

Proyecto 1

Santiago Bobadilla – 201820728

Juan José Beltrán – 201819446

~\*~

Mateo Visbal – Experto

Silvana Ruiz – Experta

Laura Mejía – Experta



28 de marzo de 2022

Semestre 2022-01

Contenido

[Introducción 2](#_Toc99386770)

[Comprensión del Negocio y Enfoque Analítico 2](#_Toc99386771)

[Problema / Oportunidad del negocio 2](#_Toc99386772)

[Requerimientos 3](#_Toc99386773)

[Entendimiento de los datos 5](#_Toc99386774)

[Limpieza de los datos 6](#_Toc99386775)

[Ajuste de Datos: Forma 6](#_Toc99386776)

[Ajuste de Datos: Lenguaje 7](#_Toc99386777)

[Bolsa de Palabras y Tf-Idf 7](#_Toc99386778)

[Modelos de Clasificación 8](#_Toc99386779)

[Naive Bayes Classifier 8](#_Toc99386780)

[Support Vector Machine 9](#_Toc99386781)

[Árbol de Decisión 10](#_Toc99386782)

[Comparación 10](#_Toc99386783)

[Recomendaciones 11](#_Toc99386784)

[Resultados 11](#_Toc99386785)

[Conclusiones 11](#_Toc99386786)

[Anexos 11](#_Toc99386787)

[Bibliografía 12](#_Toc99386788)

## Introducción

La Analítica de Textos (AT) es un campo interdisciplinario que conjuga el machine learning (aprendizaje de máquina) y el Procesamiento de Lenguaje Natural, y que tiene como objetivo procesar de manera automática grandes cantidades de textos para la extracción de conocimiento que apoye la toma de decisiones.

La medicina no es la excepción, partiendo de descripciones dispuestas por los médicos se puede establecer de manera automática la categorización de ciertas enfermedades. Con tal fin, buscamos presentar a continuación tecnología de asistencia que puede identificar, con alta precisión, la clase de problemas descritos en la descripción del médico.

## Comprensión del Negocio y Enfoque Analítico

Los resúmenes médicos describen las condiciones actuales de un paciente. Los médicos rutinariamente escanean docenas o cientos de resúmenes cada día mientras hacen sus rondas en un hospital y deben recoger rápidamente la información saliente que apunta a la enfermedad del paciente.

Buscamos automatizar (como herramienta de ayuda en la toma de decisiones) un identificador de alta precisión para la clase de problemas descritos en el resumen. Se cuenta resúmenes de 5 condiciones diferentes: enfermedades del sistema digestivo, enfermedades cardiovasculares, neoplasias, enfermedades del sistema nervioso y condiciones patológicas generales. En el siguiente enlace se encuentran los datos originales del problema. (Medical Text, n.d.)

### Problema / Oportunidad del negocio

En la unidad inmediatamente anterior se encuentra el problema al cual se enfrenta el negocio. A continuación, se presentan el objetivo general y los objetivos específicos derivados de las necesidades del negocio y enfocados en la tarea asignada a desarrollar con el trabajo.

#### Objetivo General

Ayudar al proceso de diagnóstico médico por medio de algoritmos de clasificación basados en información histórica de resúmenes de diferentes pacientes.

#### Objetivos Específicos

* Obtener el modelo con mayor precisión posible.
* Lograr una limpieza y procesamiento de datos clara y replicable para futuros resúmenes médicos.
* Entregar al negocio un modelo que permita clasificar a un nuevo paciente dado el informe del médico que lo trató.

### Requerimientos

Con el fin de realizar la producción de la herramienta de apoyo a la decisión se requiere de cierta información importante para cumplir con lo solicitado por el negocio:

1. Una muestra significativa de la información con la cual se va a realizar la clasificación. En este caso, la información debe tener una cantidad de registros suficiente para que el modelo aprenda de manera robusta, así como la categoría a la cual pertenece cada registro. Es decir, a cuál enfermedad fue clasificado el resumen.
2. Conocimiento de expertos en la temática con el fin de realizar pruebas de testeo y validación de decisiones de procesamiento de datos.

#### Tipo de Aprendizaje

Dados los requerimientos de entrada que se plantearon, y la información suministrada que cumple con dichos requerimientos, se establece que estamos trabajando con un aprendizaje supervisado. Es decir, el modelo aprende a comportarse con base en una columna objetivo (también llamada target). Aprende y mejora supervisando su aprendizaje con base en la variable de respuesta ya conocida.

#### Tarea de Aprendizaje

Dentro de la rama de aprendizaje supervisado, y dado la naturaleza del problema, se va a trabajar con la tarea de clasificación. Es decir, a partir de información explicativa poder clasificar un registro bajo una(s) de las categorías presentes en la variable de respuesta. En este caso se debe aclarar que el registro solamente pertenece a una de las categorías; y en el contexto del problema estas son mutuamente excluyentes.

#### Algoritmo e Hiper-Parámetros

Con el fin de llegar a una respuesta concreta y acertada de un modelo capaz de ser herramienta de apoyo a la decisión, se decidió probar diferentes algoritmos de clasificación con el fin de comparar resultados y poner a disposición el mejor. En ese orden de ideas se usaron tres:

1. Naïve Bayes Classifier
2. Support Vector Machine
3. Árbol de Decisión

A continuación, se procede a explicar de manera simple y resumida como funciona el algoritmo y su configuración de hiper-parámetros en caso de tener.

##### Naïve Bayes Classifier

Se basa en la aplicación del teorema de Bayes con la suposición "ingenua" de independencia condicional entre cada par de características dado el valor de la variable de clase. El teorema de Bayes planea la siguiente relación entre una variable de respuesta, dado un conjunto de variables explicativas .

Texto

Descripción generada automáticamente con confianza media

Bajo la expansión estadística de dicho teorema con base en la tarea que se busca cumplir, el algoritmo respecto a la función de:

Texto

Descripción generada automáticamente

Específicamente, dado que tenemos una tarea de clasificación no binaria, más de dos categorías, se usa un Multinomial Naive Bayes. Así se implementa el algoritmo ingenuo de Bayes para datos distribuidos multinomialmente, y es una de las dos variantes clásicas utilizadas en la clasificación de textos.

Su único hiper-parámetro es el cual nos permite ajustar ciertas características del teorema de Bayes con el fin de suavizar o apretar la relajación de la ‘ingenuidad’ al realizar alisado de Laplace o alisado de Lidstone.

En dado caso que se considere necesario, se puede definir que variables son de mayor importancia para el aprendizaje del algoritmo, así como sugerir distribución de ajuste de los datos esperada. Ambos son parámetros del modelo más su uso se sugiere bajo explicación de expertos y no por medio de pruebas.

##### Support Vector Machine

Son métodos de aprendizaje supervisados utilizados para la clasificación, regresión y detección de valores atípicos. Específicamente usamos el clasificador de SGD. El cual consiste en la técnica de optimización de descenso de gradiente estocástico. Esta técnica consiste en estimar el gradiente de la pérdida de cada muestra a la vez y actualizar el modelo en el camino con un horario de fuerza decreciente (es decir, tasa de aprendizaje). Siempre se busca minimizar la función de perdida.

Por consiguiente, los hiper-prámetros con los que se cuentan son:

* Función de perdida. La cual busca elegir entre ‘hinge’, ‘log’, ‘modified\_huber’, ‘squared\_hinge’, ‘perceptron’, or a regression loss.
* Penalty: Relacionacionado con la regularización AKA. Se puede elegir entre ‘l2’, ‘l1’, ‘elasticnet’.
* Alpha: Constante que multiplica el termino de regularización.
* Max\_iter: El número máximo de pases sobre los datos de entrenamiento (también conocidos como épocas).
* Tol: El criterio de parada.

Hay muchos más, pero se consideran estos como los más relevantes a ajustar.

##### Árbol de Decisión

El árbol de decisión es un algoritmo que se basa en la partición binaria al momento de construir su clasificación; armando la ramificación por medio de particiones de pureza. Entre mayor pureza, mejor la ramificación.

Así los hiper-parámetros a ajustar son:

* Criterion: Con el cual se va a medir la calidad del modelo, es decir que tan bien se esta ramificando. Puede ser ‘gini’ o ‘entropy’ común mente.
* Max\_depth: Que indica la mayor profundidad que puede tomar el árbol.
* Min\_sample\_split: Que indica el número mínimo de splits que deben ocurrir.

Los tres algoritmos se contrastaron y se buscó la mejor configuración de hiper-parámetros. Posteriormente se muestran los resultados y procesamiento de los datos.

## Entendimiento de los datos

En primera instancia, se deben analizar los datos entregados por la empresa con el fin de ver la calidad de los mismo, así como el debido tratamiento que se les debe dar. En ese orden de ideas empezamos definiendo y revisando las variables, sus rangos, y sus tipos.

Como se puede observar en conjunto de datos cuenta con dos columnas:

* Problems\_described ~ Target: [1 a 5] donde se estipula la enfermedad a la que pertenece ese registro. Es de tipo numérica y su información es correcta. Se revisa qué tan balanceada están las categorías de target.

Gráfico, Gráfico de barras

Descripción generada automáticamente

* Medical\_abstracts: Texto con la información médica. Es de tipo ‘string’ y si bien su información es correcta se debe corregir.

### Limpieza de los datos

Cada uno de los registros contiene ciertas palabras claves, por tanto, se debe tener una inferencia inicial de la longitud de los textos en palabras según el problema descrito.

Gráfico, Histograma

Descripción generada automáticamente

Con esto se procede a realizar dos tipos de limpieza, con sus debidas especificaciones, teniendo en cuenta que los histogramas muestran que todas las clases están homogéneamente distribuidas entre las clases en términos de longitud:

### Ajuste de Datos: Forma

1. Remover caracteres no ASCII (dado el alfabeto usado).
2. Volver toda palabra minúscula con el fin de evitar distinciones entre palabras iguales, pero con alguna mayúscula. Es decir, realizar comparación homogénea.
3. Remover los signos de puntuación, dado que no aportan nada.
4. Ponen los números en una codificación de texto, con el fin de que se puedan tener en cuenta en el análisis.
5. Eliminar las palabras que no aportan nada al significado, mejor conocidas como ‘stopwords’. Se eliminan dado que su función es más gramatical que de dar significado lógico.

Así, se obtiene la siguiente representación:

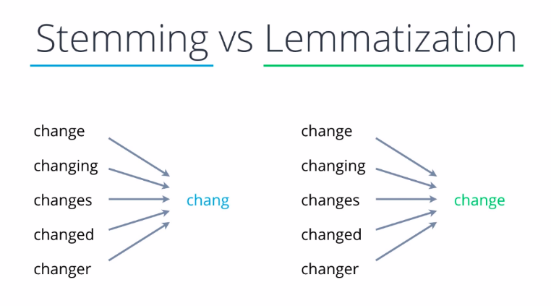
Texto

Descripción generada automáticamente

### Ajuste de Datos: Lenguaje

1. Stemming: Se reducen las palabras a lo que les da el significado, a su raíz.
2. Lemmatizing: Se busca que se encuentre una palabra con la cual se acepten las demás palabras flexionadas

Para mejor entendimiento se muestra un ejemplo de cada proceso:



Y el resultado obtenido:

Texto

Descripción generada automáticamente

### Bolsa de Palabras y Tf-Idf

Con los textos reducidos a las palabras se procede a procesar cada uno a una representación que los algoritmos entiendan. En sí, se va a usar TF-IDF, pero con el fin de lograr esa aproximación se empieza haciendo uso de una bolsa de palabras.

#### Bolsa de Palabras

Es una representación que convierte texto arbitrario en vectores de longitud fija contando cuántas veces aparece cada palabra. Se conoce como vectorización, y se realiza para cada registro; da como resultado una tabla de gran tamaño. (Zhou, 2019)

#### Tf-Idf

Term frequency-inverse document frequency es una medida que puede cuantificar la importancia o relevancia de las representaciones de cadenas (palabras, frases, lemas, etc.) en un documento entre una colección de documentos para evitar el sesgo de la longitud de los documentos (también conocido como corpus). (Capital One, 2021)

Con esta transformación los datos están listos para ser usados en cada uno de los modelos que se presentaron anteriormente. Se resalta por último que, dado que los algoritmos de clasificación funcionan mejor con datos balanceados en la variable de respuesta, se realizó un under sample con el fin de balancear de manera homogénea todas las categorías.

## Modelos de Clasificación

Los tres modelos usados con su respectivo ajuste de hiper-parámetros se encuentran a detalle en el Jupyter Nootbook anexado. Partiendo de la explicación ya dada, se van a mostrar los parámetros finales de cada modelo, así como su resultado.

Se especifica quién hizo cada uno:

1. Naïve Bayes Classifier – Santiago Bobadilla
2. Support Vector Machine – Juan José Beltrán
3. Árbol de Decisión – Ambos equitativamente

### Naive Bayes Classifier

Parámetros:

Texto

Descripción generada automáticamente

RESP: 0.642512077294686 (indica la proporción de predicciones correctas sobre el conjunto test)

Reporte:

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media

### Support Vector Machine

Parámetros:

Texto

Descripción generada automáticamente

RESP: 0.6570048309178744

Reporte:

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza media

### Árbol de Decisión

Parámetros:

Texto, Carta

Descripción generada automáticamente

RESP: 0.6417069243156199

Reporte:

Tabla

Descripción generada automáticamente

## Comparación

Si vemos las métricas generales del modelo obtenemos:

|  |  |
| --- | --- |
| Bayes | 0.642512 |
| **SVM** | **0.657005** |
| Árbol | 0.641707 |
|  |  |

Es decir, la mejor opción general es Vector Support Machine. No obstante, como se esta tratando el diagnóstico medico de pacientes (tema muy delicado) nos interesa de manera muy puntual el Recall, donde se busca sea lo mayor posible.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Recall | | | | | |
| Clase | 1 | 2 | 3 | 4 | 5 | Avg. |
| Bayes | 0.78 | **0.8** | 0.62 | 0.86 | 0.19 | 0.650 |
| Vector | **0.79** | 0.79 | **0.69** | 0.84 | **0.2** | **0.662** |
| Árbol | **0.79** | **0.8** | 0.62 | **0.87** | 0.15 | 0.646 |

Así, se reafirma que Vector Support Machine es la más adecuada. Gana en la mayoría de las categorías, y de manera global; en las que pierde es por poco, por tanto, es la mejor a usar. No obstante, las otras aproximaciones son bastante buenas, pero se le debe dar al modelo una alternativa concisa. Cabe mencionar que estas respuestas son dadas después de pasar el modelo por una validación que optimiza los hiper parámetros

### Recomendaciones

Usar Vector Support Machine en primera mano. Si la clasificación es 1 a 4, se considera confiable. No obstante, revisar las clasificadas como categoría 5. Dado que los otros modelos tampoco dan buenas métricas para la categoría 5, usar las otras aproximaciones tampoco vale la pena. De manera general, usar esta modelo solo como recomendaciones y herramienta de apoyo a la decisión. Lo que se sugiere no es verdad absoluta y vale la pena tener una segunda forma de revisión.

## Resultados

Se presentan de manera puntual en el tablero de control, y video; los cuales están anexados.

## Conclusiones

Se cumplió el objetivo general que fue lograr obtener un modelo que ayude al diagnóstico médico partiendo de resúmenes. Específicamente se obtuvo un modelo con una alta precisión, específicamente un recall aceptable. Por último, se da un procedimiento claro y lógico del procesamiento que se le debe dar a un resumen medico que llega.

Se concluye que se entrega un proyecto de calidad, que queda abierto a recomendaciones finales de los expertos.

## Anexos

1. Jupyter Notebook
2. Tablero de Control con los Resultados Específicos
3. Informe de Trabajo en grupo
4. Video
5. Datos modificados usados para la determinación de resultados

## Bibliografía

Capital One. 2021. Understanding TF-IDF for Machine Learning | Capital One. [online] Available at: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/> [Accessed 28 March 2022].

Kaggle.com. n.d. Medical Text. [online] Available at: <https://www.kaggle.com/datasets/chaitanyakck/medical-text?select=train.dat> [Accessed 28 March 2022].

Zhou, V., 2019. A Simple Explanation of the Bag-of-Words Model. [online] Medium. Available at: <https://towardsdatascience.com/a-simple-explanation-of-the-bag-of-words-model-b88fc4f4971> [Accessed 28 March 2022].