MAGISTER EN INGENIERÍA INFORMÁTICA

Modelo Predictivo para la detección temprana de incendios forestales en Chile: Mitigando riesgos y pérdidas

AUTORES

Cristian Cuevas Tapia Diego Aristizabal Bejarano
Aris Miranda Garrido José Cariqueo Pilquianti
Matías Gómez Cartes

DOCENTES

Jorge Sabattin Ortega Mailiu Díaz Peña

CURSO

202400.9019 | PROY INTEGRADOR: CIENCIA DE DATOS

Santiago, mayo 2024

Tabla de Contenidos

Santiago, mayo 2024

Resumen

El aumento de los incendios forestales en Chile representa una amenaza creciente para la biodiversidad, la seguridad de las comunidades y la economía nacional. Este proyecto de ciencia de datos se enfoca en desarrollar un modelo predictivo para la detección temprana de incendios forestales en Chile, con el objetivo de mitigar riesgos y pérdidas asociadas. Utilizando datos proporcionados por CONAF, se lleva a cabo un análisis exhaustivo que incluye la exploración de datos, preprocesamiento, análisis predictivo e interpretación de resultados.

El análisis exploratorio de datos revela patrones importantes en la distribución y correlación de variables relacionadas con la ocurrencia de incendios. De dicho análisis se observa una mayor frecuencia de incendios forestales en la zona central de Chile (regiones de Valparaíso al Biobío) y la región de la Araucanía. El foco del análisis y aplicación de modelo se realizará sobre dichas regiones y que denominaremos “Zona Central ampliada”.

Se implementa una variedad de técnicas de preprocesamiento para limpiar y normalizar los datos, seguido por la comparación de varios modelos predictivos, incluyendo regresión lineal y árboles de decisión, entre otros. La evaluación del modelo se realiza utilizando métricas estándar como el error cuadrático medio, validación cruzada y otras.

Los resultados obtenidos proporcionan información valiosa para la toma de decisiones y la planificación de medidas preventivas. Se identifican factores clave relacionados con la ocurrencia de incendios forestales y se generan escenarios futuros bajo diferentes condiciones. Además, se discute la sensibilidad del modelo a cambios en los datos de entrada y se presentan recomendaciones para futuras investigaciones.

Este proyecto de ciencia de datos ofrece una contribución significativa a la predicción de incendios forestales en la zona central de Chile, destacando la importancia del análisis de datos en la gestión de desastres naturales.

Santiago, mayo 2024

Introducción

Descripción del problema

Los incendios forestales son una preocupación creciente en Chile, especialmente para la zona central y con mayor frecuencia en los meses de verano. Estos eventos no solo causan daños materiales significativos, sino que también ponen en peligro la vida humana, la flora y fauna de la región. La frecuencia y la intensidad de los incendios forestales han aumentado en los últimos años, exacerbados por factores como el cambio climático, la actividad humana y la falta de medidas preventivas efectivas. Ante esto, es necesario desarrollar herramientas y estrategias para detectar y prevenir incendios forestales de manera temprana, para reducir su impacto devastador en el medio ambiente y la sociedad.

Objetivos del proyecto

El objetivo principal de este proyecto es desarrollar un modelo de ciencia de datos que permita detectar patrones ambientales y predecir la ocurrencia de incendios forestales en Chile. Para lograr este objetivo, se plantean los siguientes objetivos específicos:

1. Recopilar y procesar datos relevantes sobre incendios forestales, condiciones climáticas y variables ambientales.
2. Realizar un análisis exploratorio de datos para comprender la distribución y correlación de las variables, identificar datos atípicos y faltantes, y generar visualizaciones descriptivas.
3. Preparar los datos para su análisis predictivo mediante técnicas de limpieza y normalización.
4. Comparar y evaluar diferentes modelos de aprendizaje automático para predecir la ocurrencia de incendios forestales.
5. Discutir los resultados obtenidos y proporcionar recomendaciones para la prevención y mitigación de incendios forestales en Chile.

Análisis Exploratorio de Datos (EDA)

Fuentes de datos

Para el desarrollo del proyecto, se utilizó la base de datos proporcionada por los docentes del curso, pero para poder complementar datos y abordar limitantes se utilizó una segunda base de datos, disponible en la página web de la Corporación Nacional Forestal (CONAF), llamada "Estadísticas - Causas según Daño de Incendios Forestales 1987 - 2023", la cual fue fundamental para comprender el significado de códigos en la base de datos inicial y garantizar una interpretación precisa de los datos.

La combinación de la base de datos inicial y la información adicional obtenida de CONAF permitió tener una fuente de datos completa y detallada que facilita el análisis y modelización de incendios forestales.

Exploración Inicial de datos

La fase de exploración inicial de datos fue crítica para comprender la estructura y la naturaleza del conjunto de datos con el que se trabajaría. Esta etapa incluyó la identificación de la dimensión del conjunto de datos y la clasificación de los tipos de variables presentes. A continuación, se detalla cada uno de estos aspectos para nuestro proyecto de los incendios forestales en Chile:

Dimensión del Conjunto de Datos

El conjunto de datos proporcionado contiene registros de incendios forestales del año 2017 y proveniente de la CONAF. Específicamente, este conjunto incluye:

- **Número de Registros (Entradas):** En total se trata de **5.234 registros**. Cada registro corresponde a un evento de incendio individual reportado y documentado.
- **Número de Variables (Características):** El conjunto de datos incluye múltiples características que describen cada evento de incendio, tales como fecha, localización, causa, tipo de vegetación afectada, y superficie quemada. En total existen **30 variables**.

A continuación, se comparte el detalle y significado de cada variable que forma de la base de datos de incendios forestales registrado por la CONAF durante el año 2017:

#	Variable	Detalle/significado
1	id	Identificador correlativo de cada registro
2	temporada	Período estadístico de los registros (1° de julio de un año a 30 de junio del año siguiente). En este caso el valor “2016-2017” para todos los registros.
3	codreg	Código numérico que identifica cada región de Chile
4	codprov	Código numérico que identifica cada provincia de Chile
5	codcom	Código numérico que identifica cada comuna de Chile
6	ambito	Alcance del territorio afectado por el incendio (CONAF o compañía forestal)
7	numero	Número de incendios en una región
8	nombre_inc	Nombre asignado al incendio
9	utm_este	Coordenada Este en proyección UTM

#	Variable	Detalle/significado
10	utm_norte	Coordenada Norte en proyección UTM
11	inicio_c	Caminos u otros lugares cercanos al inicio del incendio
12	combus_i	Tipo de material con el que habría iniciado el incendio
13	causa_gene	Causa general del incendio
14	causa_espe	Causa específica del incendio (detalle de la causa general)
15	pino_0010	Área quemada de pinos entre 0 a 10 años de edad
16	pino_11_17	Área quemada de pinos entre 11 a 17 años de edad
17	pino_18	Área quemada de pinos entre 18 más años de edad
18	Eucalipto	Área quemada de eucaliptos
19	otras_plan	Otras áreas de plantación diferentes a pino o eucalipto
20	total_plan	Suma de las áreas quemadas: pino_0010, pino_11_17, pino_18, eucalipto y otras_plan
21	Arbolado	Área quemada de vegetación nativa
22	Matorral	Área quemada de matorrales
23	Pastizal	Área quemada de pastizales
24	total_veg	Suma de las áreas quemadas: arbolado, matorral y pastizal
25	Agricola	Área agrícola quemada
26	Desechos	Área quemada considerada como desechos
27	total_otra	Suma de las áreas quemadas: agrícola y desechos
28	sup_t_a	Total áreas quemadas (suma de total_plan, total_veg y total_otra)
29	long	Longitud
30	lat	Latitud

Tipos de Variables

El análisis de los tipos de variables es fundamental para determinar las técnicas de procesamiento y análisis adecuadas. Las variables se pueden clasificar en varias categorías, dependiendo de su naturaleza y el rol que desempeñan en el análisis:

Variables Numéricas

- **Continuas:** Variables como sup_t_a (superficie total afectada) que pueden tomar cualquier valor dentro de un rango. Estas son clave para la modelación y requieren chequeo de normalidad y posible transformación.
- **Discretas:** Por ejemplo, el número de incendios reportados por zona, que cuentan eventos y son valores enteros.

Variables Categóricas

- **Nominales:** No tienen un orden inherente, como el tipo de vegetación (e.g., bosque nativo, plantación de pino).
- **Ordinales:** Categorías con un orden específico.
- **Variables Temporales:** Fecha del incendio, que puede requerir transformación para análisis como agrupación por mes o estación.

En base a los diferentes tipos de variables de la base de datos de incendios forestales en Chile, se identifican los siguientes tipos:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5234 entries, 0 to 5233
Data columns (total 30 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          5234 non-null   int64
 1   temporada   5234 non-null   object
 2   codreg      5234 non-null   int64
 3   codprov     5234 non-null   int64
 4   codcom      5234 non-null   int64
 5   ambito      5234 non-null   object
 6   numero      5234 non-null   float64
 7   nombre_inc  5234 non-null   object
 8   utm_este    5234 non-null   float64
 9   utm_norte   5234 non-null   float64
 10  inicio_c    5234 non-null   object
 11  combus_i    5234 non-null   object
 12  causa_gene  5234 non-null   float64
 13  causa_espe  5234 non-null   object
 14  pino_0010   5234 non-null   float64
 15  pino_11_17  5234 non-null   float64
 16  pino_18     5234 non-null   float64
 17  eucalipto   5234 non-null   float64
 18  otras_plan  5234 non-null   float64
 19  total_plan  5234 non-null   float64
 20  arbolado    5234 non-null   float64
 21  matorral    5234 non-null   float64
 22  pastizal    5234 non-null   float64
 23  total_veg   5234 non-null   float64
 24  agricola    5234 non-null   float64
 25  desechos    5234 non-null   float64
 26  total_otra  5234 non-null   float64
 27  sup_t_a     5234 non-null   float64
 28  long        5234 non-null   float64
 29  lat         5234 non-null   float64
dtypes: float64(20), int64(4), object(6)
memory usage: 1.2+ MB
```

Imagen N°1: Tipos de variables de la base de datos de incendios forestales en Chile para la temporada 2016-2017.

Las variables enteras (destacadas en color rojo en imagen anterior) son: *id*, *codreg*, *codprov* y *codcom*. La primera no será considerada dado que se trata del correlativo de cada registro. Las otras variables serán tratadas como categóricas. Por otro lado, se identifican variables categóricas (destacadas en color amarillo), de las que se eliminarán:

- *temporada*: Dado que sólo guarda el mismo valor para todos los registros "2016-2017")
- *nombre_inc*; Dado que es un valor nominativo que se le da a cada incendio (nombre)
- *inicio_c*: Dado que es un camino o localidad cercano al inicio del incendio.

Se mantendrán las variables: *ámbito*, *combus_i* y *causa_espe*. Se identificó un caso especial (causa_gene), que si bien se identificó como continua, fue una interpretación errónea del código por tener valores de identificación de la clasificación, por ejemplo: 1.01, 2.01, etc.; dicha variable será tratada como categórica (al igual que causa_espe). El resto de las variables son continuas y se mantendrán (tipo float64).

Resumen de Medidas Estadísticas

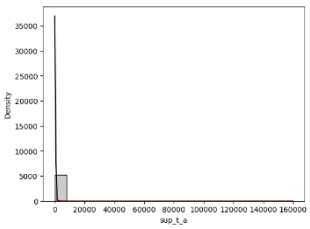
El análisis estadístico reveló que la mayoría de los incendios afectan a áreas pequeñas, aunque existe un número menor de incendios afectando a un número significativo de superficie (se identifica al revisar el último cuartil de la variable dependiente). Por otro lado, la media de la variable dependiente *sup_t_a* (1.088306e+02) es mucho mayor que su mediana (5.000000e-01), por lo que estaríamos frente a una distribución asimétricas positiva, lo que se ratifica con la prueba de normalidad de Shapiro-Wilk (*p-value* = *0.0000000000*).

	count	mean	std	min	25%	50%	75%	max
numero	5234.0	3.538999e+02	277.383422	1.000000e+00	1.260000e+02	2.890000e+02	5.280000e+02	1117.00
utm_este	5234.0	5.083689e+05	209721.431390	2.244440e+05	2.786560e+05	6.391985e+05	7.032372e+05	775622.00
utm_norte	5234.0	6.005836e+06	292761.663034	3.981678e+06	5.834825e+06	5.944400e+06	6.269925e+06	7952097.00
pino_0010	5234.0	1.359989e+01	557.103917	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	38354.42
pino_11_17	5234.0	7.302577e-01	16.932087	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	700.00
pino_18	5234.0	2.922131e+01	1064.558117	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	73619.12
eucalipto	5234.0	9.274645e+00	165.960143	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7075.10
otras_plan	5234.0	9.275621e-01	39.021636	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2732.14
total_plan	5234.0	5.375367e+01	1750.270765	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e-02	119920.49
arbolado	5234.0	1.724175e+01	354.555764	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	20891.60
matorral	5234.0	2.148128e+01	316.740351	0.000000e+00	0.000000e+00	7.000000e-02	5.000000e-01	12938.70
pastizal	5234.0	1.019263e+01	94.166918	0.000000e+00	0.000000e+00	1.000000e-02	5.475000e-01	2463.00
total_veg	5234.0	4.891566e+01	687.461608	0.000000e+00	2.000000e-02	2.650000e-01	1.500000e+00	35842.26
agricola	5234.0	5.488889e+00	143.726970	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7519.00
desechos	5234.0	6.723611e-01	14.836936	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	960.00
total_otra	5234.0	6.161250e+00	144.487575	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	7519.00
sup_t_a	5234.0	1.088306e+02	2460.163794	0.000000e+00	9.000000e-02	5.000000e-01	2.400000e+00	159812.58
long	5234.0	2.198360e+05	67002.974238	8.598404e+04	1.668695e+05	2.180593e+05	2.720998e+05	476719.00
lat	5234.0	6.003468e+06	294353.943325	3.981678e+06	5.826784e+06	5.940878e+06	6.269925e+06	7952097.00

Imagen N°2: Medidas estadísticas de variables.

Visualizaciones

Imagen N°3: Se generaron histogramas para visualizar la distribución de la superficie total afectada por incendios. Se identifica la concentración de incendios que afectan superficies pequeñas y la presencia de pocos eventos que afectan áreas significativamente mayores.



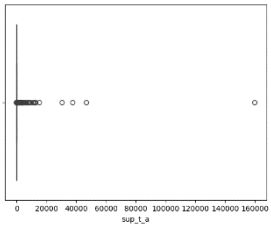
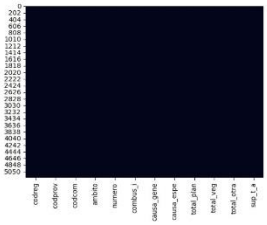


Imagen N°4: Los boxplots permiten visualizar rápidamente la mediana, los cuartiles y los valores extremos. Estos son útiles para decidir si se deben mantener estos **outliers** en el análisis, dado que pueden representar incendios de gran magnitud que son críticos para el estudio.

Imagen N°5: Se empleó un mapa de calor para visualizar la presencia de datos faltantes en el conjunto de datos. Afortunadamente, el análisis visual confirmó que no hay valores faltantes en las variables clave. Esto es importante para asegurar la calidad del modelo predictivo, ya que los datos faltantes pueden introducir sesgos si no se tratan adecuadamente.



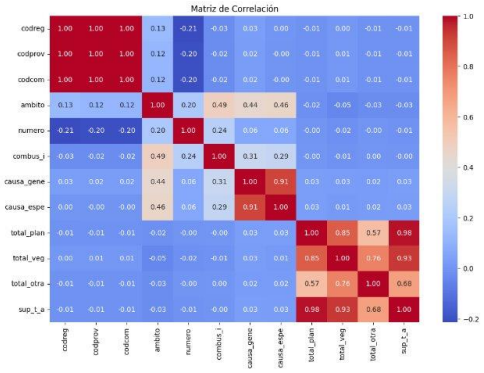


Imagen N°6: Se crearon gráficos de matrices de correlación para examinar las relaciones entre las diferentes variables como ***total_plan***, ***total_veg***, y ***total_otra***. Estos gráficos identifican variables con influencia significativa, lo cual es crucial para la selección de características en el modelo predictivo.

Análisis Predictivo

El análisis predictivo en este proyecto se enfoca en construir y evaluar modelos que puedan predecir la superficie total afectada por incendios forestales en Chile. Este proceso involucra varias etapas, comenzando con la preparación de datos, seguida de la selección y comparación de modelos, y finalizando con la evaluación del desempeño de los modelos.

Limpieza y Normalización de Datos

- **Limpieza de Datos:** El primer paso en la limpieza de datos fue eliminar variables que no aportaban al modelo (por ejemplo: *id*). Se verificaron todas las entradas para garantizar que los datos estén correctamente alineados con las definiciones de las variables, y se realizaron correcciones donde fue necesario.
- **Tratamiento de Outliers:** Dado que la variable *sup_t_a* presentaba outliers significativos que podían influir desproporcionadamente en el modelo predictivo, se optó por aplicar técnicas de truncamiento y transformaciones logarítmicas para reducir el impacto de estos valores extremos sin eliminarlos completamente del conjunto de datos, preservando así información valiosa sobre incendios severos.
- **Normalización de Datos:** Se aplicaron técnicas de normalización, especialmente la estandarización (restar la media y dividir por la desviación estándar), para que las variables numéricas tuvieran un impacto equitativo durante el entrenamiento del modelo y mejorar la convergencia de los algoritmos de aprendizaje automático.

Comparación de Modelos Vistos

- **Selección de Modelos:** Se seleccionaron varios modelos de aprendizaje automático que se habían estudiado durante el curso, incluyendo Regresión Lineal Múltiple (RLM), Perceptrón Multicapa (MLP), Árbol de Decisión (DT) y K-Nearest Neighbors (KNN).
- **Implementación de Modelos:** Para cada modelo se utilizó la base de datos “limpia” (quitando variables que se identificaron no aportaban al modelo, según la correlación de variables identificadas y el valor de p-value para cada una de ellas en cada modelo).
- **Comparación Basada en Métricas:** Los modelos se compararon basándose en métricas estándar de evaluación como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Además, se consideraron aspectos como la complejidad del modelo y el tiempo de entrenamiento para asegurar una implementación eficiente.

Evaluación del Modelo

- **Validación Cruzada:** Para asegurar que los modelos fueran robustos y generalizables, se utilizó la técnica de validación cruzada. Esto implica dividir el conjunto de datos en varios subconjuntos y entrenar el modelo en varios de estos, mientras se prueba en el restante.
- **Análisis de Resultados:** Los resultados de la validación cruzada fueron analizados para identificar el modelo con el mejor desempeño general. Se prestaron especial atención a las métricas de error para asegurar que el modelo seleccionado proporcionara predicciones precisas y consistentes.
- **Interpretación de Modelos:** Además de la precisión del modelo, también se consideró la interpretabilidad de los resultados. Por ejemplo, los modelos de árbol de decisión ofrecen una buena visualización de cómo se toman las decisiones, lo cual es valioso para explicar los resultados a las partes interesadas.

Discusión de Resultados

La discusión de los resultados obtenidos del análisis predictivo nos proporciona una oportunidad para profundizar en la interpretación de los modelos y evaluar su aplicabilidad práctica. Por consiguiente, se exploran las implicaciones de los resultados, identificando fortalezas y limitaciones de los modelos utilizados y sugiriendo áreas de mejora para investigaciones futuras.

Interpretación de Modelos y Resultados Claves

- **Importancia de las Variables:** Los modelos predictivos revelaron que la variable *total_veg*, que representa la superficie afectada sobre bosque nativo, es un predictor crucial para la superficie total afectada por incendios. Esto indica que los incendios en vegetación nativa son más significativos en términos de área afectada en comparación con otros tipos de vegetación. La relevancia de esta variable resalta la importancia de enfocar esfuerzos de prevención y respuesta en estas áreas.
- **Desempeño de los Modelos:** Entre los modelos evaluados, el Perceptrón Multicapa (MLP) y el Árbol de Decisión (DT) demostraron ser más efectivos en la predicción de la superficie afectada. Estos modelos lograron capturar mejor las complejidades y las relaciones no lineales entre las variables. Sin embargo, la Regresión Lineal Múltiple (RLM) mostró resultados que eran demasiado perfectos, lo cual es sospechoso y podría indicar un sobreajuste.

	MSE	RMSE	MAE	R^2
Linear Regression	0.0000	0.0000	0.0000	1.0000
Multi-layer Perceptron	20033.4431	141.5395	5.6518	0.9911
Decision Tree	154592.7210	393.1828	16.9169	0.9317
K-Nearest Neighbors	685421.2666	827.9017	45.5650	0.6971

Imagen N°7: Comparación de principales mediciones de errores y precisión de los modelos comparados.

- **Problemas de Sobreajuste y Generalización:** A pesar del buen rendimiento en el conjunto de datos de entrenamiento, hay preocupaciones de sobreajuste, especialmente con modelos complejos como el MLP. Esto puede limitar la capacidad del modelo para generalizar a nuevos datos. Se realizaron pruebas quitando variables lineales directamente relacionadas con la variable dependiente *sup_t_a* (subtotales), teniendo resultados con R2 y visualizaciones que nos indicaron mantener las variables originales. Es crucial implementar técnicas como la validación cruzada y ajustar la regularización en estos modelos para asegurar que sean robustos y aplicables a datos no vistos.

Limitaciones y Desafíos

- **Falta de Datos Externos:** Un desafío significativo en este estudio fue la ausencia de variables climáticas y geográficas en el conjunto de datos. Estas variables podrían influir considerablemente en la frecuencia y severidad de los incendios forestales. Su inclusión podría mejorar significativamente la precisión y la utilidad de los modelos predictivos.
- **Distribución de los Datos:** La distribución de la variable ***sup_t_a*** no es normal, lo que complica el uso de ciertos tipos de análisis estadísticos y modelos predictivos que asumen

la normalidad de los datos. Esto requirió la implementación de transformaciones de datos y la selección de modelos que pueden manejar tales distribuciones.

Conclusiones y Recomendaciones Futuras

Inclusión de más variables

Para futuras investigaciones, sería beneficioso expandir el conjunto de datos para incluir variables como condiciones climáticas, patrones de viento, y otras características geográficas. Estas variables podrían proporcionar insights más profundos sobre los patrones y causas de los incendios forestales (dado que los datos usados en este proyecto fueron limitados y con foco en subconjuntos de áreas afectadas por incendios)

Modelos híbridos y enfoques avanzados

Considerar la implementación de modelos híbridos que combinen elementos de varios modelos predictivos para mejorar la precisión y la robustez. Además, explorar técnicas avanzadas de aprendizaje automático, como el aprendizaje profundo y los modelos ensemble, podría ofrecer mejoras significativas en la predicción de incendios forestales.

Evaluación continua y ajuste de modelos

Es esencial establecer un protocolo para la evaluación continua de los modelos predictivos a medida que se disponga de nuevos datos. Esto ayudaría a mantener la relevancia y precisión del modelo a lo largo del tiempo.

Referencias Bibliográficas

- Corporación Nacional Forestal (CONAF). (1964 - 2023). Datos anuales de incendios forestales. CONAF. https://www.conaf.cl/incendios-forestales/incendios-forestales-en-chile/estadisticas-historicas/