Proyecto 1 Inteligencia de Negocios

\* Sergio Molano

\* María Paula Téllez

\* Santiago Hernández

**Comprensión del negocio y enfoque analítico**

|  |  |  |  |
| --- | --- | --- | --- |
| Oportunidad/problema Negocio | | El cáncer es una enfermedad sumamente agresiva e intrusiva en el cuerpo, que en cuestión de años puede llevar a la muerte, debido a su severidad, muchos pacientes optan por hacer parte de ensayos clínicos, en su afán de remover esta enfermedad. Sin embargo, los ensayos clínicos no siempre son seguros, por lo que la solución, para algunos pacientes, podría volverse peor que la enfermedad. Es debido a esto, que es necesario poder determinar quien es apto para hacer parte de un ensayo clínico y quien no. Es necesario poder realizar esta selección de forma automática y normalizada para todos los pacientes ya que no sería justo que alguien no apto, sea elegido y sufra consecuencias complicadas o por el contrario alguien apto no sea elegido y no se le ofrezca la oportunidad de participar en el ensayo clínico que podría, eventualmente salvarle la vida. Frente a esta problemática, pensamos solventarla mediante el uso de técnicas de machine learning mediante el uso de datos previos de pacientes que fueron o no elegidos para estos ensayos. | |
| Descripción del requerimiento desde el punto de vista de aprendizaje de máquina | | Para poder solventar este problema, analizando los datos provistos, creemos que la mejor solución es utilizar un aprendizaje supervisado, ya que contamos con la fortuna de tener los datos etiquetados en si es apto o no para los ensayos, adicionalmente, vemos que la variable etiquetada es discreta debido a esto tenemos una tarea de clasificación, en este caso, clasificar si es o no, apto para los ensayos clínicos. Con lo anterior en cuenta, tenemos que lograr usar y entrenar un modelo de clasificación que sea capaz de recibir ciertas condiciones médicas y poder en base a eso, clasificar al paciente como apto o no apto. Para esto, se realizarán 3 diferentes modelos con 3 diferentes algoritmos de clasificación, para finalmente elegir el de mejor desempeño y poder usarlo para solventar esta problemática. | |
| Detalles de la actividad de minería de datos | | | |
| Tipo aprendizaje | Tarea de aprendizaje | | Algoritmo e hiper-parámetros utilizados (con la justificación respectiva) |
| aprendizaje supervisado | Clasificación con Arboles de decisión | | El algoritmo usado fue arboles de decisión. Los parámetros para considerar son el criterio de decisión en los nodos los cuales son gini o entrophy, y la profundidad máxima del árbol. |
| aprendizaje supervisado | Clasificación con knn | | El algoritmo usado fue KNearest-Neighbors. Los hiperparametros fueron n\_neighbors y p. |
| aprendizaje supervisado | Clasificación con random forest | | El algoritmo usado fue random forest. Los hiperparametros fueron gini o entrophy y la profundidad máxima. |

Se decidió usar la tarea de aprendizaje de clasificación debido a que se desea estimar una variable categórica, con la tarea de clasificación se puede construir un modelo predictivo que podrá ser aplicado para determinar, con una cierta precisión, la elegibilidad de un paciente para ensayos clínicos de cáncer a partir de texto descriptivo.

**Comprensión de los datos y preparación de los datos.**

Los datos obtenidos era un data set que contenía 2 columnas, label y study\_and\_condition, la primera era la variable etiquetada y la segundas la información sobre los estudios y el tipo de cáncer del paciente. La muestra de los datos tenía 12,000 registros y la longitud de la explicación tenía una media de 173 caracteres lo cual es bastante, demostrando que los mensajes eran bastante descriptivos, en la ilustración 1, se puede observar un poco más de cifras.

Una captura de pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza mediaLos registros tenían problemas en el formato, por ejemplo, el estudio y la condición venían en la misma columna, pero según las reglas de negocio, estaban separados por un punto (.), debido a esto, tuvimos que realizar de forma manual, la separación de esta única columna y dividirla en 2, mediante un Split en el punto:

Ilustración

Pasando de una tabla así:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

A la siguiente:

\_Texto

Descripción generada automáticamente con confianza baja

Texto

Descripción generada automáticamenteAdicionalmente a esto, podemos ver como la columna etiquetada ‘label’ tiene ruido:’\_\_label\_\_’ lo cual impide la conversión de esta columna a numérica, por esto, se decidió quitar esta parte, dejar únicamente los 1 y 0 y posteriormente pasar la columna a numérica:

Tabla

Descripción generada automáticamente con confianza media

Adicionalmente, revisamos la existencia de nulos, en este caso no había. Finalmente se generó un reporte y estos fueron los datos más relevantes:

Interfaz de usuario gráfica, Texto, Aplicación

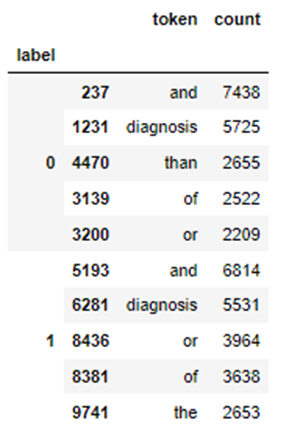
Descripción generada automáticamente

Para la preparación de datos, se tuvieron que realizar varias tareas debido a que se están manejando cadenas de texto y estas pueden presentar complicaciones al momento de pasarlas a un modelo. La primera limpieza que se realizó fue limpiar todos los caracteres no ascii ya que estos no aportan nada de valor y seguramente son un error que estén en los valores, se unificó el tema de mayúsculas y minúsculas para que palabras como Hola y hola no sean tomadas de forma diferente, se pasó todo a minúsculas, se removieron todos los signos de puntuación y los números dejaron de estar en formato numérico :6 a estar en la representación textual: seis.

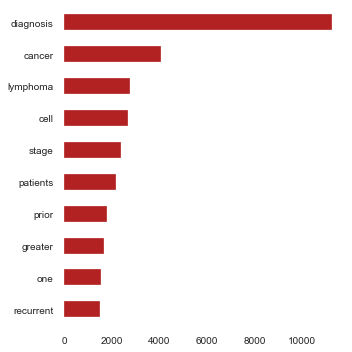
Otra parte fundamental para el procesamiento de texto no es únicamente la parte ortográfica y de presentación, es también muy importante revisar la parte semántica y el significado de las palabras, para esto, se realizó una transformación en donde se eliminaron todas las ‘stopwords’ que son palabras que son comunes y no se desean para el procesamiento de datos, estas están en un diccionario y únicamente se buscan para eliminarse. Otro cambio importante fueron las contracciones, en inglés, muchas personas hablan y escriben con contracciones, esto es un problema para el análisis ya que se podría tomar como 2 palabras diferentes cuando son lo mismo, por ejemplo: ‘i´ll’ y ‘i Will’, por eso se usó un diccionario para reemplazar todas las contracciones por su representación completa.

Finalmente, para manejar las diferentes formas o tiempos gramaticales y que queden de forma unificada se hicieron las transformaciones de ‘stem\_words’ para las palabras y ‘lemmatize\_verbs’ para los verbos. Dejando de esta forma el texto listo para los modelos.

Después de realizar los pasos de preparación de los datos se procede a analizar su comportamiento con el fin de hacerse una idea de los posibles resultados a obtener con los modelos que serán implementados más adelante. En este, se realiza un analisis de la frecuencia de las palabras, se hace un conteo de las palabras totales por etiqueta (label), las palabras distintas por paciente. Luego de esto, se observan las palabras más utilizadas en cada etiqueta, obteniendo el siguiente resultado en la ilustración 2.

Esto con el fin de identificar si son palabras relevantes a la hora de predecir si una persona puede ser seleccionada para un ensayo clínico contra el cáncer. En la tabla anterior puede observarse que los términos más frecuentes en todos los usuarios se corresponden con palabras que no aportan información relevante sobre el texto. A estas palabras se les conoce como stopwords. Estas palabras deben ser eliminadas para poder obtener un buen análisis de los datos. Luego de realizar esto, se presenta una gráfica con las palabras más usadas en los textos descriptivos, como se puede observar en la ilustración 3.

Ilustración

Luego se realiza un Tf-idf, la cual nos ayuda a saber cuan relevante es una palabra para un documento en una colección. En nuestro caso se creará un matriz tf-idf en la que cada columna es un término, cada fila un texto y el valor de intersección el tf-idf correspondiente. Esta matriz representa el espacio n-dimensional en el que se proyecta cada texto. Esta transformación de los datos es necesaria para pasar la información a los algoritmos de clasificación.

Ilustración

**Modelado y evaluación:**

A partir de la comprensión y la preparación de los datos se procede a aplicar los modelos de clasificación seleccionados para determinar si un paciente es elegible para un ensayo clínico para el tratamiento del cáncer a partir de un texto. Es importante mencionar que los datos se separarán en training set y test set.

**Árbol de clasificación (María Paula Tellez):**

Un árbol de clasificación es un modelo de predicción. Utiliza la medida de impurezas de Gini para clasificar los registros en las categorías del campo objetivo. Las predicciones se basan en combinaciones de valores en los campos de entrada. Se realiza una búsqueda con KFold y GridSearch para encontrar el mejor modelo. Al utilizar el algoritmo KFold, se define un k igual a 10, para que este divida los datos en 10 particiones y se obtiene un resultado de los mejores valores de hiperparámetros a continuación:



Utilizando estos hiperparámetros, se construye y se presentan el desempeño del modelo. Se obtienen las siguientes métricas sobre el conjunto de entrenamiento:

Texto

Descripción generada automáticamente

y las siguientes métricas sobre el conjunto de prueba:

Texto

Descripción generada automáticamente

Se puede observar que el modelo en general presenta unas métricas buenas, con una precisión de 0.70. Si bien no son las mejores, pueden ser aceptables para el negocio. De las métricas obtenidas, podemos ver que el modelo presenta un alto nivel de precisión, lo cual nos puede indicar que está en capacidad de generalizar, por lo que cuando el modelo diga que el paciente es elegible para el ensayo clínico del cáncer, esto será correcto el 70% del tiempo. A partir de las métricas de los conjuntos de prueba se puede observar que el modelo no está sobre ajustado, por lo que se puede usar para generalizar.

**KNearest-Neighbors (Sergio Molano):**

El método de los k vecinos más cercanos (en inglés, k-nearest neighbors, abreviado k-nn es un método de clasificación supervisada (Aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad F(x/Cj) de las predictoras x por cada clase C\_{j}.

Se realiza una búsqueda con KFold y GridSearch para encontrar el mejor modelo (se tienen en cuenta los hiperparámetros). Al realizar la búsqueda de los mejores hiperparametros, se obtiene como resultado los siguientes:



Al obtener el mejor modelo con estos hiperparametros, se obtienen los siguientes desempeños, sobre el conjunto de entrenamiento:

Texto

Descripción generada automáticamente

Y sobre el conjunto de prueba:

Texto

Descripción generada automáticamente

Como se puede observar el modelo creado usando el algoritmo KNN genero unas métricas aceptables para el negocio, aunque no son muy altas pueden ser usadas por el negocio, con una precisión de 0.86. A partir de las métricas de los conjuntos de prueba se puede observar que el modelo no está sobre ajustado, por lo que se puede usar para generalizar.

**Random Forest (Santiago Hernández):**

Random forest (o random forests) también conocidos en castellano como '"Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia. Se realiza una búsqueda con KFold y GridSearch para encontrar el mejor modelo. Al realizar el KFold y el GridSearch se obtienen los siguientes hiperparametros del mejor modelo:



Utilizando estos hiperparámetros, se construye y se presentan el desempeño del modelo. Se obtienen las siguientes métricas sobre el conjunto de entrenamiento:

Texto

Descripción generada automáticamente

Se obtienen las siguientes métricas sobre el conjunto de prueba:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Como se puede observar a partir de las métricas obtenidas usando el algoritmo Random Forest, el modelo presenta una alta precisión (0.82) y exactitud (0.82) , por lo cual puede ser usado por el negocio para determinar la elegibilidad de un paciente para ensayos clínicos de cáncer. Al obtener una precisión de 0.82 se dice que cuando el modelo prediga que un paciente es elegible para ensayos clínicos de cáncer, esto será correcto un 82% del tiempo. A partir de las métricas de los conjuntos de prueba se puede observar que el modelo no está sobre ajustado, por lo que se puede usar para generalizar.

**Resultados:**

Al obtener las métricas asociadas con cada uno de los algoritmos usados de clasificación, se llega a la conclusión de que en general, todos los modelos tienen un buen desempeño. Las métricas sobre los conjuntos de prueba nos indican que los modelos no están sobre ajustados, y es por esto que pueden ser usados para generalizar.

El modelo con una mayor exactitud fue Random Forest, lo que indica que es el mejor modelo para predecir la elegibilidad de un paciente para ensayos clínicos de cáncer con un 0.82% de exactitud. También, Random Forest fue el algoritmo que arrojo el modelo con mayor recall y precisión, esto quiere decir que es el modelo que predice con la menor cantidad de valores falsos negativos. A partir de las métricas obtenidas en cada uno de los modelos, se puede concluir que el mejor modelo es el Random Forest, es el modelo que puede ser usado por el negocio para determinar la elegibilidad de un paciente para ensayos clínicos de cáncer a partir de textos descriptivos. Utilizar textos descriptivos y construir modelos de predicción a partir de ellos, puede ser una ventaja competitiva para el negocio ya que como se observó a lo largo del proyecto, es posible utilizar un modelo como Random Forest para determinar la elegibilidad de un paciente para ensayos clínicos de cáncer a partir de texto descriptivo. Para concluir, se puede decir que el modelo Random Forest sería de gran utilidad para la ciencia a la hora de elegir los pacientes correctos para un ensayo clínico teniendo en cuenta un texto descriptivo. Teniendo en cuenta esto se podrían salvar muchas vidas si se agilizara el proceso de selección de individuos para encontrar una eventual cura del cáncer.

**Tablero de control:**

**Interfaz de usuario gráfica

Descripción generada automáticamente**

Trabajo en equipo:

Líder de proyecto: Sergio Molano

Líder de negocio: Santiago Hernández

Líder de datos: Maria Paula Tellez

Líder de analítica: Santiago Hernández

Cada uno de los integrantes del grupo aporto al proyecto de forma equitativa. Hicimos una reunión para definir como nos repartiríamos el trabajo. Todos realizamos un análisis inicial de que datos escogeríamos y porque, y como utilizaríamos los modelos vistos en clase sobre textos descriptivos, aproximadamente duramos 3 horas realizando esto. La limpieza y preparación de los datos la hicimos todos en una reunión, en la cual tardamos 5 horas y cada uno luego realizo su propio algoritmo en lo cual cada uno tardo 3 horas aproximadamente. Luego de esto, cada uno se encargó en su mayoría de alguna entregable, Santiago del tablero de control y del video, Maria Paula del documento y Sergio de la presentación.

# Referencias

AURIML. (2018). *Clinical Trials on Cancer*. Obtenido de Kaggle: https://www.kaggle.com/datasets/auriml/eligibilityforcancerclinicaltrials

-https://es.wikipedia.org/wiki/Tf-idf#:~:text=Tf%2Didf%20(del%20ingl%C3%A9s%20Term,un%20documento%20en%20una%20colecci%C3%B3n.

- https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-classification-tree

- https://es.wikipedia.org/wiki/K\_vecinos\_m%C3%A1s\_pr%C3%B3ximos

- https://es.wikipedia.org/wiki/Random\_forest