Predicción de las Causas de los Incendios Forestales en EE.UU.

Francisco Muñoz Guajardo
Departamento de Ingeniería Matemática
Universidad de Chile, FCFM

***Abstract*—En este informe se explorarán métodos para predecir las causas que producen un incendio dados algunos parámetros del mismo, como lo es el origen geográfico del incendio, la fecha en que ocurrió, la extensión del incendio, entre otras cosas.**

Se probarán algoritmos de clasificación, tales como *Naïve Bayes*, *KNN*, *Discriminante Lineal/Cuadrático* y *Random Forest* en distintos escenarios. El resultado de los experimentos son exactitudes que bordean el $50-60\%$.

I. Introducción

A. Contexto y Motivación

En el último tiempo se ha visto que han ocurrido un gran número de incendios en todo el mundo. Cómo lo es en el caso de Chile, el Amazonas (Brasil) o Australia. Es por esto, que se plantea utilizar las herramientas de Aprendizaje de Máquinas para poder ayudar a combatir y prevenir los incendios que están aconteciendo actualmente.

En este informe se trabajará con un *dataset* que cuenta con aproximadamente dos millones de incendios ocurridos en EE.UU. de forma histórica. Una descripción más detallada acerca de los datos que se ocuparán se puede encontrar en II.

B. Preguntas realizadas

Dadas algunas propiedades intrínsecas de un incendio, como lo es su ubicación geográfica, la fecha, hora de ocurrencia y algunos otros factores que se detallarán más adelante,

- ¿Es posible predecir las causas de un incendio?
- ¿Es posible saber si un incendio es causado por error humano, por la naturaleza o con fines maliciosos?

C. Objetivos

Realizadas las preguntas a responder, se propondrán los siguientes objetivos:

- Predecir las causas específicas de un incendio dados los datos que nos entrega el *dataset*.
- Predecir las causas generales de un incendio dados los datos que nos entrega el *dataset*.

D. Resumen del trabajo realizado en la primera presentación

En la primera presentación se empezó a buscar algún proyecto interesante, en el que se pueda aplicar las herramientas aprendidas en el curso.

Escogido el tema, se planteó la pregunta: "*¿Se puede predecir el área total y extensión de un incendio?*" Esta pregunta fue reemplazada por la pregunta "*¿Es posible predecir las causas de un incendio?*". La razón de este reemplazo es que se consideraba que para la primera pregunta se necesitan datos meteorológicos, los cuales no se cuentan actualmente o son muy difíciles de obtener.

Luego, se empezaron a buscar base de datos interesantes, de forma que pueda responder a las preguntas. Finalmente, se encontraron dos bases de datos: la correspondiente a los incendios de EE.UU. (la cual se encuentra bastante detallada) y la correspondiente a la de CONAF, que es chilena. Finalmente, se prefirió la base de datos de EE.UU., pues la base de datos chilena posee pocos datos, además de que solamente se encuentra disponible en Excel, dificultando la extracción y manipulación de los datos.

E. Resumen del trabajo realizado en la segunda presentación

En la segunda presentación, se seleccionaron las columnas que se consideraran relevantes de la base de datos de EE.UU. Paso seguido, se empezó a preparar el *dataset* para su posterior manipulación, eliminando algunos datos que no fueran importantes, modificando el formato de los datos y agregando nuevas columnas al *dataset*.

Luego, se realizó un análisis exploratorio de los datos, observando la distribución del número de incendios con respecto a las causas, a los estados (de EE.UU.), a los días de la semana y con respecto a la hora. Luego se intentó observar la correlación entre las columnas.

Finalmente, se implementaron algunos modelos básicos ocupando la base de datos, como lo es *Naïve Bayes*, *K-Vecinos más Cercanos*, *Random Forest*, entre otros.

II. Descripción de los Datos

A. Descripción preliminar de los datos

Se cuenta con un *dataset*, obtenido de Kaggle, con un registro de 1.880.465 incendios históricos ocurridos en Estados Unidos, desde 1992 hasta 2015. La información original se obtuvo a partir del *U.S. Department of Agriculture* [1].

El *dataset* se encuentra en extensión *.sqlite*, de donde ocuparemos la tabla "*Fires*" que posee en total 37 columnas, dentro de las cuales, existen columnas para describir el ID (como lo es ***FOD_ID*** O ***FPA_ID***), las siglas (como lo es ***STATE***), los nombres (como lo es ***SOURCE_SYSTEM_TYPE***), algunos parámetros numéricos (como lo es ***FIRE_SIZE***), parámetros temporales (como lo es ***CONT_DATE*** o ***CONT_TIME***),

parámetros espaciales (como lo es ***LATITUDE***), entre otros[1]. Dentro de las cuales, la columna de etiquetas (que en nuestro caso viene a ser lo mismo que las causas) que se desea predecir es la llamada ***STAT_CAUSE_DESCR***.

Para más información acerca de las columnas del *dataset*, se puede consultar la documentación proporcionada en la página de Kaggle.

B. Descripción de las columnas a ocupar

Del total de columnas que posee la tabla, ocuparemos aquellas que estén relacionadas con la fecha (como la hora y la fecha estimada en que se inició el incendio y el momento en que se controló), la geografía (como las siglas del estado en que se originó el incendio, y la latitud y longitud del incendio), el tamaño del incendio y la descripción de la causa del incendio (que es la variable objetivo a predecir).

El resto de variables no representan información relevante para predecir las causas de un incendio. Por ejemplo, el ID del incendio, o del departamento que logró controlar el incendió, no aporta información importante a la causa del incendio, pues no es una propiedad "*natural*" del mismo (es más, puede resultar en un sesgo a la hora de predecir).

A continuación se describirán brevemente las columnas que se ocuparán:

1) ***FIRE_YEAR***: Corresponde al año en que se originó el incendio. Es una variable temporal.
2) ***DISCOVERY_DATE***: Es la fecha (estimada) en que se originó el incendio. Es una variable temporal que inicialmente es del tipo float.
3) ***DISCOVERY_TIME***: Es la hora (estimada) en que se originó el incendio. Es una variable temporal que inicialmente es de tipo cadena.
4) ***STAT_CAUSE_DESCR***: Es la causa que originó el incendio. Es una variable categórica de tipo cadena.
5) ***CONT_DATE***: Es la fecha en que se logró extinguir el incendio. Es una variable temporal que inicialmente es del tipo float.
6) ***CONT_TIME***: Es la hora en que se logró extinguir el incendio. Es una variable temporal que inicialmente es de tipo cadena.
7) ***FIRE_SIZE***: Es la extensión (área) que tuvo el incendio (medido en acres[2]). Es una variable numérica.
8) ***FIRE_SIZE_CLASS***: Es una clasificación que utilizan los bomberos para describir la extensión del incendio. Es una variable categórica.
9) ***LATITUDE***: Es la latitud del incendio. Es una variable numérica.
10) ***LONGITUDE***: Es la longitud del incendio. Es una variable numérica.
11) ***STATE***: Es el estado en que se originó el incendio. Es una variable categórica.

[1]Que no viene al caso describir cada una de las columnas del *dataset*, dado a que posee una cantidad abrumadora de columnas.

[2]Un acre corresponde a 4046.85 metros cuadrados.

C. Descripción del dataset con las columnas a ocupar

Restringido a las columnas mencionadas anteriormente, se observa que el *dataset* posee un total de 3.535 filas duplicadas.

También se observa que las siguientes columnas poseen valores nulos:

- ***DISCOVERY_TIME***, con un total de 882.638 valores nulos,
- ***CONT_DATE***, con un total de 891.531 valores nulos, y
- ***CONT_TIME***, con un total de 972.173 valores nulos.

Además, se observa que las columnas relacionadas con las fechas se encuentran en formato de fecha juliana[3] y que las columnas relacionadas a las horas son cadenas, se encuentran en formato "HHMM", donde HH corresponde al número de horas y MM corresponde al número de minutos.

D. Descripción de las etiquetas de la columna objetivo

A continuación, describiremos brevemente las etiquetas que posee la columna que se desea predecir:

- **Miscellaneous**: de causa Miscelánea.
- **Children**: Causado por infantes.
- **Lightning**: Causado por un rayo.
- **Smoking**: Causado por un cigarrillo mal apagado.
- **Arson**: Causado con fines maliciosos.
- **Equipment Use**: Causado por el uso de un equipamiento.
- **Debris Burning**: Causado por quema de escombros.
- **Campfire**: Causado por un campamento.
- **Railroad**: Causado por un ferrocarril.
- **Missing/Undefined**: Que tiene datos faltantes o que no están definidos.
- **Powerline**: Causado por la linea eléctrica.
- **Fireworks**: Causado por fuegos artificiales.
- **Structure**: Causado por Estructura.

Cabe destacar, que en la documentación de la base de datos no aparece una descripción detallada de las categorías, con lo que, varias de las descripciones de estas causas fueron inferidas por el autor.

III. Tratamiento de los Datos

A. Limpieza preliminar de los datos

Como se comentó en la sección II-C, el *dataset* posee algunos datos duplicados o con valores nulos. Para remediar esto, se procede a eliminar estas filas, resultando en un *dataset* de 890.821 filas, aproximadamente la mitad de datos que poseía el *dataset* originalmente.

B. Tratamiento de las columnas de fecha y hora

Como se comentó en la sección II-C, las fechas se encuentran en formato de fecha juliana, y el tiempo se encuentra en formato "HHMM".

Para tratar con estos formatos, primero se transforma las fechas en formato gregoriano[4], y las horas se transforman

[3]Esto corresponde al número de días transcurridos desde el mediodía del 1° de enero del año 4713 a. C.

[4]Es decir, en un formato AAAA-MM-DD, donde AAAA es el año, MM es el mes y DD es el día.

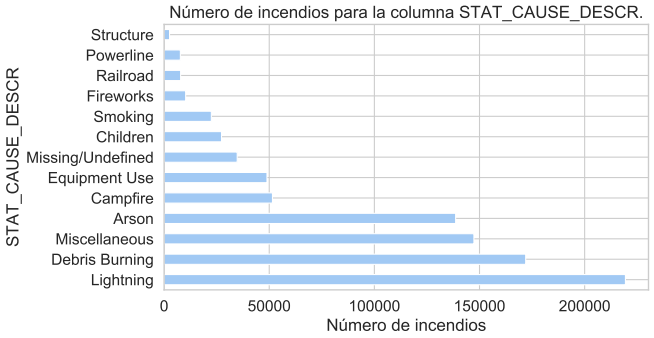

Fig. 1. Gráfico de Causas vs. Número de incendios.

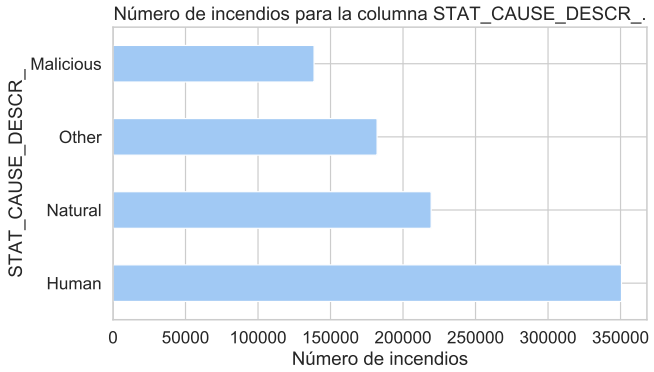

Fig. 2. Gráfico de Causas Nuevas vs. Número de incendios.

al siguiente formato: "HH:MM:SS", con SS los segundos (donde como no se posee la información de los segundos, se rellenan con 00). Paso seguido, se utiliza la fecha y la hora para transforlas en datos del tipo *DateTime*. Esto se realiza para las columnas de sufijo ***DISCOVERY_**** y ***CONT_****. De esta forma, cuatro de las columnas que se tenían originalmente, se resumen en dos columnas, las cuales se nombraron ***DISC_DATE_TIME*** y ***CONT_DATE_TIME***.

C. Creación de nuevas columnas

A partir de las nuevas columnas creadas, {***DISC,CONT***}***_DATE_TIME***, se crean las siguientes columnas:

- {***DISC,CONT***}***_MONTH***: Columna que indica el mes de descubrimiento/contención del incendio.
- {***DISC,CONT***}***_DAY_OF_WEEK***: Columna que indica el día de la semana del descubrimiento/contención del incendio.
- ***DT_FIRE***: Columna que indica el número de horas utilizadas en controlar el incendio.

D. Ordenación del dataset con respecto a la fecha

Por razones que se explicarán más adelante (específicamente en la sección V-C), el *dataset* se organiza con respecto a la columna ***DISC_DATE_TIME***, es decir, se organiza con respecto a la fecha y a la hora.

E. Creación de nuevas categorías

Más adelante se observará que existen categorías bastante mayoritarias frente a otras. Para solucionar este problema, se propondrá una nueva categorización, que consiste en crear cuatro nuevas categorías:

- **Natural**: Causas naturales. **Lightning** entra en esta categoría.
- **Malicious**: Causas maliciosas. **Arson** entra en esta categoría.
- **Other**: Causas que no se especifican bien. **Miscellaneous** y **Missing/Undefined** entran en esta categoría
- **Human**: Hecha por causas humanas, y que probablemente fueron accidente. El resto de las causas que no han sido mencionadas entran en esta categoría.

IV. Análisis Exploratorio de los Datos

A. Distribución de las causas de un incendio

La primera pregunta que se plantea es: ¿Cómo están distribuidos el número de incendios con respecto a las causas del incendio? Para responder a la interrogante, se procede a graficar el número de incendios con respecto a sus causas. El resultado del gráfico se puede observar en la Fig. 1.

Se puede observar que las principales causas de incendio son producidos por rayos (**Lightning**), quema de escombros (**Debris Burning**), de forma miscelánea (**Miscellaneous**) y por razones maliciosas (**Arson**). Mientras que el resto de causas tienen un papel minoritario.

También, otra de las primeras observaciones que se pueden rescatar, es que el *dataset* se encuentra *desequilibrado* (i.e., que existen clases bastante más númerosas que otras). Esto causará que a la hora de realizar clasificación, las clases con mayor número de incendios (como **Lightning**) queden beneficiadas frente a las clases minoritarias (como **Structure**), resultando en *overfitting*.

Para solucionar este problema, se crearán cuatro nuevas categorías, como se explica en III-E. estas son: **Natural**, **Malicious**, **Other** y **Human**. En la Fig. 2 se puede observar la nueva distribución que resulta de la creación de las nuevas categorías.

Se puede observar a primera vista que la nueva categorización resulta en un *dataset* más balanceado, y que además, la causa humana resulta ser la nueva clase mayoritaria.

B. Distribución de los incendios a lo largo de los años

Otra interrogante que se puede plantear es: ¿Cómo se distribuyen los incendios a lo largo de los años? Para responder a esta pregunta, se elaboró un gráfico del número de incendios vs. el año de ocurrencia. El resultado del gráfico se puede observar en la Fig. 3.

Las primeras apreciaciones que se observan es que en los últimos años (desde el 2011 hasta el 2015) se tuvo un alza importante de incendios. Quizás, esto se puede deber a que en los últimos años se cambió la forma en que se contabilizaban los incendios, o quizás que en esos años ocurrieron incendios con el interés de ocupar el terreno.

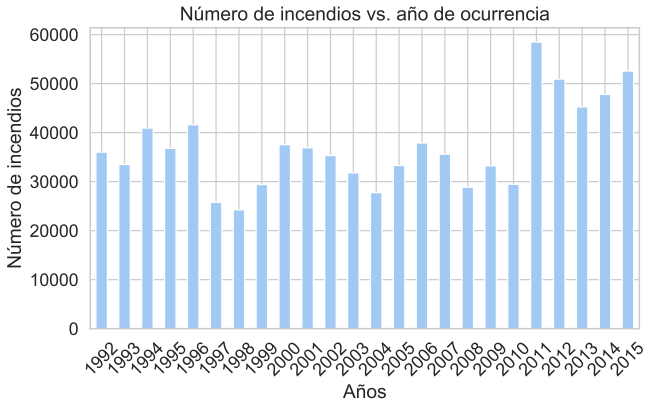

Fig. 3. Gráfico de Número de incendios vs. Año de ocurrencia.

Fig. 4. Número de las causas (menores) de los incendios a lo largo de los años.

Además, también se observa que hay años que son minoría comparativamente, como por ejemplo, en el año 2011 ocurrieron más del doble de incendios que en el año 1998.

Es importante tomar en cuenta estas observaciones, pues puede afectar en la predicción del clasificador en gran medida, pues este no estará consciente de las "rachas" de incendios que ocurrieron en los últimos años.

Luego, una interrogante natural que surge es: ¿Cómo se distribuyen las causas de los incendios a lo largo de los años? Para responder a esta pregunta, se elaboran los gráficos para las causas con un número de incendios bajo, medio y alto. Estos gráficos se encuentran en las Figuras 4, 5 y 6 respectivamente.

Se puede observar que existen varias causas de incendio que no se mantienen constantes a lo largo de los años. Un caso como esto es la causa **Missing/Undefined**, que en los primeros años ocurren varios incendios en que entraban en esta categoría, luego, desde el año 1997 hasta el año 2010 no hay mucha actividad hasta que en el año 2011, es una de las mayores causas (entre **Smoking**, **Children**, **Equipment Use** y **Campfire**) de incendio.

Estos comportamientos se puede deber a la forma en cómo las agencias contabilizaban los incendios a lo largo de los años. Sin embargo, es muy importante tener en cuenta estas irregu-

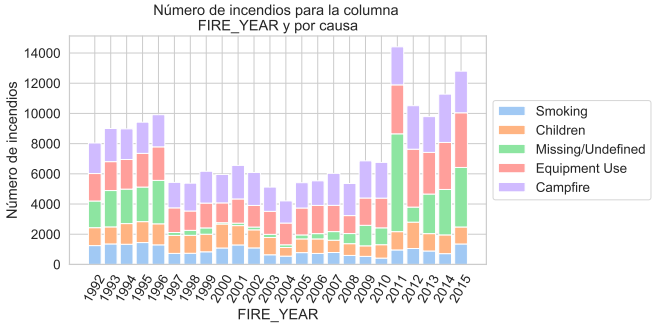

Fig. 5. Número de las causas (medias) de los incendios a lo largo de los años.

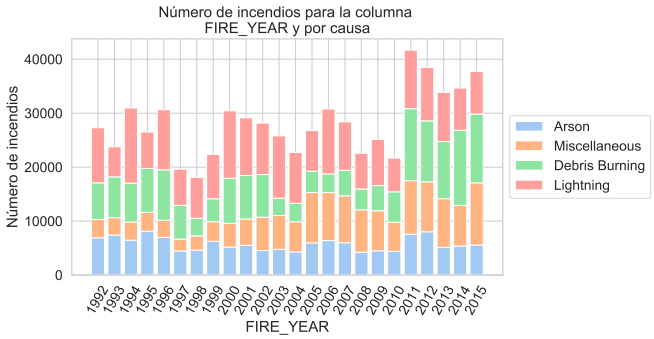

Fig. 6. Número de las causas (mayores) de los incendios a lo largo de los años.

laridades temporales, pues los modelos se pueden "memorizar" patrones para ciertos años, para después cambiar radicalmente su comportamiento, deteriorando las predicciones.

V. Metodología

A. Modelos

Se proponen cinco modelos para la clasificación, escogidos en base a los algoritmos vistos en el curso. Estos son: *Naïve Bayes* (NB), *K-Vecinos más Cercanos* (KNN), *Discriminante Lineal* (DL), *Discriminante Cuadrático* (DC) y *Random Forest* (RF).

1) Preparación del modelo de Naïve Bayes: Para el modelo de *Naïve Bayes*, no se seleccionarán hiperparámetros. Se asumirá un prior uniforme, pues de todas formas, se considera que se tiene una gran cantidad de datos como para que el prior se vuelva irrelevante.

2) Preparación del modelo de KNN: Para el modelo de los *K-Vecinos más Cercanos*, se empezará por normalizar los datos y escogiendo con anterioridad un k óptimo, de forma heurística.

La metodología para encontrar este k óptimo, es probando con una lista finita de valores tentativos (e.g.: $k \in \{1, 10, 50, 100, 150, 200\}$) y evaluando cuál posee el mayor *score*. Realizado esto, se procede a realizar una nueva iteración en la vecindad en que se encontraba el k anterior (e.g.: si resultó que $k = 50$ en el ejemplo anterior, entonces se probará en un conjunto acotado que esté entre 10 y 100).

3) Preparación del modelo lineal/cuadrático: No se realizará una preparación previa para estos modelos, ni se ajustarán sus hiperparámetros.

4) Preparación del modelo de Random Forest: Dada a la gran cantidad de datos que posee el dataset, y al gran costo computacional que supone este algoritmo, resulta muy difícil ajustar los hiperparámetros que posee, con lo que se prefiere conservar los hiperparámetros que viene por defecto con el modelo.

B. Escenarios

Se propondrá distintos escenarios en los que se aplicarán los cinco modelos descritos anteriormente.

1) Todos los datos: En este escenario se considerarán todos los datos que contiene el *dataset* posterior a su depuración, sin realizar un "preprocesamiento". Se realizará esto para comparar la eficiencia que poseen los otros escenarios frente a "no realizar nada".

2) A partir del año 2011: Como se observó en la Fig. 3, ocurre una alza importante en estos años, dando una regularidad significativa frente a los años anteriores. Es por esto que se probarán los distintos modelos en el escenario de los últimos años a partir del 2011.

3) Agrupación de las categorías: Cómo se habló en la sección IV, existe un desbalance importante en las categorías de las causas de un incendio. Es por esto que se considerará en realizar una recategorización de los datos y además, al igual que en el escenario anterior, se considerarán los últimos años a partir de 2011.

C. Preparación del Conjunto de Entrenamiento y Test

Dado que los datos poseen una dependencia temporal, se dividirá un conjunto de entrenamiento y un conjunto de test de forma que el primer 90% de los datos serán asignados para el entrenamiento y el 10% restantes serán asignados para el test (recordando que los datos se encuentran ordenados por fecha, como se menciono en la sección III-D).

La razón de esta división es que se busca predecir el futuro en base al pasado, y si simplemente se toma una muestra aleatoria del *dataset*, si después se busca predecir una causa, puede "aprenderse" la respuesta conociendo el futuro para predecir la causa. Es más, esta fue una falla que se tuvo en la primera presentación.

Mientras que la razón de esta proporción $90/10$ es que, si es que el modelo se llegase a utilizar para predecir una causa real, es esperable que se utilicen todos los datos históricos para predecir una nueva causa.

VI. Resultados

El resumen de las exactitudes (*accuracy*) de cada uno de los modelos frente a cada uno de los escenarios se encuentra en la Tabla I.

Dada a la gran cantidad de modelos que se manejan, se adjuntarán los reportes de clasificación de cada uno de los modelos por cada uno de los escenarios que se encuentra. Los reportes para el Escenario 1, Escenario 2 y Escenario 3 se pueden encontrar en el Apéndice A.

TABLA I
TABLA RESUMEN DE LOS MODELOS

Modelos	Escenario 1	Escenario 2	Escenario 3
Naive Bayes	0.2215	0.1400	0.3574
Discriminante Cuadrático	0.2451	0.1512	0.3870
Discriminante Lineal	0.3465	0.3233	0.4210
KNN	0.3978	0.4266	0.5459
Random Forest	0.4562	0.5277	0.6467

VII. Discusión y Conclusiones

En términos generales, se observa de la Tabla I que en los tres Escenarios, el modelo con la mayor exactitud corresponde a RF, seguido del modelo de KNN y del DL. Mientras que los modelos con menor desempeño se encuentra el DC y NB, siendo este último el modelo con el peor desempeño.

Que el modelo del DL tenga un desempeño "decente" resulta algo sorprendente, pues indica que las causas se pueden separar a través de hiperplanos (o mejor dicho, que las regiones de decisión correspondan a poliedros). Esto parece indicar que existe cierto comportamiento lineal entre las causas de un incendio.

Igualmente, resulta sorprendente que el modelo de DC tenga en general un peor desempeño que el DL, pues este modelo supone ser más flexible que el DL y que en general, contiene de manera natural al modelo de DL.

A continuación se realizarán las conclusiones para cada uno de los Escenarios:

A. Conclusiones acerca del Escenario 1

Podemos observar que en el primer escenario, las predicciones de la mayoría de los modelos son insatisfactorias. Inclusive el modelo de RF, que posee un 45.6% de exactitud, es un porcentaje relativamente bajo.

Esto se puede deber a que el modelo se aprendió ciertos patrones del pasado (digamos, de los años 1992 a 2010) para después ver que en los años futuros, donde todo el escenario cambió radicalmente, como se comentó en la sección IV-B.

Además, de las Tablas II hasta la VI, se observa que por lo usual, la causa **Lightning** es la que posee mayor *recall*. Es lo esperable, pues es la causa con mayor número de incendios, y para los modelos les resulta más fácil decir que pertenece a la clase mayoritaria que a las clases minoritarias.

Comentando además, que en varios de los modelos, las causas minoritarias vagamente se encuentran predichas. Como ocurre en el caso de **Railroad**, **Smoking** o **Structure**.

En conclusión, el Escenario 1 es bastante susceptible a que sufra de *overfitting*.

B. Conclusiones acerca del Escenario 2

Se observa claramente que la exactitud mejoró considerablemente para los modelos de RF y KNN al considerar los últimos años. Sin embargo, los modelos de DL, DC y NB tuvieron una disminución en la exactitud con respecto al Escenario 1, en especial los modelos de DC y NB.

Sin embargo, observando las Tablas VII hasta la XI, se observa que este Escenario sufre el mismo problema que el Escenario 1: las causas mayoritarias se predicen bastante bien, mientras que las causas minoritarias prácticamente no se predicen. Esto ocurre en prácticamente todos los modelos.

C. Conclusiones acerca del Escenario 3

Se observa que la exactitud mejoró considerablemente en todos los modelos con respecto a los Escenarios 1 y 2. Esto se puede deber, como se comentó en la sección IV-A, a que esta nueva categorización se encuentra mejor equilibrada que el dataset anterior.

Observando la Tabla XII, RF logra un *recall* del 78.1% para las causas naturales, frente a un 68.4% de *recall* para la causa humana. Esto resulta sorprendente, pues para este *dataset*, la causa humana es la categoría mayoritaria, con lo que se esperaría que esta causa terminara con un *recall* mayor.

Otra razón por la este hecho resulte interesante, es que esto significa que este modelo predice bastante bien cuando un incendio es causado por un rayo (pues la causa natural solo posee la causa **Lightning**), aunque esto se puede deber a que en ciertos estados de EE.UU. sea más recurrente que ocurran incendios por rayos, y que el modelo simplemente se haya aprendido estos patrones.

Además, cabe mencionar que las causas maliciosas poseen un *recall* del 28%, el cual resulta bastante bajo, sin embargo, posee una *precisión* del 55.4%, lo cual es bastante decente. Esto último se puede interpretar como, a pesar de que el modelo no predice muy bien esta clase, cuando lo hace, es relativamente confiable (pues una *precisión* del 50% sigue siendo un valor bajo).

Además, de la Tabla XIV, observamos que el modelo de DL entrega una *precisión* y *recall* nula para la causa maliciosa. Esto resulta interesante, pues se puede interpretar como, a pesar de que las razones humanas, naturales y otras causas se pueden clasificar por medio de poliedros, es difícil clasificar a las causas maliciosas dentro de un poliedro, siendo este inexistente.

Finalmente, observemos que, a partir de las Tablas XV y XVI, los modelos de DC y NB poseen un *recall* muy alto (cerca del 90% en ambos casos), pero una *precisión* baja (cerca del 40%). Esto se puede deber a que, como esta es la categoría mayoritaria, los modelos simplemente se aprendieron sus patrones respectivos, considerando vagamente al resto de categorías.

D. Comentarios finales

En conclusión, la respuesta a *"¿Es posible predecir las causas de un incendio?"* resultó insatisfactoria. Pues a pesar de que el modelo de RF para el Escenario 2 posea un 52.8% de exactitud, esta cantidad aún resulta insuficiente para afirmar que es posible predecir las causas de un incendio. Sin embargo, los resultados del experimento resultaron ser bastantes modestos.

Mientras que la respuesta a la pregunta "*¿Es posible saber si un incendio es causado por error humano, por la naturaleza o con fines maliciosos?*", la respuesta aún sigue inconclusa. Pues el modelo de RF para el tercer escenario, a pesar de poseer un *f1-score* del 64.2% para las causas humanas, y un 76.7% para las causas naturales (resultados que son bastantes buenos) aún sigue siendo insatisfactorio la predicción de las causas maliciosas, que posee un *f1-score* del 37.2%.

VIII. Futuras mejoras

Quizás, reuniendo imágenes satélites de incendios, reuniendo datos atmosféricos y utilizando métodos más sofisticados de Aprendizaje de Máquinas, se pueda responder a una de las preguntas que incentivaron este trabajo: *¿Es posible predecir cuánto se extenderá un incendio?* En este trabajo no se abordó esta pregunta, pues se consideró que no se poseían los datos ni las herramientas suficientes para responder a esta interrogante.

Otra pregunta que resulta interesante a responder, pero por tema de tiempo no se logró contestar es la siguiente: *¿Se puede predecir si un incendio fue causado con fines maliciosos?* Quizás, ocupando el Discriminante Lineal de Fisher se puede responder satisfactoriamente a esta pregunta.

Visto que reducir la cantidad de años mejoró considerablemente los modelos de predicción, quizás si se empiezan a probar distintos intervalos de años, se pueda mejorar el rendimiento de los modelos.

Por último, quizás se podría crear un escenario en dónde se intente predecir únicamente las causas humanas, pues los modelos de DC y NB del Escenario 3 poseen un *recall* bastante alto, en donde lo que se podría intentar hacer, es primero predecir si un incendio fue por causas humanas o no, y luego intentar predecir cuál de todas las posibles causas humanas corresponde esta causa.

Referencias

[1] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

APÉNDICE A
REPORTES DE CLASIFICACIÓN

A continuación se presentarán los distintos reportes de clasificación por cada modelo y por cada escenario descrito:

A. Escenario 1

Los resultados de la clasificación en el Escenario 1 utilizando *Random Forest*, *K-Vecinos más cercanos*, *Discriminante lineal*, *Discriminante Cuadrático* y *Naïve Bayes* se encuentran en las Tablas II, III, IV, V y VI respectivamente.

B. Escenario 2

Los resultados de la clasificación en el Escenario 2 utilizando *Random Forest*, *K-Vecinos más cercanos*, *Discriminante lineal*, *Discriminante Cuadrático* y *Naïve Bayes* se encuentran en las Tablas VII, VIII, IX, X y XI respectivamente.

C. Escenario 3

Los resultados de la clasificación en el Escenario 3 utilizando *Random Forest*, *K-Vecinos más cercanos*, *Discriminante lineal*, *Discriminante Cuadrático* y *Naïve Bayes* se encuentran en las Tablas XII, XIII, XIV, XV y XVI respectivamente.

TABLA II
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 1 UTILIZANDO RANDOM FOREST

	precision	recall	f1-score
Arson	0.283	0.358	0.316
Campfire	0.348	0.251	0.291
Children	0.236	0.110	0.150
Debris Burning	0.537	0.584	0.560
Equipment Use	0.236	0.133	0.170
Fireworks	0.510	0.308	0.384
Lightning	0.650	0.839	0.733
Miscellaneous	0.394	0.459	0.424
Missing/Undefined	0.239	0.160	0.192
Powerline	0.102	0.004	0.008
Railroad	0.128	0.017	0.031
Smoking	0.138	0.022	0.037
Structure	0.258	0.019	0.035
accuracy			0.456
macro avg	0.312	0.251	0.256
weighted avg	0.421	0.456	0.430

TABLA III
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 1 UTILIZANDO KNN

	precision	recall	f1-score
Arson	0.217	0.318	0.258
Campfire	0.183	0.113	0.140
Children	0.187	0.078	0.110
Debris Burning	0.479	0.553	0.513
Equipment Use	0.187	0.047	0.074
Fireworks	0.346	0.233	0.279
Lightning	0.524	0.822	0.640
Miscellaneous	0.372	0.379	0.375
Missing/Undefined	0.003	0.001	0.002
Powerline	0.222	0.001	0.003
Railroad	0.035	0.005	0.008
Smoking	0.000	0.000	0.000
Structure	0.000	0.000	0.000
accuracy			0.398
macro avg	0.212	0.196	0.185
weighted avg	0.338	0.398	0.354

TABLA IV
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 1 UTILIZANDO DISCRIMINANTE LINEAL

	precision	recall	f1-score
Arson	0.280	0.082	0.126
Campfire	0.026	0.000	0.000
Children	0.000	0.000	0.000
Debris Burning	0.389	0.658	0.489
Equipment Use	0.000	0.000	0.000
Fireworks	0.000	0.000	0.000
Lightning	0.334	0.811	0.473
Miscellaneous	0.280	0.201	0.234
Missing/Undefined	0.000	0.000	0.000
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.000	0.000	0.000
accuracy			0.346
macro avg	0.101	0.135	0.102
weighted avg	0.237	0.346	0.258

TABLA V
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 1 UTILIZANDO DISCRIMINANTE CUADRÁTICO

	precision	recall	f1-score
Arson	0.201	0.110	0.142
Campfire	0.169	0.039	0.064
Children	0.030	0.449	0.057
Debris Burning	0.405	0.705	0.514
Equipment Use	0.000	0.000	0.000
Fireworks	0.067	0.438	0.116
Lightning	0.565	0.220	0.316
Miscellaneous	0.467	0.051	0.092
Missing/Undefined	0.000	0.000	0.000
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.033	0.014	0.020
accuracy			0.245
macro avg	0.149	0.156	0.102
weighted avg	0.319	0.245	0.217

TABLA VI
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 1 UTILIZANDO NAÏVE BAYES

	precision	recall	f1-score
Arson	0.185	0.103	0.132
Campfire	0.097	0.068	0.080
Children	0.033	0.422	0.061
Debris Burning	0.371	0.706	0.486
Equipment Use	0.000	0.000	0.000
Fireworks	0.066	0.540	0.118
Lightning	0.500	0.087	0.148
Miscellaneous	0.433	0.041	0.076
Missing/Undefined	0.333	0.000	0.000
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.037	0.028	0.032
accuracy			0.221
macro avg	0.158	0.153	0.087
weighted avg	0.312	0.221	0.177

TABLA VII
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 2 UTILIZANDO RANDOM FOREST

	precision	recall	f1-score
Arson	0.419	0.359	0.386
Campfire	0.377	0.340	0.358
Children	0.194	0.071	0.104
Debris Burning	0.453	0.580	0.509
Equipment Use	0.327	0.133	0.189
Fireworks	0.592	0.360	0.448
Lightning	0.704	0.835	0.764
Miscellaneous	0.458	0.532	0.492
Missing/Undefined	0.566	0.552	0.559
Powerline	0.091	0.007	0.013
Railroad	0.517	0.161	0.246
Smoking	0.301	0.042	0.073
Structure	0.200	0.017	0.031
accuracy			0.528
macro avg	0.400	0.307	0.321
weighted avg	0.499	0.528	0.502

TABLA VIII
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 2 UTILIZANDO KNN

	precision	recall	f1-score
Arson	0.329	0.251	0.285
Campfire	0.200	0.112	0.144
Children	0.000	0.000	0.000
Debris Burning	0.342	0.576	0.429
Equipment Use	0.278	0.057	0.095
Fireworks	0.345	0.203	0.256
Lightning	0.553	0.803	0.655
Miscellaneous	0.424	0.418	0.421
Missing/Undefined	0.293	0.180	0.223
Powerline	0.154	0.004	0.009
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.000	0.000	0.000
accuracy			0.427
macro avg	0.224	0.200	0.194
weighted avg	0.376	0.427	0.382

TABLA IX
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 2 UTILIZANDO DISCRIMINANTE LINEAL

	precision	recall	f1-score
Arson	0.000	0.000	0.000
Campfire	0.000	0.000	0.000
Children	0.000	0.000	0.000
Debris Burning	0.252	0.792	0.382
Equipment Use	0.000	0.000	0.000
Fireworks	0.000	0.000	0.000
Lightning	0.410	0.790	0.540
Miscellaneous	0.191	0.059	0.090
Missing/Undefined	0.000	0.000	0.000
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.000	0.000	0.000
accuracy			0.323
macro avg	0.066	0.126	0.078
weighted avg	0.180	0.323	0.209

TABLA X
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 2 UTILIZANDO DISCRIMINANTE CUADRÁTICO

	precision	recall	f1-score
Arson	0.173	0.036	0.059
Campfire	0.333	0.003	0.006
Children	0.018	0.573	0.034
Debris Burning	0.271	0.729	0.395
Equipment Use	0.167	0.002	0.003
Fireworks	0.094	0.143	0.113
Lightning	0.595	0.102	0.174
Miscellaneous	0.104	0.005	0.009
Missing/Undefined	0.221	0.056	0.089
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.050	0.008	0.014
accuracy			0.151
macro avg	0.156	0.127	0.069
weighted avg	0.277	0.151	0.119

TABLA XI
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 2 UTILIZANDO NAÏVE BAYES

	precision	recall	f1-score
Arson	0.210	0.029	0.051
Campfire	0.000	0.000	0.000
Children	0.018	0.579	0.035
Debris Burning	0.246	0.702	0.365
Equipment Use	0.154	0.003	0.006
Fireworks	0.092	0.168	0.119
Lightning	0.622	0.068	0.123
Miscellaneous	0.110	0.007	0.013
Missing/Undefined	0.219	0.063	0.098
Powerline	0.000	0.000	0.000
Railroad	0.000	0.000	0.000
Smoking	0.000	0.000	0.000
Structure	0.000	0.000	0.000
accuracy			0.140
macro avg	0.129	0.125	0.062
weighted avg	0.266	0.140	0.103

TABLA XII
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 3 UTILIZANDO RANDOM FOREST

	precision	recall	f1-score
Human	0.605	0.684	0.642
Malicious	0.554	0.280	0.372
Natural	0.753	0.781	0.767
Other	0.625	0.598	0.611
accuracy			0.647
macro avg	0.634	0.586	0.598
weighted avg	0.644	0.647	0.640

TABLA XIII
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 3 UTILIZANDO KNN

	precision	recall	f1-score
Human	0.515	0.663	0.580
Malicious	0.434	0.116	0.183
Natural	0.621	0.684	0.651
Other	0.530	0.422	0.470
accuracy			0.546
macro avg	0.525	0.471	0.471
weighted avg	0.539	0.546	0.529

TABLA XIV
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 3 UTILIZANDO DISCRIMINANTE LINEAL

	precision	recall	f1-score
Human	0.414	0.772	0.539
Malicious	0.000	0.000	0.000
Natural	0.437	0.569	0.494
Other	0.389	0.026	0.050
accuracy			0.421
macro avg	0.310	0.342	0.271
weighted avg	0.377	0.421	0.328

TABLA XV
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 3 UTILIZANDO DISCRIMINANTE CUADRÁTICO

	precision	recall	f1-score
Human	0.387	0.905	0.542
Malicious	0.160	0.123	0.139
Natural	0.613	0.148	0.239
Other	0.409	0.068	0.116
accuracy			0.387
macro avg	0.392	0.311	0.259
weighted avg	0.431	0.387	0.298

TABLA XVI
REPORTE DE CLASIFICACIÓN PARA EL ESCENARIO 3 UTILIZANDO NAÏVE BAYES

	precision	recall	f1-score
Human	0.372	0.892	0.526
Malicious	0.172	0.175	0.173
Natural	0.573	0.069	0.123
Other	0.280	0.036	0.064
accuracy			0.357
macro avg	0.349	0.293	0.221
weighted avg	0.376	0.357	0.250