística***  Con una exactitud de: ***80,25%*** |

# **Comprensión y preparación de los datos**

* Habían dos columnas y 12.000 registros. La primera columna correspondía a la etiqueta de si el paciente había sido seleccionado para el ensayo. 0, aplica; 1, no aplica. La segunda columna correspondía al texto.
* No se encontraron datos nulos.
* Los datos se encontraban balanceados, habiendo igual cantidad de datos para cada etiqueta: 6.000 para 0, e igual cantidad para 1.

asd

**Tokenización: Sacar las palabras más importantes del texto**

* Separación en dos bloques.
* Pasar todo a minúscula.
* Eliminar palabras de parada como conectores y caracteres que no fueran palabras.
* Del segundo bloque se hizo separación en una lista de palabras.

**Lematización: Para ver que tanto afecta cada palabra al estudio**

* Queda una lista con únicamente las raíces de las palabras.
* Facilita la identificación de la influencia de cada palabra.

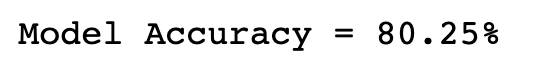
# **Modelado y evaluación**

* Regresión logística:
* Descripción del algoritmo:

Regresión logística es un algoritmo de clasificación. Para la implementación del algoritmo no era estrictamente necesario normalizar los datos por lo que no se realizó este paso

* Resultados cuantitativos:

Para el reporte de métricas al negocio se realizó un promedio de los resultados que arroja aplicar el modelo con las métricas obtenidas con anterioridad al conjunto de prueba dando como resultado: Se reporta una exactitud del **80.25%**

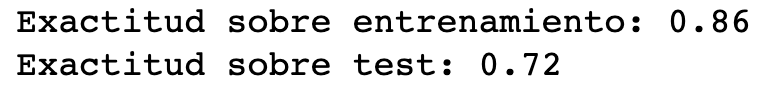


* El mejor modelo fue el de Regresión logística con una métrica de exactitud de ***80,25%***. Se debe aclarar que, aunque el modelo de *Naive Bayes* arrojo resultados similares, lo correcto para el negocio seria utilizar este modelo debido a que no consume tantos recursos.
* KNN:
* Descripción del algoritmo:

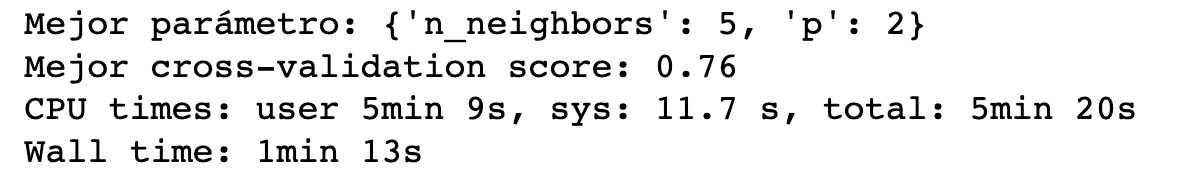
El algoritmo de KNN es un algoritmo de clasificación

* Resultados cuantitativos:

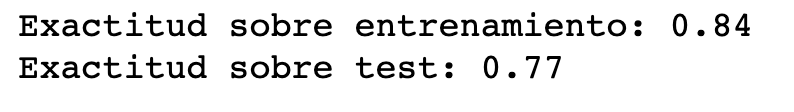
Para la implementación del modelo KNN se elimina el ruido, se tokenizan las palabras y normalizamos los datos. Después se implementó el algoritmo KNN. Obtuvimos los siguientes resultados:



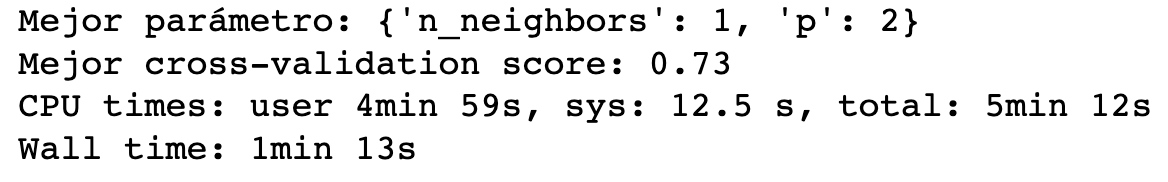
Aplicamos búsqueda con hiperparámetros de máximo 10 vecinos obteniendo como resultado que el n debe ser 5 para los vecinos y debemos utilizar la función euclidiana.



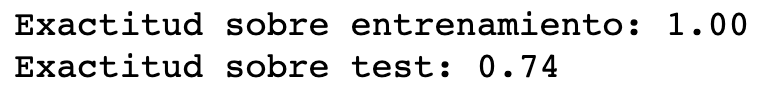
Los resultados aplicando el mejor estimador fueron:



Luego de normalizar los resultados obtuvimos que ahora el mejor numero de n vecinos es de 1 y que debemos usar la función Manhattan:



Aplicamos el modelo y obtuvimos una precisión de **74%**:

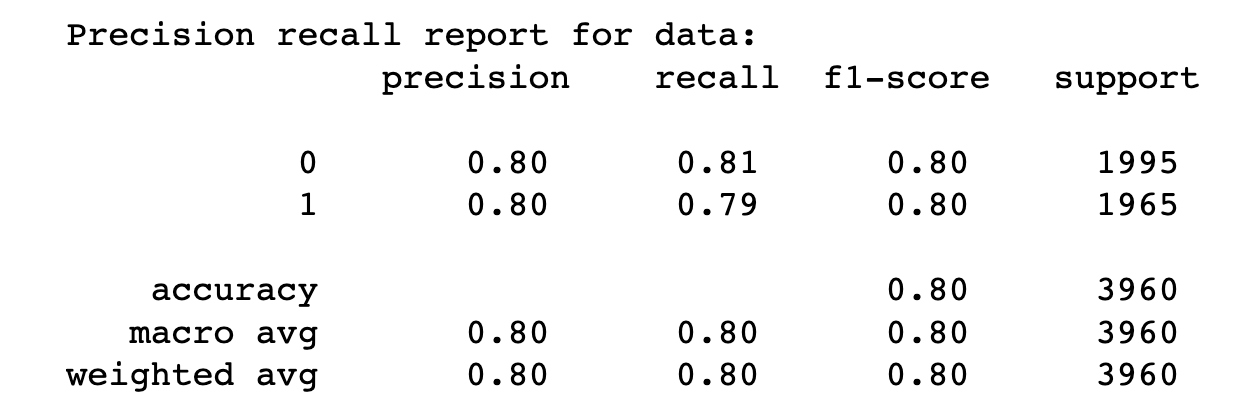


* Naive Bayes:
* Descripción del algoritmo:

es un algoritmo de clasificación para problemas de clasificación binarios (dos clases) y multiclase. La técnica es más fácil de entender cuando se describe utilizando valores de entrada binarios o categóricos.

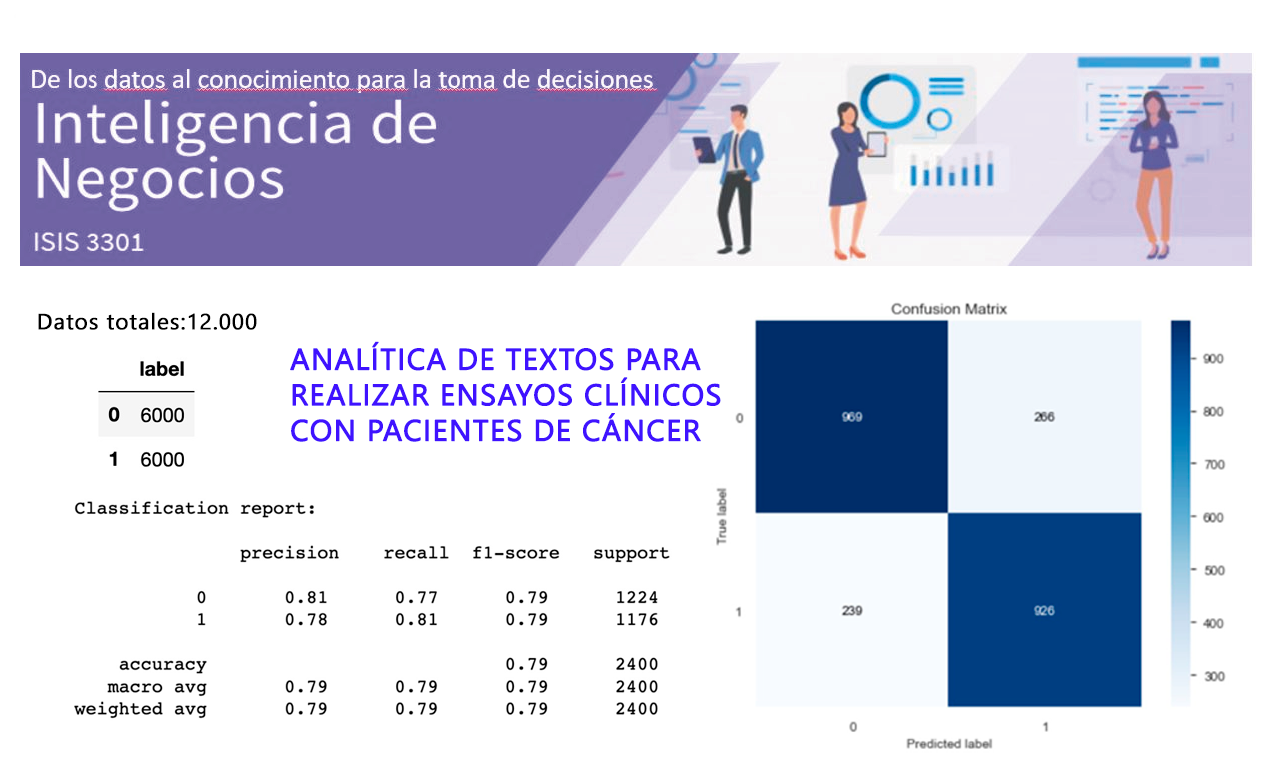
* Resultados cuantitativos:

Para el reporte de métricas al negocio se realizó un promedio de los resultados que arroja aplicar el modelo con las métricas obtenidas con anterioridad al conjunto de prueba dando como resultado una exactitud del **80%**



# **Resultados**

* **¿Porque es útil la información?:** Nos permite identificar si un paciente va a ser un candidato para ensayos clínicos de cáncer. Para esto vemos que tipo de palabras se usaron en el reporte clínico que se le realizó.



# **Trabajo en equipo**

* **Daniel Santamaría:**
* **Rol:** Líder de negocio, líder de analítica
* **Algoritmo:** Naive Bayes
* **Tareas Realizadas:**
* Desarrollo del algoritmo Naive Bayes
* Verificación de que la solución se alinea con las necesidades del negocio.
* Elección del modelo según las métricas obtenidas en los algoritmos.
* **Retos y como se solucionaron:**
* Falta de entendimiento del problema: Se nos dificulto entender que era lo que necesitaba el negocio. Investigamos para preparar correctamente los datos.
* **Puntuación:**33.3
* **Gabriel Serna:**
* **Rol:** Líder de datos
* **Algoritmo:** Regresión logística
* **Tareas Realizadas:**
* Desarrollo del algoritmo Regresión logística
* Limpieza y preparación de datos
* **Retos y como se solucionaron:**
* La correcta manipulación de los datos y poder convertir los str en floats fue un problema que al principio no entendíamos y nos tenía frenados hasta que recordamos la clase de la profe y aplicamos la vectorización.
* **Puntuación:**33.3
* **Daniela Manrique:**
* **Rol:** Líder del Proyecto
* **Algoritmo:** KNN
* **Tareas Realizadas:**
* Desarrollo del algoritmo KNN
* Planeación de las fechas de reunión del equipo y coordinación de la división del trabajo.
* **Retos y como se solucionaron:**
* Falta de tiempo debido a la alta carga de trabajo de los integrantes del equipo. Se solucionó programando las reuniones con bastante tiempo de antelación.
* **Puntuación:**33.3

# **Referencias**

* <https://www.aprendemachinelearning.com/regresion-logistica-con-python-paso-a-paso/>
* <https://gitlab.virtual.uniandes.edu.co/ISIS3301/practicas/tree/master/ProcesamientoTextos>