Causalidad y predicción de incendios forestales en Argentina

Barreto, Romina
Ing.Industrial UTN FRBA

Pulella, M. Belén
Ing.Industrial UTN FRBA

Vigón, Lucila
Ing.Industrial UTN FRBA

***Abstract* — El objetivo de este proyecto es identificar las causas de los incendios producidos en Argentina y evaluar la factibilidad de un modelo de machine learning que permita la predicción de los mismos.**

Key Words: Incendios forestales, Machine Learning, Modelos de Regresión, auto-scaling.

I. Introduccion

Recientemente, han estado ocurriendo en muchos países del mundo, especialmente en Argentina, incendios de grandes magnitudes. Los mismos, generan consecuencias tanto en las personas, como en la flora y fauna del lugar afectado. El objetivo de este estudio es analizar las causas de ocurrencia de estos incendios, ya que no siempre son causados "naturalmente", sino de manera intencional, por el accionar del hombre. De poder lograr este objetivo, se buscarán realizar planes de acción para prevenirlos.Se busca, a través de un modelo de regresión, predecir las cantidades de incendios que ocurrirán a futuro habiendo identificado las zonas con mayor riesgo de ocurrencia.

Esto podría aportar a instituciones gubernamentales una idea general de dónde se estarían generando los focos de incendios dentro del territorio nacional. A partir de esta información, podrían optimizar la destinación de recursos en pos de evitar o combatir los mismos.

II. Data Set

En búsqueda del objetivo, se exploró la base de datos de la Dirección Nacional de Bosques del Ministerio de Ambiente y Desarrollo Sostenible. Así, se han encontrado dos muestras de datos con información muy enriquecedora para el análisis exploratorio. El primero de ellos cuenta con 999 instancias y 24 features, brinda información acerca de la causalidad del incendio por provincia y por año, mientras que el segundo (585 x 8), la superficie afectada por tipo de vegetación, por provincia y por año. Cabe aclarar que ambos datasets son entre los años 1993 y 2017.

III. análisis exploratorio de datos

A. Procesamiento de la base de datos

Comenzando con el análisis, el primer paso fue conocer cómo estaba formada nuestra matriz de datos para luego continuar con la estructuración y limpieza de la misma. Finalmente visualizar los datos.

Al buscar NaNs (datos nulos), nos encontramos con que había features completamente nulas y que no brindaban ningún tipo de información, las mismas fueron eliminadas. A su vez, observamos valores nulos que no debían ser eliminados a los que se les asignó el valor 0.

Una vez finalizado, se cambió el nombre a las features, para que coincidan en ambas matrices de datos, como es el caso de la feature "Año". Lo mencionado, nos permitió aplicar "Concat" por columnas y obtener una única matriz de datos compuesta por las dos matrices iniciales. Dicha nueva matriz resultó tener: 584 samples y 13 features.

B. *Estadísticas descriptivas*

Una vez depurada nuestra base, pudimos emplear diferentes herramientas de visualización y arribar a conclusiones.

Teniendo en cuenta la cantidad de incendios por provincia a lo largo de los años, pudimos concluir que Buenos aires y Rio Negro son las provincias con mayor cantidad de incendios, tal como se puede ver en el siguiente mapa de calor:

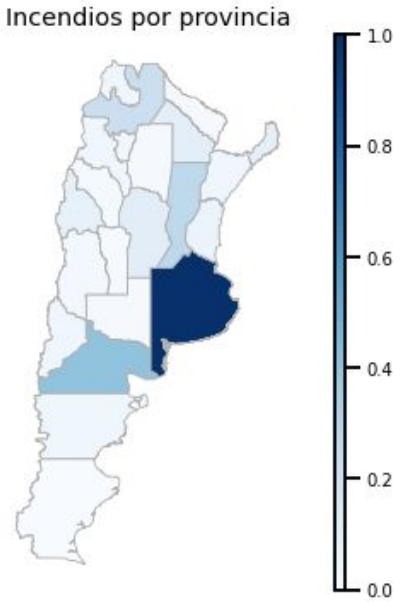


Figura 1. Cantidad de incendios por provincia en todo el período de análisis.

A su vez, pudimos observar que la mayoría de estos incendios se deben a causas desconocidas y en una segunda instancia, han ocurrido como por negligencia.

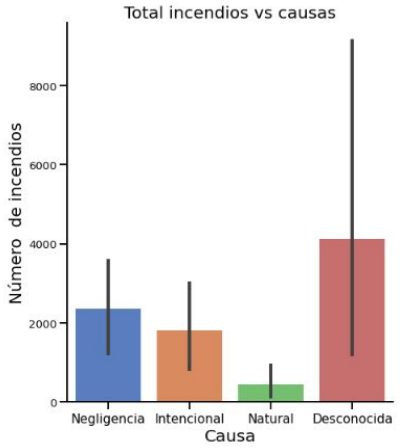


Figura 2. Causalidad de los Incendios en el período bajo análisis

Realizando el mismo análisis, pero ahora teniendo en cuenta la superficie afectada, llegamos a las mismas dos provincias. Rio Negro en primer lugar con la mayor cantidad de superficies afectadas y en segundo lugar Buenos Aires.

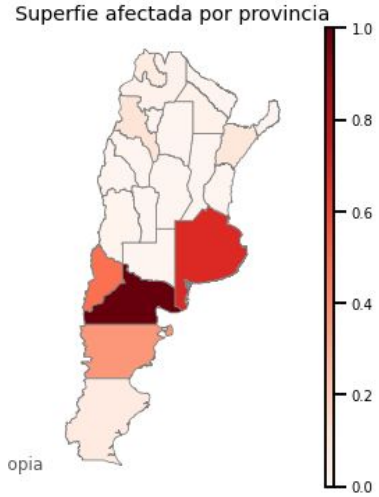


Figura 3. Hectáreas afectadas por provincia en todo el período de análisis.

En este caso, notamos que el tipo de superficie más afectada por incendios en argentina en el período bajo análisis corresponde a pastizales.

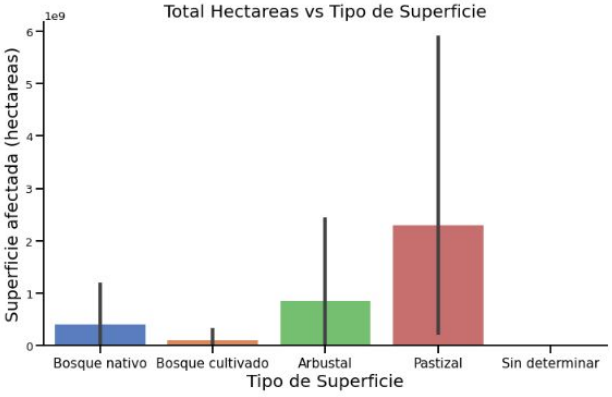


Figura 4. Clase de superficie afectada en todo el período de análisis.

Entrando a un detalle más profundo para ambas provincias, visualizamos la cantidad de incendios en función de su causalidad. Así pudimos concluir que la mayoría de esos incendios se deben a causas desconocidas. En el caso de Buenos Aires, también, una gran mayoría de los incendios son ocasionados por la negligencia del ser humano en su accionar.. Tal como se puede ver en el siguiente gráfico, lo mencionado anteriormente, es una tendencia que se mantiene a lo largo de los años de estudio (1993-2017)

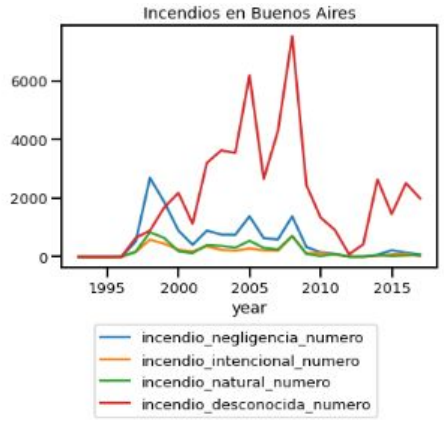


Figura 5. Causalidad de los incendios en Buenos Aires

Por su parte, la provincia Rio Negro, la gran mayoría de los incendios que tuvieron lugar se debieron a la intencion y negligencia del hombre.

Figura 6. Causalidad de los incendios en Río Negro

Siguiendo con la investicación de los incendios ocurridos en Rio Negro, pudimos concluir que la mayor cantidad de superficie afectada por los incendios fueron los pastizales en primer lugar y en segundo lugar arbustral.

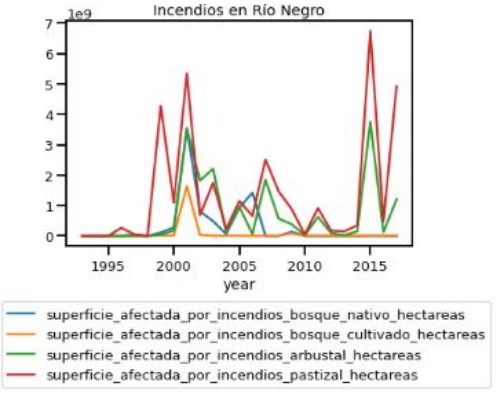


Figura 7. Hectáreas afectadas por provincia

IV. Materiales y métodos

Con la finalidad de cumplir con el objetivo planteado, se han empleado diferentes herramientas

del entorno de la programación. La principal fue Google Collaboratory, donde se ha desarrollado todo el código y, también, se han importado diferentes librerías. Para poder manipular los datos se han importado numpy y pandas, para el ploteo y visualización de los datos se ha importado: matplotlib , seaborn, plotly y geopandas (para georeferenciar los datos), y por último, para ejecutar los diferentes modelos regresión se ha importado la librería de Scikit learn.

A. Auto-scaling

Decidimos aplicar auto-scaling. El mismo, asume que cada feature responde a una distribución de probabilidad normal y busca estandarizar afectando los valores por la media y el desvío estándar.

$$x_i' = \frac{(x_i - \mu)}{\sigma}$$

Cada feature queda con una media=0 y un desvío=1 a lo largo de todas las muestras.

B. Polynomial Features

Durante el desarrollo de cada uno de los modelos, hemos decidido emplear lo conocido como polynomial features. Lo mencionado, consiste en elevar las características (features) existentes a un exponente, lo cual ayuda a realizar mejores predicciones.

C. Regresión

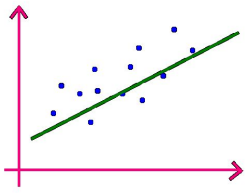
La regresión es un tipo de aprendizaje supervisado, el cual es empleado, cuando los datos tienen etiquetas que pertenecen al campo de los reales. Como nuestros datos cumplian con lo mencionado se ha tomado la decisión de aplicar diferentes modelos de regresión a fin de encontrar aquella función regresora F(x) que mejor se ajuste a nuestros datos minimizando el error.

D. Modelos de Regresión

Existen diferentes modelos de regresión dentro del machine learning:

a. Regresión Lineal

El modelo de regresión lineal, consiste en encontrar una F(X) lineal, la cual se construye calculando parámetros “w” asociados a cada dimensión/feature. Así dichos parámetros “w” determinarán el valor que tomará la variable dependiente “y”



$$F(x, w) = y$$

Figura 8. Regresión Lineal. Imagen tomada de los apuntes de Cluster AI.

b. Support Vector Regression (SVR)

El modelo Support Vector Regression construye hiperplano lineal y determina un margen (maximizado) como función de costo, tratando de que todas las muestras estén dentro de dicho margen.

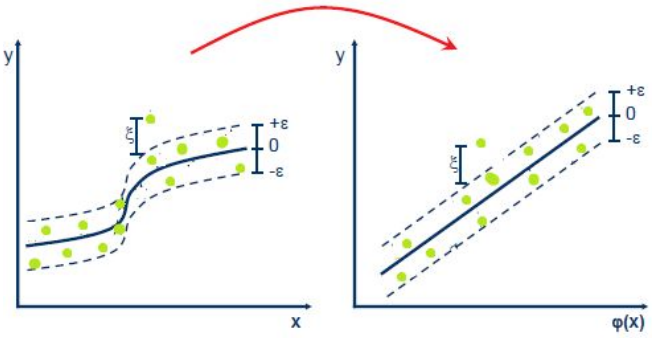


Figura 9. Modelo SVR. Imagen tomada de los apuntes de Cluster AI.

c. KNN Regression

Este tipo de modelo de regresión, a diferencia de los dos anteriores, no tiene que aprender parámetros w para definir f(x). En este caso, para determinar la y de una muestra, se tiene en cuenta la distancia euclídea de las k muestras vecinas más cercanas. En otras palabras, el y a predecir se determina por la interpolación de los y en los K vecinos.

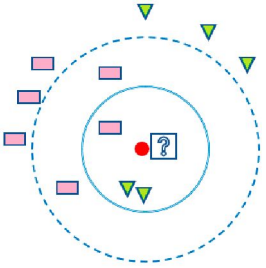


Figura 10. Modelo KNN. Imagen tomada de los apuntes de Cluster AI.

d. Random Forest Regression

Este modelo es un algoritmo que utiliza las características de múltiples árboles de decisión para formar las predicciones.

El algoritmo tiene una gran desventaja porque provoca un ajuste excesivo. Como ventaja podemos mencionar que, el Random Forest es más rápido y robusto que otros modelos de regresión.

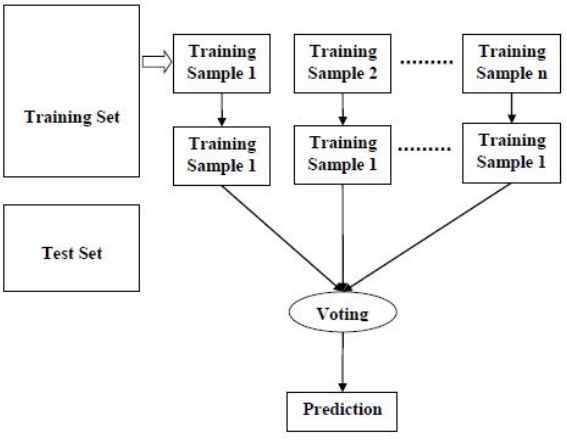


Figura 11. Modelo Random Forest.

E. Regularización

Dentro del modelo de regresión lineal, existe “Ridge Regression” el cual regulariza la regresión lineal, imponiendouna penalización a los parámetros w. De esta manera, hace que estos tiendan a cero en caso de que no sean tan importantes.

Así entonces, cuanto mayor sea la penalización más parámetros del vector w se aproximan a cero.

$$\hat{w}^{\text{ridge}} = \underset{w}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{d} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right\}$$

F. Medición de Resultados

Existen diferentes métricas para medir los resultados obtenidos en cada uno de los modelos. Ellas son R2. MSE y MAE:

a. R2 (Error medio al cuadrado)

El R2 explica la proporción de la varianza de “y” que explica el modelo de regresión. El R2 toma valores entre 0 y 1 y es independiente de la escala de “y”.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

siendo

$$\text{TSS} = \sum_{i=0}^{i=n} (y_i - \bar{y})^2$$

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

b. MSE (Mean Square Error):

El error cuadrático medio es sensible a predicciones muy malas por lo cual puede ser problemático en datos ruidosos. Se obtiene de la siguiente manera:

$$\boldsymbol{MSE} = \frac{\sum (\hat{y}_t - y_t)^2}{n}$$

c. MAE (Mean Average Error):

La media del error es mejor cuando no se quiere penalizar fuerte errores grandes. Se calcula de la siguiente manera:

$$\boldsymbol{MAE} = \frac{|\sum (\hat{y}_t - y_t)|}{n}$$

V. Resultados

Llevando a cabo los diferentes modelos, observaremos las mediciones del error para cada uno, algunas de ellas fueron:

Modelo	Features	R2	MSE	MAE
Lineal	Poly	0.55	590891.82	427.03
Ridge	Poly	0.44	747214.52	439.81
SVR	Poly	0.53	621794.87	382.873
KNN	Poly	0.18	1092518.36	455.72
Random	Liner	0.757	324041.37	455.72

En base a esto, hemos concluido que la mejor estimación fue obtenida al emplear el Modelo de **Random Forest Regression**

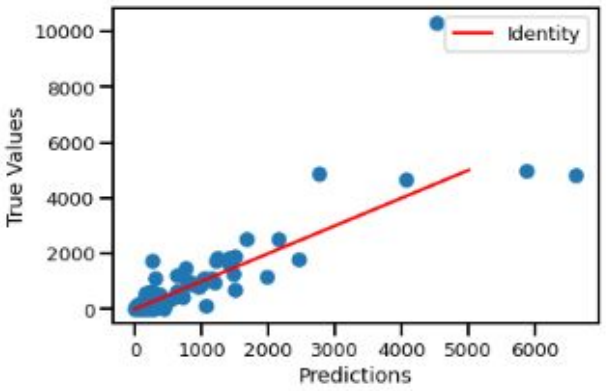


Figura 12. Regresión obtenida del Random Forest

VI. Conclusión

A partir de los datos abiertos de Ambiente y Desarrollo Sostenible hemos logrado construir un modelo de regresión robusto, que permite predecir, con un cierto margen de error los incendios, que tendrán lugar de aquí en adelante a ocurrir en el futuro.

También, pudimos analizar lo ocurrido en cada una de las provincias en cuanto a cantidad de incendios y superficie afectada, donde Buenos Aires y Río Negro eran las más perjudicadas. Como conclusión con respecto a ello, puede que esto se deba a que Buenos Aires, sea una de las provincias que más mide este tipo de situación y como consecuencia se ve reflejado en los registros con una gran diferencia con respecto al resto de las provincias de Argentina.

VII. Referencias

[1] Dirección Nacional de Bosques del Ministerio de Ambiente y Desarrollo Sostenible http://datos.ambiente.gob.ar/lt/dataset/incendios-forestales

[2] Apuntes de Cluster AI

[3] Numpy: https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Numpy_Python_Cheat_Sheet.pdf

[4] Scikit learn: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[5] Papers: https://ieeexplore.ieee.org/document/5158210 https://ieeexplore.ieee.org/document/8949588 https://ieeexplore.ieee.org/document/9092926/