CLASIFICACIÓN Y PREDICCIÓN DE TIPOS DE ARBOLES EN LA RESERVA ROOSEVELT USANDO DATOS CARTOGRÁFICOS

ALEJANDRO ANDRES MENDEZ MIRANDA

Proyecto de Grado entregado a la Facultad de Ingeniería para optar al grado académico de Magíster en Data Science

Profesor guía: DRA. DANIELA OPITZ

Facultad de Ingeniería
Universidad del Desarrollo
Chile

10 de enero de 2022

Índice

Resumen

Debido a las restricciones en reservas naturales es complejo realizar el estudio de estas. En este trabajo presentamos distintos modelos de machine learning para predecir el tipo de cobertura de árbol utilizando el dataset Forest Cover Type, de la Roosevelt National Forest en Colorado. Este set de datos contiene 581.012 observaciones, 54 atributos con información cartográfica y 7 categorías de árboles a predecir, cada instancia corresponde a un área de 30x30m donde la categoría tenga predominancia. Para realizar las predicciones se utilizaron dos modelos de machine learning: Random Forest y LightGBM, Se experimentó utilizando la función de pérdida Focal Loss y adicionando información sintética de las categorías minoritarias utilizando redes CTGAN. Con este último enfoque se alcanzó un valor para la métrica F1 de 0.943 y accuracy de 0.966. Un análisis de la interpretabilidad del modelo reveló uno de los atributos más importantes para predecir la cobertura de arboles es la *Elevación, Distancia horizontal a carreteras y Distancia horizontal a puntos de incendios.*

1. Introducción

La conservación y estudio de la naturaleza es uno de los desafíos más importantes que enfrenta la humanidad en siglo XXI. Actualmente, esta importancia se ha hecho patente debido al cambio climático, cuyo efecto más visible es el incremento de desastres naturales y el cambio que experimentan la flora y fauna [11]. Cada año, por efecto directo o indirecto del ser humano y del cambio climático, desaparecen miles de hectáreas de árboles y dejan de existir especies de animales. La deforestación producida principalmente por la actividad humana, produce la reducción de la población de árboles nativos, desplazamiento de animales, alteración de la biodiversidad y la generación de nuevos ecosistemas a una velocidad muy alta y peligrosa, y en muchos casos, conllevando a la muerte de la vida natural [28].

Una de las medidas tomadas por los países para conservar parte de la naturaleza es la selección de áreas protegidas. En estas áreas la presencia humana se limita parcial o totalmente, lo que permite generar ecosistemas donde el control de la flora y fauna es un efecto natural del ecosistema y no producto de la intervención humana [15]. A pesar de lo efectivas que son las restricciones de acceso en la conservación de áreas protegidas, dichas restricciones generan barreras para el estudio de la diversidad, el estudio del desarrollo de las distintas especies y la estimación de poblaciones. En vista de estas limitantes, la ciencia está constantemente buscando y desarrollando nuevas metodologías que permitan estudiar áreas protegidas [27].

Uno de los procedimientos más utilizados para estudiar ecosistemas naturales es el censo forestal, que permite estimar la población de árboles, predecir la cantidad de biomasa vegetal viva y muerta y estimar los posibles biomas que se generan para los animales, entre otros [9]. Para esta tarea se han diseñado muchas metodologías, donde una de las más importantes se basa en el método de conteo presencial en áreas designada con formas circulares, cuadradas o rectangulares para luego extrapolar los resultados a las demás zonas de estudio [12]. Sin embargo, estas metodologías se han vuelto un tanto obsoletas, ya que actualmente es sabido que en los ecosistemas raramente las distribuciones siguen asunciones de independencia, homocedasticidad y normalidad [3].

En vista de las limitantes de los métodos tradicionales para estudiar ecosistemas ha sido necesario incorporar métodos más complejos, tales como algoritmos de machine learning, que manejan de mejor forma las distintas relaciones entre los datos[7]. La actual dificultad para adquirir información de áreas protegidas puede ser mitigada con el uso de algoritmos de aprendizaje automático tales como árboles de decisiones junto con imágenes satelitales en sus distintos espectros de onda. Estos métodos

por ejemplo, pueden realizar predicciones de la especie de árbol predominante en un área específica (o tipo de cobertura) de la imagen con muy buenos resultados llegando a métricas de accuracy del 92 % [25, 26, 21, 8]. Otra metodología alternativa que no ha sido muy explorada, es la utilización de datos cartográficos, tales como la elevación, cercanía de ríos, iluminación, etc, que está recopilada por distintas instituciones dedicadas al estudio de las distintas zonas geográficas [5].

Considerando la importancia de la conservación de la naturaleza, los límites de estudio de las áreas naturales protegidas y el avance en los modelos matemáticos de predicción de categorías, en este trabajo se analizarán distintos algoritmos de aprendizaje automático para predecir tipos de coberturas de árboles utilizando datos cartográficos provenientes del dataset Forest Cover Type, de la Reserva Forestal Nacional Roosevelt en Colorado y estimar si el modelo es suficiente para realizar las predicciones comparado con utilizar análisis de imágenes y analizar el uso de esta información en otras metodologías.

2. Hipótesis y objetivos

2.1. Hipótesis

La hipótesis principal de este trabajo es que usando datos cartográficos de coberturas de árboles tales como elevación, inclinación y sombreado de relieve a lo largo del día, entre otros, de la reserva nacional Roosevelt National Forest, y algoritmos de machine learning, es posible desarrollar un modelo de predicción de la especie predominante de arboles en una área específica de la misma reserva forestal y analizar la importancia de las variables en las predicciones finales y estimar el uso de estas variables en otros modelos.

2.2. Objetivos

2.2.1. Objetivo general

El objetivo general es generar un modelo de predicción de tipos de cobertura de arboles, robusto y con un alto nivel de accuracy (exactitud), F1 y AUC (Área bajo la curva ROC). Además de analizar la interpretabilidad del modelo e importancias de las variables.

2.2.2. Objetivos específicos

Para lograr el objetivo general se plantean los siguientes objetivos específicos:

- Desarrollar un análisis exploratorio de los datos que servirán de input para el modelo.
- Seleccionar algoritmos apropiados para implementarel modelo predictivo.
- Obtener métricas de validación.
- Probar distintos parámetros y mejoras del modelo en un proceso iterativo de búsqueda.
- Analizar la influencia de las distintos parámetros del modelo e influencia de las variables.
- Analizar la interpretabilidad del modelo generado.

3. Descripción de datos

Se utilizó como fuente de datos el set de datos Forest Cover Type de la Reserva Forestal Nacional Roosevelt en Colorado, creado por Blackard [2] y disponibilizado por UCI Machine Learning [13]. Este dataset fue obtenido por medio de la combinación de observaciones aéreas a larga escala por el US Forest Service (USFS) y datos de cartográficos otorgados por el US Geological Survey (USGS), para las áreas silvestres Rawah (296.28 km^2), Comanche Peak (273.89 km^2), Neota (39.04 km^2) y Cache la Poudre (38.17 km^2) sumando un total 647.38 km^2 . Las zonas geográficas descritas pertenecen a la reserva y pueden ser visualizadas en la Figura 1.

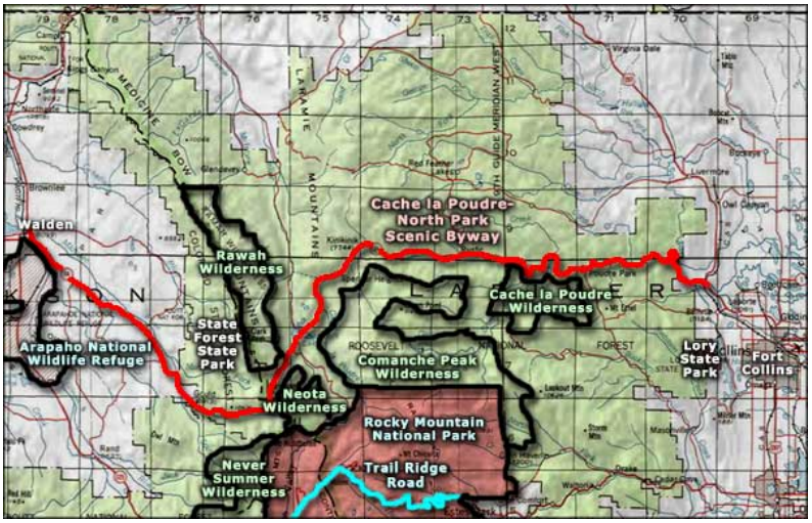

Figura 1: Mapa de áreas silvestres ubicadas en la Reserva Nacional Roosevelt.

Las áreas de estudio fueron seleccionadas debido a que estas han experimentado prácticamente nula intervención humana, por lo cual el crecimiento de árboles en esta zona es principalmente resultado de un proceso ecológico natural.

El set de datos consta de 54 atributos (variables o columnas) y 581.012 observaciones o instancias. Cada observación corresponde a información de un cuadrante $30x30\,m$ para la especie de árbol predominante en dicha área, ignorando la presencia árboles minoritarios en estos cuadrantes, si se realiza la estimación del área cubierta serían $522.91\,km^2$ o un 80 % del área total. La variable objetivo o cobertura de árbol hace referencia a siete especies distintas predominante en la región: Lodgepole pine (*Pinus contorta*), Spruce/fir (*Picea engelmannii* y *Abies lasiocarpa*), Ponderosa pine (*Pinus ponderosa*), Douglas-fir (*Pseudotsuga menziesii*), Aspen (*Populus tremuloides*), Cottonwood/Willow (*Populus angustifolia, Populus deltoides, Salix bebbiana* y *Salix amygdaloides*) y Krummholz (*Picea engelmannii , Abies lasiocarpa* y *Pinus aristata*). Estas especies tienen las siguientes características:

- Lodgepole Pine: Vive en suelos bien drenados en altas elevaciones en bosques de su misma especie. El fuego mata la mayor parte de la especie, pero se vuelve a establecer rápido ya que los conos se abren con el calor y las semillas son liberadas.

- Spruce/Fir: Vive en suelos arenosos bien drenados, en zonas húmedas y en los arrollos de las montañas. Muy suceptible al fuego debido a su delgada corteza.

- Ponderosa Pine: Vive en hábitats secos, con suelos de poco nutrientes. Resistente al fuego debido a la dureza de la corteza y autopodado de sus ramas. Además, sus hojas contienen un alto contenido de humedad.
- Douglas-fir: Vive en suelos rocosos de laderas húmedas en bosques de su misma especie o con otras coníferas. Dada su corteza con alto contenido en resinas lo hace particularmente suceptible al fuego. Algunos se han adaptado a los incendios al aumentar el grosor de su corteza.
- Aspen: Ubicado en muchos tipos de suelo, en especial en los bien drenados, arenosos y en pendientes de grava. Muere fácilmente con el fuego, pero recoloniza los suelos rápidamente.
- Cottonwood/Willow: Vive en llanuras aluviales, cercanos a riachuelos; en bosques húmedos y en bajas alturas, se encuentran solos o con sauces. Muy suceptible al fuego.
- Krummholz: Vive en alturas extremas, en bosques distorcionados por el viento. Suceptibles a los incendios al debido a que sus ramas se encuentran muy cercanas al suelo.

A continuación se describe cada una de la variables del set de datos y se presenta un resumen en la Tabla 1, Tabla 2 y Tabla 3. Además, una visualización de la heterogeneidad entre los distintas Clases de suelos junto a las categorías de Tipo de suelo se puede observar en la Figura 2. Las variables del set de datos son:

- Elevación: Altura en la que se encuentra la cobertura de árbol.
- Exposición: Dirección del compás por el cual el sol enfrenta una pendiente o cerro.
- Distancias vertical/horizontal: Distancia a cuerpos de agua, carreteras o zonas de incendios.
- Sombreado de relieve: Cantidad de iluminación a distintas horas del día.
- Área silvestre: Zonas designadas en la reserva Roosevelt, mostradas en la Figura 1.
- Clase de suelo: Designación dada según el tipo de roca y clima en distintos terrenos.
- Tipo de cobertura: Designada según el árbol predominante en un cuadrante definido.

Se debe destacar que la Elevación se obtuvo directamente del *USGS digital elevation model (DEM)*. Utilizando esta representación gráfica de como es la elevación del terreno se obtuvieron los

demás datos. Exposición, Inclinación y Sombreado fueron obtenidos utilizando estandares GIS para el análisis de superficie y sombreado, por el Environmental System Research Institute. Las distancias horizontales fueron calculadas utilizando distancia Euclidiana con los puntos más cercanos a cuerpos de aguas y carreteras del USGS, mientras que para los puntos de incendio se utilizó los puntos iniciales de ignición.

Tabla 1: Atributos del set de datos, x en la sigla hace referencia a la categoría del atributo

Atributo	**Sigla**	**Tipo**	**Medida**
Exposición	Asp	Cuantitativa	Azimut
Elevación	Elv	Cuantitativa	Metros
Inclinación	Slope	Cuantitativa	Grados
Distancia horizontal a cuerpos de agua	HdH	Cuantitativa	Metros
Distancia vertical a cuerpos de agua	VdH	Cuantitativa	Metros
Distancia horizontal a carreteras	HdR	Cuantitativa	Metros
Sombreado de relieve a las 9am	Hs9am	Cuantitativa	Índice 0 a 255
Sombreado de relieve medio día	HsNoon	Cuantitativa	Índice 0 a 255
Sombreado de relieve a las 3pm	Hs3pm	Cuantitativa	Índice 0 a 255
Distancia horizontal a puntos de incendio	HdFP	Cuantitativa	Metros
Área silvestre (4 columnas binarias)	Wax	Cualitativa	0 (Ausencia) o 1 (Presencia)
Clase de suelo (40 columnas binarias)	Stx	Cualitativa	0 (Ausencia) o 1 (Presencia)
Tipo de cobertura (7 tipos)	Cty	Valor objetivo	1 a 7

Tabla 2: Áreas silvestres de la reserva Roosevelt

Área silvestre	**Sigla**	**Cantidad**
Rawah	Wa1	260796
Neota	Wa2	29884
Comanche Peak	Wa3	253364
Cache la Poudre	Wa4	36968

Como se observa en la Figura 2, diagrama que nos muestra la proporción de las variables categóricas en el área silvestre Cache la Poudre, la cantidad de datos disponible por tipo de cobertura está desbalanceada, tal que la especies Spruce/Fir y Lodgepole Pine concentran el mayor área. Además se observa que los distintos tipos de suelos concentran cantidades distinta de datos, siendo para este

Tabla 3: Tipo de cobertura de árbol.

Cobertura	Cantidad
Spruce/Fir	211840
Lodgepole Pine	283301
Ponderosa Pine	35754
Cottonwood/Willow	2747
Aspen	9493
Douglas-fir	17367
Krummholz	20510

caso la clase de suelo 10 la predominante.

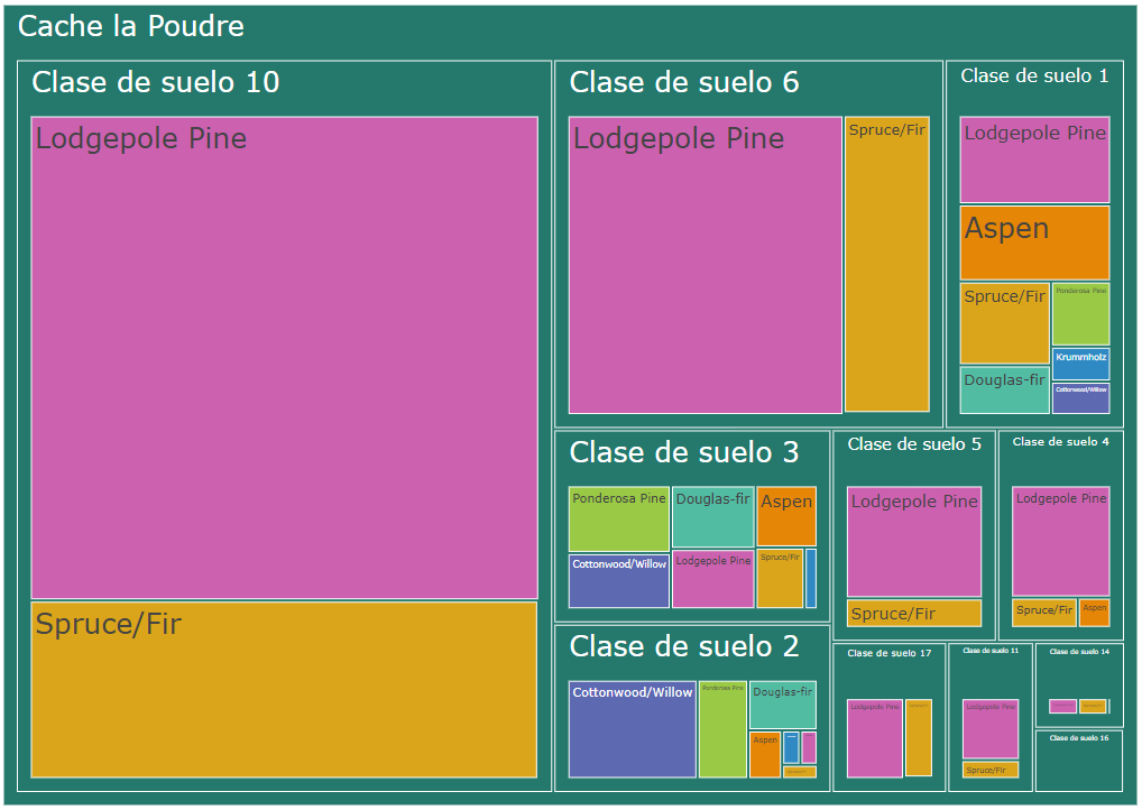

Figura 2: Treemap de las variables categóricas dentro del dataset. Se realizan tres separaciones, de exterior a interior: Área silvestre Cache la Poudre, Clase de suelo y Tipo de cobertura. El tamaño de los contenedores varía según su frecuencia absoluta dentro del contenedor mayor.

Por otro lado, la clase de suelo hace referencia a una codificación realizada por el Unite States Forest Service (USFS) mediante los valores del Ecological Landtype Units (ELUs). Los valores

para esta reserva forestal se presentan en la Tabla 4; una mejor especificación de que representan los códigos ELUs se presentan en la Tabla 5. El primer dígito del código representa una zona climática, mientras que el segundo dígito una zona geológica, el tercer y cuarto dígito representan una codificación especial para la reserva. Por ejemplo, la Clase de suelo 2 le corresponde el código 2703, el primer dígito 2 nos indica que la zona climática es *Montaña Baja*, mientras que el dígito 7 nos indica una zona geológica *Ígnea y metamórfica*.

Clase de suelo	**Código ELUs**	**Clase de suelo**	**Código ELUs**
1	2702	21	7103
2	2703	22	7201
3	2704	23	7202
4	2705	24	7700
5	2706	25	7701
6	2717	26	7702
7	3501	27	7709
8	3502	28	7710
9	4201	29	7745
10	4703	30	7746
11	4704	31	7755
12	4744	32	7756
13	4758	33	7757
14	5101	34	7790
15	5151	35	8703
16	6101	36	8707
17	6102	37	8708
18	6731	38	8771
19	7101	39	8772
20	7102	40	8776

Tabla 4: Codigos Elus para las distintas Clases de suelo.

1er Dígito	**Zona Climática**	**2do Dígito**	**Zona geológica**
1	Montaña baja seca	1	Aluvión
2	Montaña baja	2	Glacial
3	Montaña seca	3	Esquisto
4	Montaña	4	Arenisca
5	Montaña seca y montaña	5	Sedimentario Mixto
6	Montaña y subalpino	6	No especificado
7	Subalpino	7	Ígneo y metamórfico
8	Alpino	8	Volcánico

Tabla 5: Especificación de los dígitos del código ELUs.

Ahora, considerando las variables continuas, podemos analizar la distribución de los datos utilizando gráficos de violin. Estos gráficos se muestran en las Figuras 3, 4, 5.

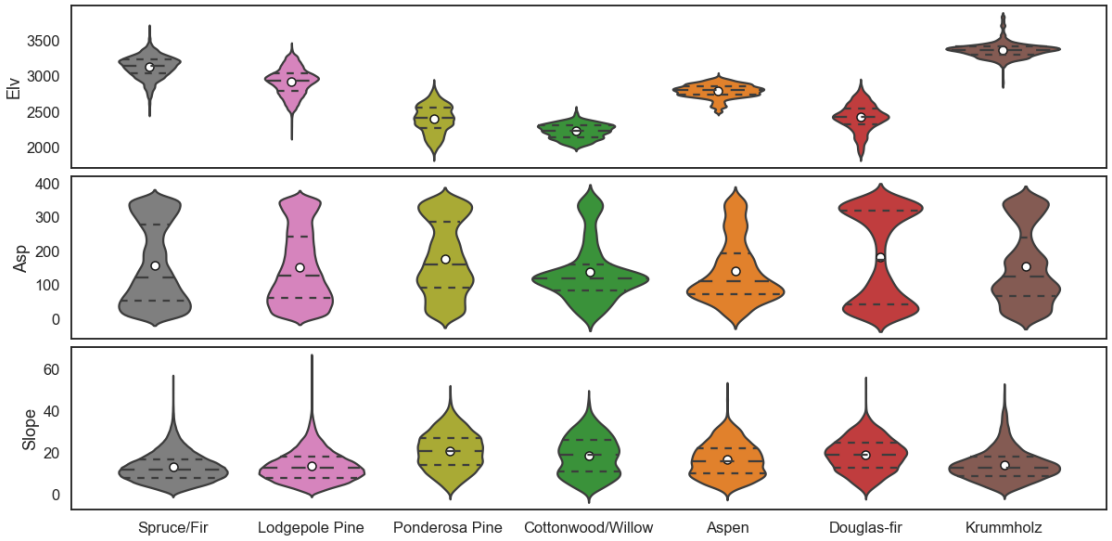

Figura 3: Gráficos de violín para las variables Elevación (Elv), Exposición (Asp) e Inclinación (Slope). El punto blanco indica la media de los valores, mientras que las lineas discontinuas indican el primer, segundo y tercer cuartil.

Como se observa en la Figura 3, de las tres variables, la que genera mayor diferencia entre los gráficos es la Elevación (Elv). Exciste por ejemplo una clara separación entre los árboles que están a mayor altura como Spruce/Fir y Krummholz comparado con los que están a menor altura, como Cottonwood/Willow, Ponderosa Pine y Douglas-fir. Además, algunas especies tienden a concentrar sus valores más cercanos a la media, generando una distribución más puntiaguda, como es el caso de la especie Krummholz. En cuanto a la Exposición (Asp) se generan distintos tipos de distribución, Spruce/Fir con Lodgepole Pine, Cottonwood/Willow con Aspen, Ponderosa Pine con Krummholz; mientras que Douglas-Fir es algo diferente. Esta variable nos muestra que los distintos árboles prefieren estar a un lado del cerro o del contrario. Finalmente la variable Pendiente (Slope) nos muestra que la mayoría de los árboles prefieren bajos valores de pendiente, siendo las especies Ponderosa Pine, Cottonwood/Willow, Aspen y Douglas Fir las que como media están por sobre las otras especies, aunque se debe señalar que la especie Lodgepole Pine llega a valores más extremos.

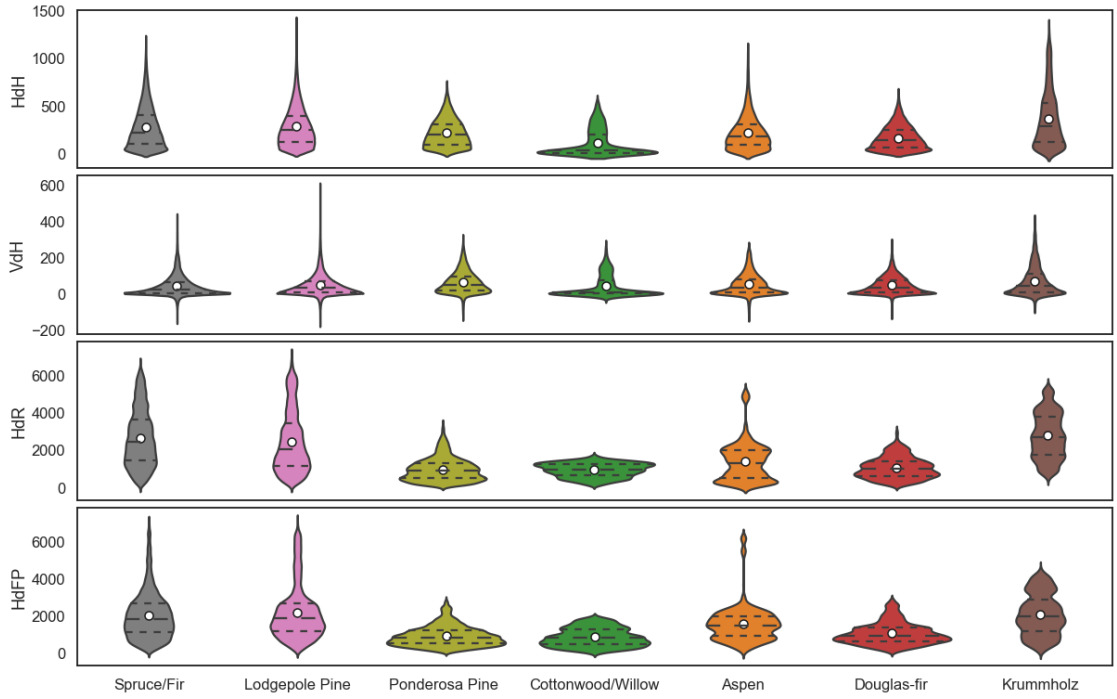

Figura 4: Gráficos de violín para las variables Distancia horizontal a cuerpos de agua (HdH), Distancia vertical a cuerpos de agua (VdH), Distancia horizontal a carreteras (HdR) y la Distancia horizontal a puntos de incendio (HdFP). El punto blanco indica la media de los valores, mientras que las lineas discontinuas indican el primer, segundo y tercer cuartil.

En cuanto a la distribución de distancias mostradas en la Figura 4, nos damos cuenta que en general las distribuciones son parecidas, la mayoría de los árboles tienen preferencia a estar a distancias bajas, pero en el caso de la categoría Cottonwood/Willow se aprecia una preferencia especial por estar más cerca a cuerpos de agua. En cuanto a la distancia a puntos de incendio y carreteras, se puede esperar cierta correlación entre ellas y con el valor de Elevación, debido a que es esperable que a mayores alturas mayor sea la distancia a carreteras, que usualmente no se construyen en grandes alturas) y a incendios (Existen pocas fuentes de incendio que puedan darse naturalmente a grandes alturas).

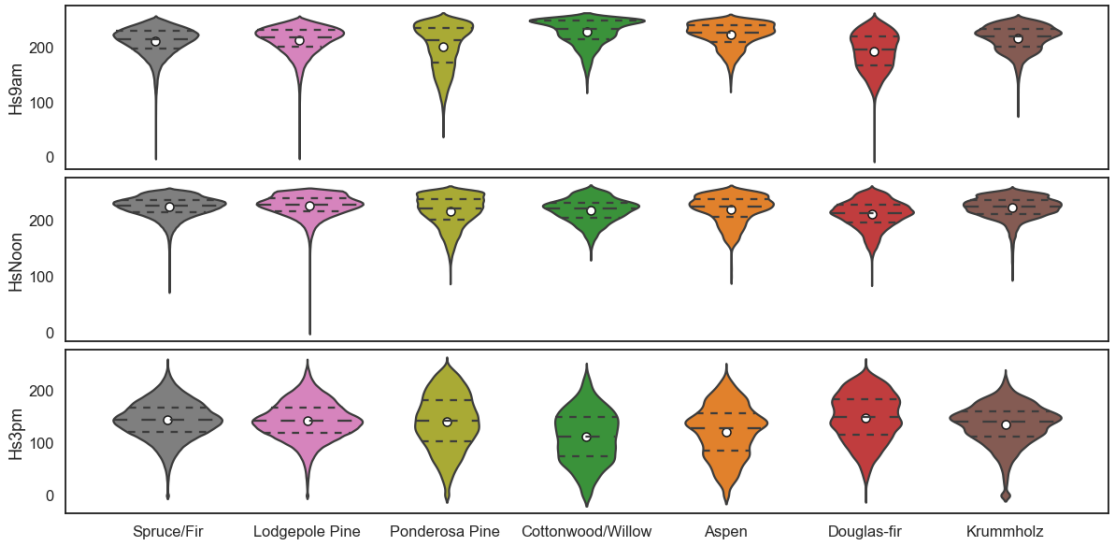

Figura 5: Gráficos de violín para las variables Sombtreado del relieve a las 9am (Hs9am), al medio día (HsNoon) y a las 3pm (Hs3pm). El punto blanco indica la media de los valores, mientras que las lineas discontinuas indican el primer, segundo y tercer cuartil.

Finalmente, al analizar el Sombreado del relieve se encuentran distribuciones muy parecidas. Se espera que estas variables tengan correlación con la variable Exposición (Asp), debido a que la dirección de la montaña donde se encuentran emplazados tiene directa relación con el ángulo con el cual inside la luz.

Finalmente si se analiza la correlación entre las variables mostradas en la Figura 6, son pocas las variables con un alto nivel de correlación, exceptuando casos como el sombreado del relieve a las distintas horas (Hs9am, HsNoon, Hs3pm) junto con la variable Exposición (Asp), esperable como se nombró anteriormente. Otras correlaciones ya moderadas, es entre la Elevación (Elv) y la Distancia horizontal a ríos (HdH) y carreteras (HdR), nuevamente, como se explicó anteriormente, aunque no son tan altas como uno esperaría, lo que indica que de estas variables se espera implicancia en las predicciones.

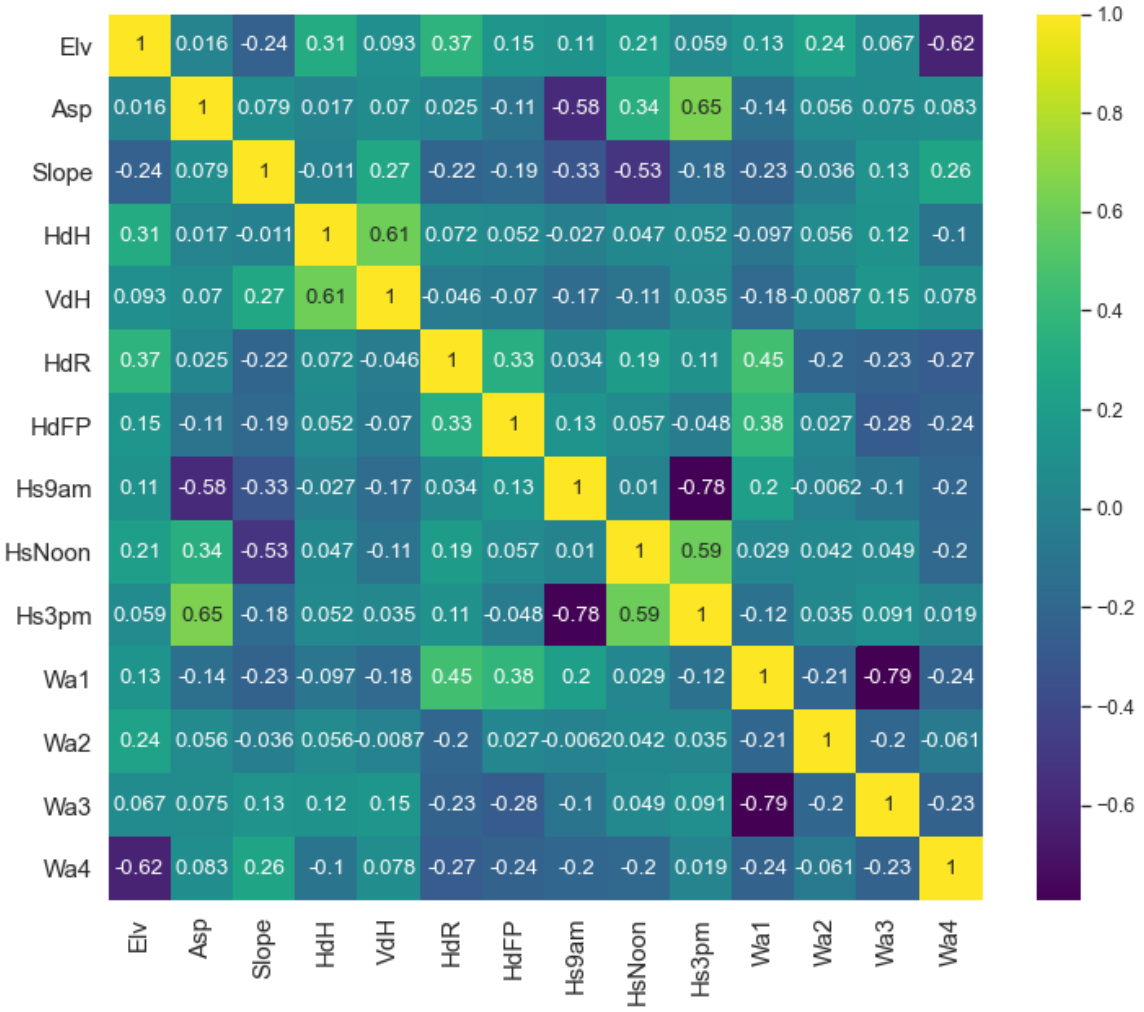

Figura 6: Heatmap de las correlaciones entre las distintas variables. No se incluye Clase de suelo.

4. Marco Conceptual

A continuación se se presentará el marco conceptual para este trabajo, cubriendo las áreas más importantes para comprender el informe y señalando las herramientas utilizadas para resolver el problema de clasificación.

4.1. Modelos

Para resolver el problema de clasificación, se implementaron los algoritmos de machine learning utilizando el lenguaje de programación Python en su versión 3.8 junto con las librerías Scikit Learn y LightGBM y sus modelos Random Forest [23] y Light Gradient Boosting Machine (LGBM) [14] respectivamente. A continuación, se describen los algoritmos utilizados:

4.1.1. Random Forest

Random forest es uno de los modelos de ensamblado más clásico y poderoso utilizando como base los árboles de decisión. Para evitar el sobre ajuste se usa la técnica llamada bagging, que consiste de la selección aleatoria de un subset del conjunto de datos disminuyendo la varianza sin incrementar el sesgo y reduciendo la correlación entre los distintos árboles. En problemas de clasificación, la clase final es determinada por votación por mayoría. Debido a que está basado en árboles de decisión hereda muchas de sus cualidades, por ejemplo, no requiere escalado ni one-hot encoding de las variables, tiene un buen rendimiento al utilizar set de datos desbalanceados y puede capturar relaciones no lineales entre variables dependientes e independientes. Una de las desventajas de utilizar Random Forest es el gran uso de recursos computacionales al ir incrementando el número de estimadores, en especial memoria, comparado con otros modelos más simples como una regresión logística [23].

4.1.2. Light Gradient Boosting Machine

LightGBM nació como parte del proyecto de Microsoft's Distributed Machine Learning Toolkit (DMTK). Está diseñado para ser rápido y distribuido, resultando en un entrenamiento más rápido que Random Forest y utilizando menos memoria. Soporta GPU y entrenamiento en paralelo para manejar gran cantidad de datos. Está basado en árboles de decisión de gradiente descendente (Gradient boosting decision tree o GBDT). A diferencia de otros métodos basado en GBDT, que necesitan para cada variable escanear todas las instancias para estimar la ganancia de información de todos los posibles puntos de separación, LightGBM utiliza el algoritmo GOSS (Gradient-based One-Side Sampling), donde las variables con altos gradientes, que contribuyen más a la ganancia de información, son elegidos por sobre aquellas variables con bajos gradientes. Además del algoritmo GOSS utiliza el algoritmo EFB (Exclusive Feature Building) para combinar variables y así reducir la complejidad al momento de entrenar [14].

Para este trabajo se implementó Focal Loss [16] como función de pérdida, ampliamente utilizada en redes neuronales para datos no balanceados, pero no implementada en la librería Lightgbm. Esta función de pérdida se puede observar en la ecuación 1.

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{1}$$

Esta ecuación es similar a la función de pérdida de entropía cruzada, pero se adicionan parámetros para dar o quitar importancia a las distintas categorías durante el cálculo de la función de pérdida. Los parámetros en la función Focal Loss son dos: alfa (α) y gamma (γ). Alpha funciona como peso para las distintas categorías, pero no ayuda en considerar el desbalance de ellas, mientras que gamma adiciona o quita valor de pérdida a las predicciones según que tan segura sea la predicción: mientras mayor sea el valor de gamma más importancia se le dará a las categorías mal clasificadas.

4.2. Generación de datos utilizando CTGAN

Las redes GANs (Generative Adversarial Networks) son modelos utilizados para generar información sintética, principalmente imágenes, con resultados muy similares a que fuese real. Las redes GANs utilizan como base dos redes que forman una estructura generador-discriminador, de tal forma que la red generadora crea información falsa y la discriminadora evalúa si lo es. En cada iteración del algoritmo, la red generadora generará información falsa que a cada iteración es más parecida a la real, mientras que la red discriminadora mejorará su capacidad de distinguir la información real de la falsa. Una visualización de este proceso, se presenta en la Figura 7.

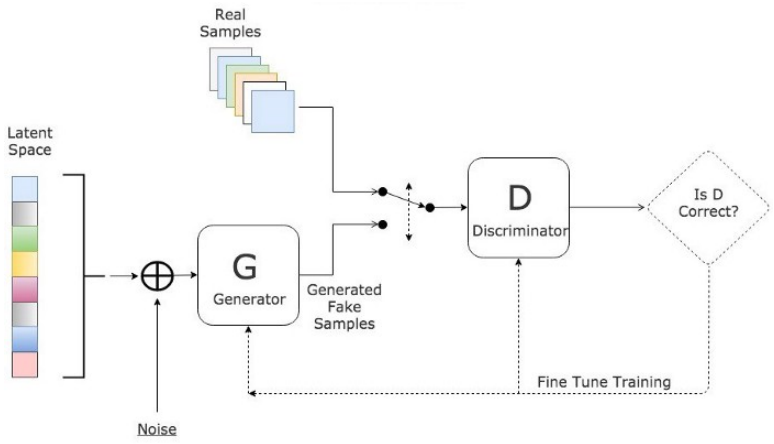

Figura 7: Estructura generador-discriminador de las redes GANs. A partir de un tensor aleatorio se genera información sintética, que luego ingresa junto a información real al discriminador. Con los resultados se realiza propagación hacia atrás para ajustar los parámetros del modelo con tal de reducir una función de pérdida definida por el problema a enfrentar.

Las redes GANs son muy usadas en generación de imágenes, aunque en los últimos años su uso se ha expandido y ha incluido la utilización y generación de datos tabulares [32, 31, 1], obteniendo mejores resultados que otras metodologías como redes bayesianas [31]. La red GAN que ha obtenido mejores resultados es el modelo CTGAN (Conditional Tabular GAN), que realiza una mezcla gaussiana variacional para normalizar datos continuos y realiza un entrenamiento por muestreo para evitar la desaparición de las categorías minoritarias y con la posibilidad de generar predicciones condicionales.

4.3. Métricas e interpretabilidad del modelo

Para la evaluación de cada uno de los modelos se utilizó como métrica los valores de accuracy, F1 y área bajo la curva ROC (AUC). Se analizaron tanto los valores promedio de cada métrica, sin ajustarlos por los pesos, como los valores por categoría. Esto último nos permite analizar los resultados tomando importancia de las categorías minoritarias.

Para la interpretación de los resultados entregados por los modelos realizamos un análisis de los valores Shapley utilizando la librería SHAP [18]. Los valores Shapley permiten entender qué variables (atributos o columnas) tienen más influencia en las predicciones de un modelo de machine learning. La definición de los valores Shapley descansa en la teoría de juegos y busca responder

cuál es la manera más justa de repartir una recompensa entre los jugadores que ganaron un juego considerando dos propiedades descritas a continuación:

- Aditividad: La suma de las distintas recompensas es la suma la recompensa del juego final.
- Consistencia: El jugador que más contribuyó debe obtener una recompensa mayor.

En el caso de ganancia de dinero en un juego f de recompensa en equipo de M jugadores, los valores Shapley serían la cantidad de dinero x ganado por cada participante i del juego según su contribución a este. La ganancia se calcula promediando la contribución marginal de cada jugador considerando todas las posibles secuencias S de los jugadores. Para el caso de machine learning podemos hacer una analogía y reemplazar el término jugadores con el termino características o atributos, el término recompensa con predicción y el término juego con modelo y entender como un atributo único i genera un aporte a la predicción x. El calculo de los valores Shapley puede se puede resumir en la ecuación 2.

$$\begin{array}{c}\text{Valor Shapley para el atributo i o}\\ \text{contribución del atributo i}\\ \text{en la predicción del modelo f(x)}\end{array} = \begin{array}{c}\text{Promedio sobre todos}\\ \text{los subset de atributos}\\ S \subseteq M/\{i\}\end{array} \left(f(S \cup \{i\}) - f(S) \right) \quad (2)$$

5. Metodología

La metodología consistió en la implementación de los algoritmos de machine learning y la generación de modelos de manera iterativa. Los algoritmos utilizados y los detalles del procedimiento realizado se describen a continuación.

El procedimiento realizado se basó primero, en probar si el problema planteado se puede solucionar utilizando modelos basados en árboles de decisión y segundo, en encontrar mejoras para el modelo, con especial enfoque a mejorar las métricas de evaluación de las categorías minoritarias o en su defecto, mejorar la métrica F1.

En primera instancia se abordó el problema de desbalanceo de las categorías a predecir. Para ello se eligió un set de datos de entrenamiento y validación que contenga una cantidad equilibrada de los datos considerando una separación de entrenamiento y validación del 50 % estableciendo que cada clase tenga aproximadamente la misma proporción. Además, se seleccionaron los parámetros de los distintos algoritmos adecuados a muestras no balanceadas y se utilizó validación cruzada de 5k-folds.

Los modelos fueron seleccionados según el problema de clasificación a enfrentar. Existen varios modelos, desde los más simples como una regresión logística a otros como métodos basados en árboles de decisión y gradiente descendiente. Para este trabajo se usaron estos últimos, debido a que funcionan bien en sistemas altamente no lineales como es el crecimiento de bosques [19]. para generar una comparación entre la búsqueda o no de hiperparámetros. Se eligieron parámetros en común:

- N° de estimadores (*n_estimators*): Al aumentar el número de árboles se crea un modelo agregado más robusto, con menos varianza y que ayuda a reducir el error en las predicciones, pero con el coste de un mayor tiempo de cálculo.
- N° de hojas (*max_leaf_nodes- num_leaves*): Al incrementar la máxima profundidad incrementa el número de posibles combinaciones atributo/valor llevando a más interacciones entre variables. Selecciona su separación según a la contribución global en la función de pérdida.
- Profundidad (*max_depth*): Efecto similar al n° de hojas, pero la separación es según la contribución a la función de pérdida de cada rama en particular, aprendiendo más lento de los errores que al aumentar el número de hojas [29].
- Peso de las categorías:

La búsqueda de hiperparámetros se realizó de forma codiciosa en búsqueda de la mejor métrica F1, que implica probar todas las combinaciones de hiperparámetros. Además, para permitir la comparación, se generaron modelos que solo consideraban el cambio del hiperparámetro n° de estimadores a 200 . Los valores utilizados en la búsqueda codiciosa se presentan en la Tabla 6.

Hiperparámetro	**Valores**
N° estimadores	100-200-300-600-1000-2000-6000
N° de hojas	20-50-100-200-500
Profundidad	10-20-30-50

Tabla 6: Valores utilizados en la búsqueda codiciosa de hiperparámetros para los modelos Random Forest y LightGBM.

A partir de los resultados, se realizó un análisis de la influencia de los hiperparámetros en las métricas del modelo. Se analizaró como el aumento de estos tres hiperparámetros en solitario y en conjunto afectaron las métricas de evaluación de los modelos, utilizando rangos similares a los

presentados en la Tabla 6. Además, se consideró el estudio de seleccionar distintas proporciones en el set de entrenamiento, variando de 10 hasta 90 % para analizar el efecto de los datos en el entrenamiento final y fijando hiperparámetros en los valores de 1000 estimadores, 200 hojas y 10 en profundidad. A partir de los resultados del análisis se procedió a buscar mejoras del modelo con tal de mejorar las predicciones de las categorías minoritarias.

Para buscar estas mejoras se utilizó la implementación de la función de pérdida Focal Loss en el algoritmo LightGBM [4] con parámetros α y γ de 0.5 y 1.5 respectivamente; se analizó su rendimiento en cuanto a métricas de evaluación y se comparó con los modelos ya generados.

Para manejar el desbalanceo de la información se generó información sintética utilizando redes CTGAN. Para entrenar el modelo generativo se utilizó toda la información de entrenamiento sin considerar la de testeo. Se optó por generar solo las dos categorías minoritarias como condición: Aspen y Cottonwood/Willow. Para evitar los efectos de la aleatoriedad de la información sintética se generó cinco set de datos sintéticos y se combinó con el set de entrenamiento, se realizó el entrenamiento utilizando el algoritmo LightGBM en cada uno de estos dataset, se obtuvo las métricas y se promediaron para estos cinco dataset. Como resultado final se utilizó la función de pérdida Focal Loss junto a la información sintética generada y se comparó los modelos generados.

Finalmente para el análisis de la interpretabilidad del modelo se utilizó la librería SHAP. A diferencia de otras metodologías de análisis de importancia de las variables como *permutation feature importance* que analizan la importancia según el descenso en el rendimiento del modelo, la metodología SHAP se basa en la magnitud de las atribuciones de características [20], lo que nos además permite generar gráficas más ilustrativas de como las variables influyen a través de las distintas categorías a predecir.

Para realizar los cálculos se utilizaron todos los datos y el mejor modelo generado hasta el momento. Con los valores Shapley se realizó el análisis general del modelo, el análisis particular por categoría y cuatro predicciones en específico. El análisis se basó en como afectaron los atributos (o columnas) en la predicción final, analizando si fue un aporte negativo o positivo en la predicción.

6. Resultados

En esta sección, presentamos los resultados de los modelos generados. Los resultados son presentados según las siguientes secciones:

- Resultados de Random Forest y LightGBM.
- Resultados de LightGBM utilizando Focal Loss como función de pérdida.
- Resultados generando información sintética con CTGAN.
- Análisis del modelo utilizando valores Shapley.

6.1. Resultados de Random Forest y LightGBM

A continuación se presentarán los modelos iniciales generados y sus resultados. Se generaron dos modelos basado en el algoritmo Random Forest: RF1 y RF2, y dos modelos basados en LightGBM: LGBM1 y LGBM2. Los hiperparámetros de los modelo base, RF1 y LGBM1, y los hiperparámetros encontrados con búsqueda codiciosa se resumen en la Tabla 7.

Modelo	**Hiperparámetro**	**Valor**
RF1	n_estimators	200
RF2	class_weight	balanced
	max_depth	50
	max_leaf_nodes	None
	n_estimators	200
LGBM1	n_estimators	200
LGBM2	class_weight	balanced
	max_depth	50
	num_leaves	500
	n_estimators	6000

Tabla 7: Hiperparámetros utilizados en la generación de modelos. RF1 y LGBM1 tomaron como base 200 estimadores; RF2 y LGBM2 se obtuvieron mediante búsqueda codiciosa de hiperparámetros.

Matrices de confusión y métricas

A continuación presentamos los resultados para los modelos RF1, RF2, LGBM1 y LGBM2. La Tabla 8 contiene las métricas de evaluación para el set de datos de entrenamiento y testeo obtenidas para RF1, RF2, LGBM1 y LGBM2. Observamos que para los modelos RF1, RF2 y LGBM2 se obtienen altos valores (mayor a 0.9) para las métricas accuracy, F1 y AUC, siendo el modelo LGBM2 el que presenta mejores resultados para el set de Test (o prueba) con 0.969 de accuracy, 0.945 para

F1 y 0.999 para AUC. De esta tabla también observamos que tres de los cuatro modelos, RF1, RF2 y LGBM2, puede presentan sobre ajuste (overfitting) para el set de datos de entrenamiento al obtener métricas de accuracy 1. Este comportamiento es discutido más adelante en la subsección *Influencia de los hiperparámetros en el sobre ajuste.*

	Accuracy		**F1**		**AUC**	
Modelo	Entrenamiento	Test	Entrenamiento	Test	Entrenamiento	Test
RF1	1.0	0.945	1.0	0.912	1.0	0.997
RF2	1.0	0.946	1.0	0.912	1.0	0.997
LGBM1	0.888	0.869	0.911	0.844	0.985	0.977
LGBM2	1.0	0.969	1.0	0.945	1.0	0.999

Tabla 8: Resultados obtenidos al utilizar Random Forest y LightGBM

Para analizar las predicciones de los modelos RF2 y LGBM2 por categoría presentamos en la Figura 8 y en la Tabla 9, que contienen las matrices de confusión y las métricas de evaluación para las distintas categorías respectivamente. A partir de la Figura 8 notamos que la diagonal de las matrices de confusión contienen la mayor parte de las clasificaciones, lo que indica un buen desempeño del modelo. Sin embargo, se observa que para las categorías Cottonwood/Willow y Aspen existe una gran proporción de falsos positivos y verdaderos negativos, que se refleja en métricas bajas de F1 para esas categorías (ver Tabla 9). Debido al desbalanceo de los datos se tomó preferencia el analizar la métrica F1 por sobre la métrica accuracy.

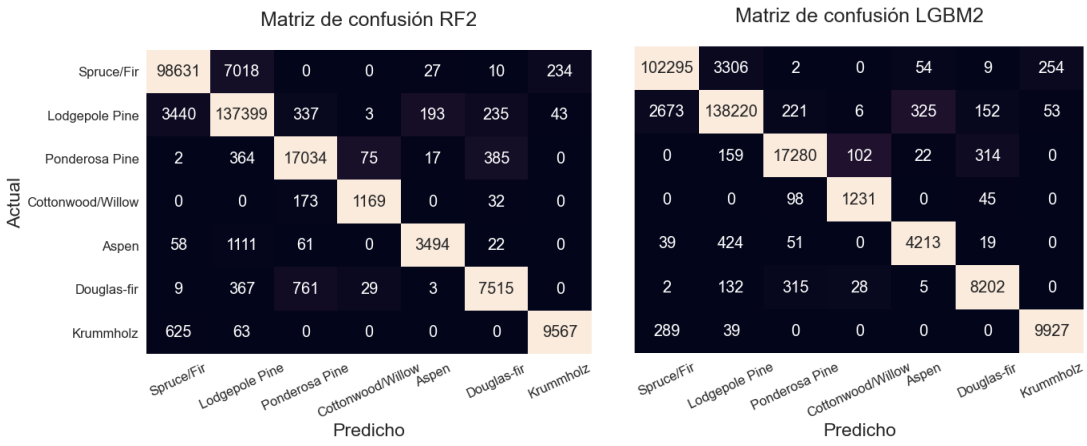

Figura 8: Matriz de confusión de los modelos RF2 y LGBM2

	RF2		**LGBM2**	
Categoría	F1	AUC	F1	AUC
Spruce/Fir	0.945	0.994	0.969	0.998
Lodgepole Pine	0.954	0.992	0.974	0.997
Ponderosa Pine	0.939	0.999	0.964	1.0
Cottonwood/Willow	0.878	0.999	0.898	0.999
Aspen	0.826	0.998	0.900	0.999
Douglas-fir	0.892	0.998	0.941	0.999
Krummholz	0.951	1.0	0.969	1.0
Promedio	0.912	0.997	0.945	0.999

Tabla 9: Valores F1 y AUC para las distintas categorías de los modelos RF2 y LGBM2

Al analizar las métricas de F1 y AUC para las distintas categorías mostradas en la Tabla 9, observamos en su mayoría valores sobre 0.9, siendo el valor máximo de F1 de 0.974 para la categoría Lodgepole Pine en el modelo LGBM2. Para la categoría Cottonwood/Willow, que corresponde a las categoría con menor cantidad de datos, se obtuvo valores F1 menores o cercanos a 0.9 tanto para el modelo RF2 y LGBM2. Además, para todas las categorías se obtuvieron altos valores de la métrica AUC. Estos resultados revelan la posibilidad de resolver el problema de clasificación utilizando modelos basados en árboles de decisión pero con la dificultad de separar las categorías minoritarias por sobre las otras categorías, es por esto que más tarde en el informe en las subsecciones *Resultados de LightGBM utilizando Focal Loss como función de pérdida* y *Resultados generando datos sintéticos con CTGAN* se analizó posibles metodologías para mejorar estas métricas y se tomará especial cuidado al tomar valores tan extremos de los hiperparámetros.

Influencia de los hiperparámetros en el sobre ajuste

Al analizar los valores de accuracy, F1 y AUC para los modelos RF1, RF2 y LGBM2 presentados en la Tabla 8, observamos puede existir sobreajuste para los datos de entrenamiento. A pesar de que no existen reglas estrictas para seleccionar hiperparámetros, es sabido que un aumento en los valores de ciertos hiperparámetros generan sobreajuste en los modelos, generando un efecto perjudicial en las predicciones que se refleja en la disminución de las métricas de evaluación para el set de datos de testeo.

Para estudiar el efecto de los parámetros (e hiperparámetros) en el modelo y analizar si al aumentarlos existe un efecto perjudicial en las predicciones, se realizó la exploración en las métricas de evaluación del set de testeo al cambiar los parámetros, para visualizar posibles descensos en las

métricas. Esta exploración se realizó utilizando un modelo de LightGBM variando los parámetros: n° de estimadores (*n_estimators*), n° de hojas (*num_leaves*) y produnfidad (*max_deph*), hiperparámetros conocidos en la práctica por generar sobre ajuste. Además, se exploró el efecto de seleccionar distintas proporciones del set de entrenamiento para estudiar la dificultad de generalizar los datos. Los resultados de la exploración se pueden observar en la Figura 9.

En primera instancia, si analizamos el efecto de seleccionar distintas proporciones del set de entrenamiento, se evidencia que a menor proporción, peores resultados, pero para ser solo 10 % de los datos para entrenar, se parte con una buena generalización de estos, con accuracy por sobre 0.9 y métricas F1 cercanas a 0.9. Por otro lado, si se analizan los hiperparámetros, tanto el número de estimadores y hojas por si solos, no logran ajustar perfectamente los datos de entrenamiento, pero se logran métricas de evaluación cercanas a 1.0, evidenciando que están a pasos de generar sobre ajuste, pero no teniendo un efecto perjudicial en las métricas de evaluación del set de testeo. El mayor cambio de las métricas se observa entre 100 - 2000 estimadores y 20 - 200 hojas. Por otro lado el hiperparámetro profundidad del árbol no genera un cambio en las métricas.

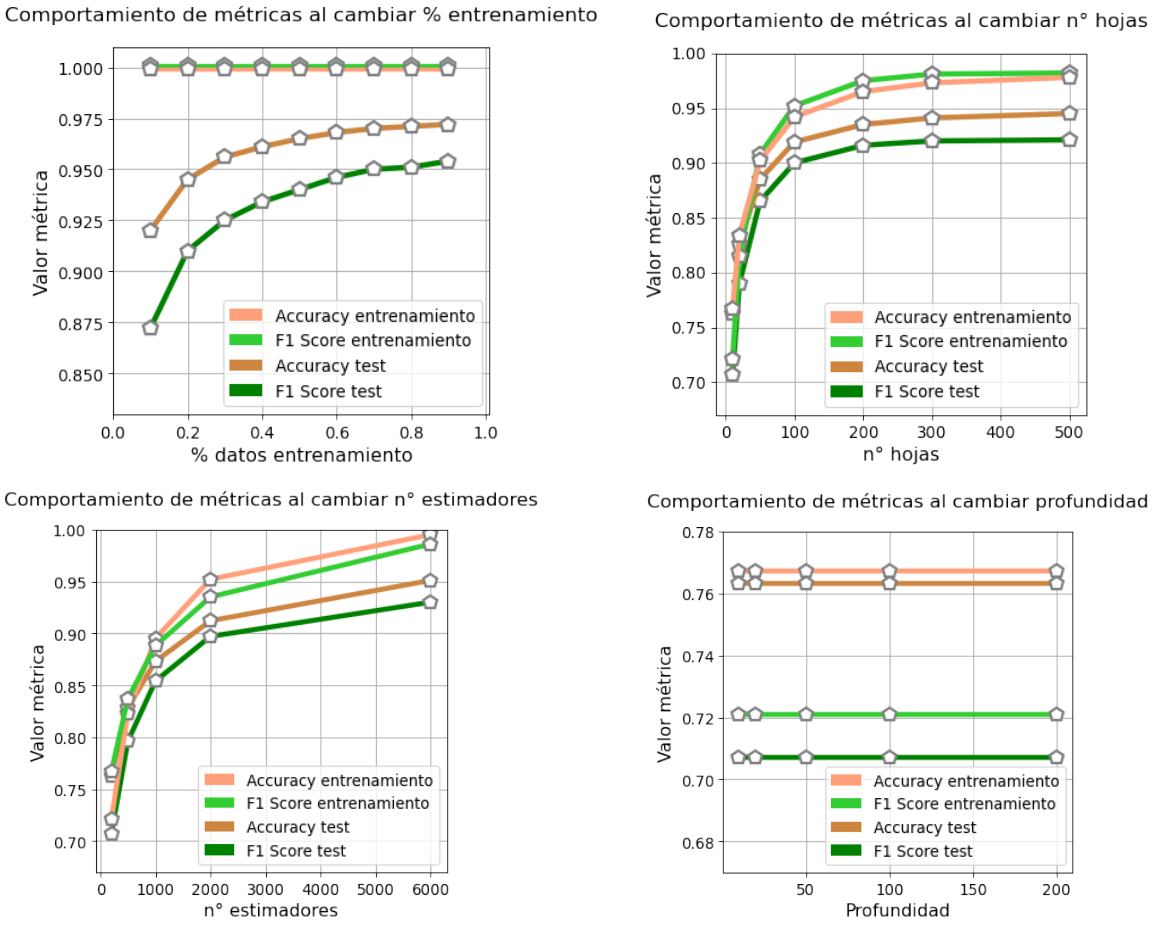

Figura 9: Variación de métricas accuracy y F1 score en datos de testeo al variar hiperparámetros de sobre ajuste para el modelo Ligth Gradient Boosting Machine.

Análisis de hiperparámetros en conjunto

Ya conociendo la existencia de dos hiperparámetros que pueden generar sobre ajuste: número de estimadores y hojas, se realizó su análisis en conjunto. Este estudio se puede observar en la Figura 10, donde se evaluó el comportamiento de las métricas F1 y accuracy al incrementar el número de estimadores y hojas. Dado los resultados anteriores, se realizó el análisis entre 100 - 1000 estimadores y 20 - 200 hojas. Además se marcó los pentágonos de color amarillo cuando se obtiene una métrica de accuracy por sobre 0.96 en el set de entrenamiento.

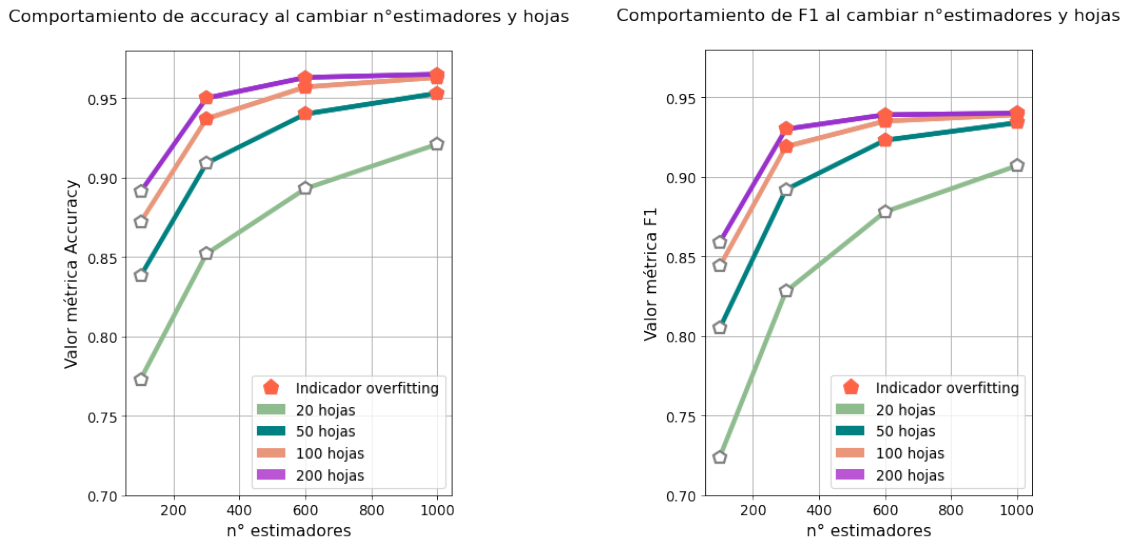

Figura 10: Comparación en métricas de testeo para distintos valores de hojas y número de estimadores.

Se observa que a mayor cantidad de hojas y estimadores siempre mejoran las métricas de evaluación y que en conjunto generarán un modelo con sobre ajuste perfecto en el set de entrenamiento. Llega un punto donde el aumento de hojas no genera un gran aumento comparado con el aumentar el número de estimadores, lo que si, en tiempo de ejecución, es mucho más lento aumentar los estimadores que el número de hojas. Además, el sobre ajuste se presencia cuando consideramos sobre las 20 hojas, existiendo un gran salto en las métricas al pasar de 20 a 50 hojas. Además, al considerar valores entre 100 y 200 hojas se llega rápidamente, en una menor cantidad de estimadores, al sobre ajuste de los datos de entrenamiento.

Con los resultados obtenidos evidenciamos que aun existiendo la presencia de sobre ajuste, este no genera un efecto perjudicial a la hora de realizar las predicciones. Se generaron modelos con bajo sesgo y varianza, debido posiblemente al tipo de datos utilizado, donde árboles crecen en condiciones similares a los de su misma especie, no generando errores, por el contrario, generando mejoras en el modelo al sobre ajustar los datos en el set de entrenamiento. Con esto en mente, pero siendo precavidos y evitando tomar valores de hiperparámetros tan extremos, para los siguientes análisis se tomará como estándar utilizar LightGBM como modelo y fijando el hiperparámetro profundidad en 10, el número de hojas en 200 y llegar hasta el límite de 1000 en el número de estimadores,

además, esto ayudará en reducir el tiempo de cálculo al momento de analizar la interpretabilidad del modelo, ya que la complejidad computacional tiene como máximo la función $O(TLD^2)$, donde T es el número de estimadores, L el de hojas y D la profundidad.

Resultados por categoría

Finalmente se realizó el análisis de la métrica de F1 sobre las distintas categorías al variar el número de estimadores y fijar en 200 el número de hojas como se observa en la Figura 11 a la izquierda, donde se utiliza la función de pérdida predefinida Log Loss. De la figura se observa un rápido estancamiento de la métrica F1 luego de los 300 estimadores para la mayoría de las categorías, exceptuando Cottonwood/Willow que presenta una leve disminución. Además, las dos categorías con menos datos, Aspen y Cottonwood/Willow, presentan una peor evaluación con curvas por debajo de las otras categorías con valores bajo 0.9 en la métrica F1. Se debe notar nuevamente que no existe un efecto perjudicial generalizado con la existencia de sobre ajuste, exceptuando la categoría Cottonwood/Willow.

6.2. Resultados de LightGBM utilizando Focal Loss como función de pérdida

A pesar de que las métricas de evaluación obtenidas por los modelos anteriores fueron en general altas, para algunas categorías minoritarias como Aspen o Cottonwood/Willow no se obtuvieron métricas de evaluación mayores a 0.9. En vista de estos resultados, decidimos mejorar el modelo de LightGBM incorporando una función de pérdida Focal Loss, que no está implementada oficialmente en el algoritmo LightGBM.

En la Figura 11 presentamos la relación entre el hiperparámetro número de estimadores y la métrica F1 para todas las categorías utilizando la función de pérdida Log Loss (panel izquierdo) y la función de pérdida Focal Loss (panel derecho).

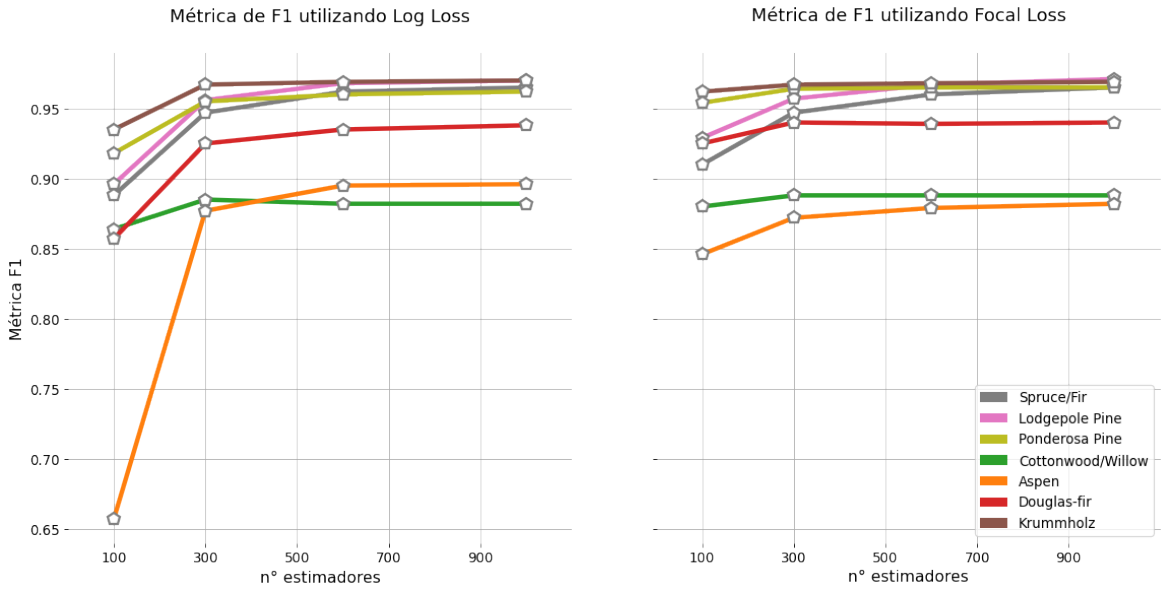

Figura 11: Relación entre el número de estimadores y F1 por categoría. El panel izquierdo utiliza como función de perdida Log Loss y mientras que el panel derecho utiliza la función Focal Loss. El número de hojas para ambos casos corresponde a 200.

A partir de la Figura 11 observamos que para valores del hiperparámetro *número de estimadores* entre los 100 y 300, el modelo que utiliza la función de pérdida Focal Loss genera mejores resultados, pero luego, para valores superiores a 300, observamos que el valor de F1 para las categorías minoritarias no es significativamente superior al valor obtenido utilizando la función Log Loss.

En vista del comportamiento de los modelos exhibido en la la Figura 11, generamos modelos con un valor para los hiperparámetros *número de estimadores*, *número de hojas* y *profundidad* de 1000, 200 y 10 respectivamente, uno usando la función de pérdida predefinida Log Loss y otro con Focal Loss. Las métricas de evaluación para estos modelos con estos hiperparámetros son desplegadas en la Tabla 10 con el nombre LogL y FocalL.

A parir de la Tabla 10 observamos que la mayoría de las categorías presentaron valores similares, mientras que las minoritarias Cottonwood/Willow subió 0.006 puntos, fue en prejuicio de la categoría Aspen que bajó 0.014 puntos, terminando con métricas promedio iguales, al menos hasta el tercer decimal. Con estos resultados no se puede concluir que utilizar la función Focal Loss llevó a una mejora, pero si podemos concluir que si es mejor utilizando una menor cantidad de estimadores.

6.3. Resultados generando datos sintéticos con CTGAN

Para manejar el desbalanceo de la información se propuso generar información sintética utilizando redes CTGAN a partir de los datos de entrenamiento. Se generó 3000 nuevos datos para Cottonwood/Willow y 3000 nuevos datos para Aspen . Al intentar con más información, o utilizar otras categorías, los modelos generados fueron peores o similares a utilizar la información original, por lo que estos casos no fueron considerados.

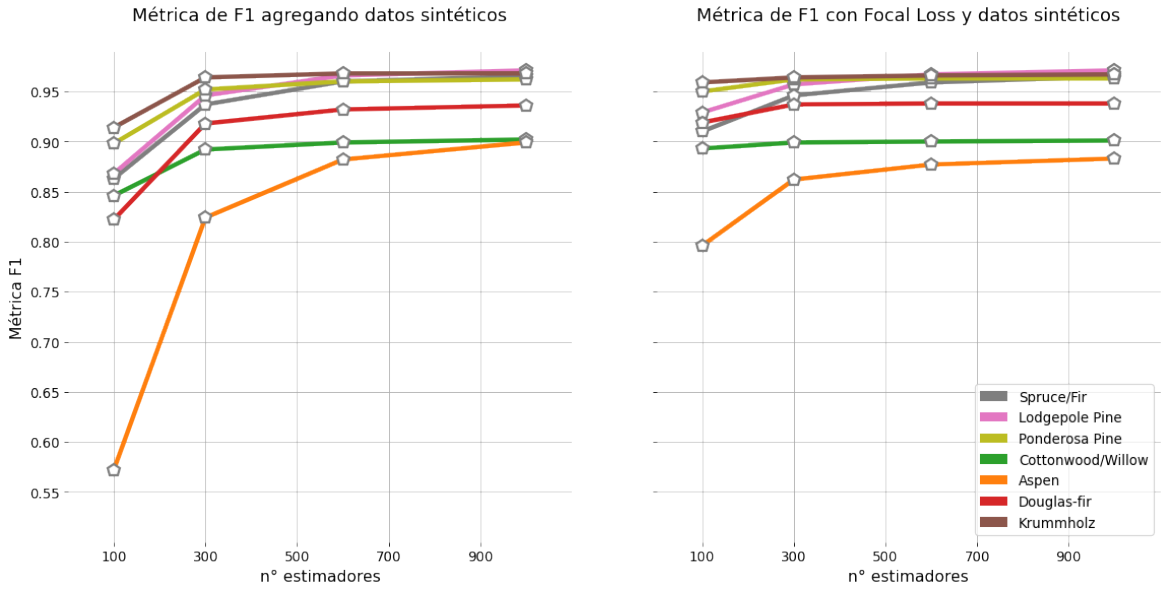

Figura 12: Métrica de F1 para las distintas categorías y valor fijo de n° hojas en 200 utilizando datos generados con CTGAN. A la izquierda utilizando Log Loss y al a derecha Focal Loss.

Los resultados obtenidos se pueden observar en la Figura 12. Estos resultados fueron similares a los obtenidos sin utilizar información sintética en cuanto a su tendencia, pero en cuanto a las métricas de evaluación se lograron mejores resultados finales que sin utilizar información sintética. Los resultados finales utilizando los mismos hiperparámetros que la comparación anterior se pueden observar en la Tabla 10 con el nombre GanLogL y GanFocalL, aunque debido a que la separabilidad de los datos a este punto es compleja, solo fue de 0.03 puntos en la métrica de F1 promedio y manteniéndose la métrica de accuracy comparado con el modelo base LogL. Con esta mejoría se logró llevar la métrica F1 de la categoría Cottonwood/Willow de 0.882 a 0.902 y la categoría Aspen de 0.896 a 0.899. Por otro lado, al intentar nuevamente con Focal Loss, los resultados fueron similares

a los anteriores: no tan bueno como utilizando la función de pérdida predefinida, pero mejores resultados a bajo número de estimadores.

Categoría	**LogL**	**FocalL**	**GanLogL**	**GanFocalL**
Spruce/Fir	0.965	0.965	0.965	0.965
Lodgepole Pine	0.970	0.971	0.971	0.971
Ponderosa Pine	0.962	0.965	0.962	0.963
Cottonwood/Willow	0.882	0.888	0.902	0.901
Aspen	0.896	0.882	0.899	0.883
Douglas-fir	0.938	0.940	0.936	0.938
Krummholz	0.970	0.969	0.968	0.967
F1 Promedio	0.940	0.940	0.943	0.941
Accuracy Promedio	0.965	0.965	0.966	0.965

Tabla 10: Valores F1 y accuracy para modelos generados utilizando GAN y/o Focal Loss. Se utilizó como base 1000 n°de estimadores, 200 hojas, 10 de máxima profundidad.

6.4. Análisis del modelo utilizando valores Shapley

Para realizar el análisis de interpretabilidad del modelo y salir de la caja negra, se realizó el análisis de los valores Shapley con la librería SHAP utilizando el modelo GanLogL entrenado con datos sintéticos. Debido a que realizar el análisis completo sería muy extenso, ya que se debe analizar las siete predicciones para cada fila de datos (son 7 probabilidades a analizar), solo se realizó para casos significativos. Se partirá realizando un análisis general, luego por cada categoría y finalmente se analizó 4 predicciones realizadas por el modelo, dos para la categoría Spruce/Fir y dos para Cottonwood/Willow, una correcta y una errada.

Análisis general

El análisis general se basó en considerar todos los valores Shapley para entender la contribución de los distintos atributos (o columnas del dataset) en la predicción de las siete categorías a predecir del dataset. Como estas contribuciones pueden ser negativas o positivas, se tomó la media de los valores absolutos de los valores Shapley calculados por la librería SHAP. Los resultados se pueden observar en la Figura 13, donde además se separa el aporte por categoría.

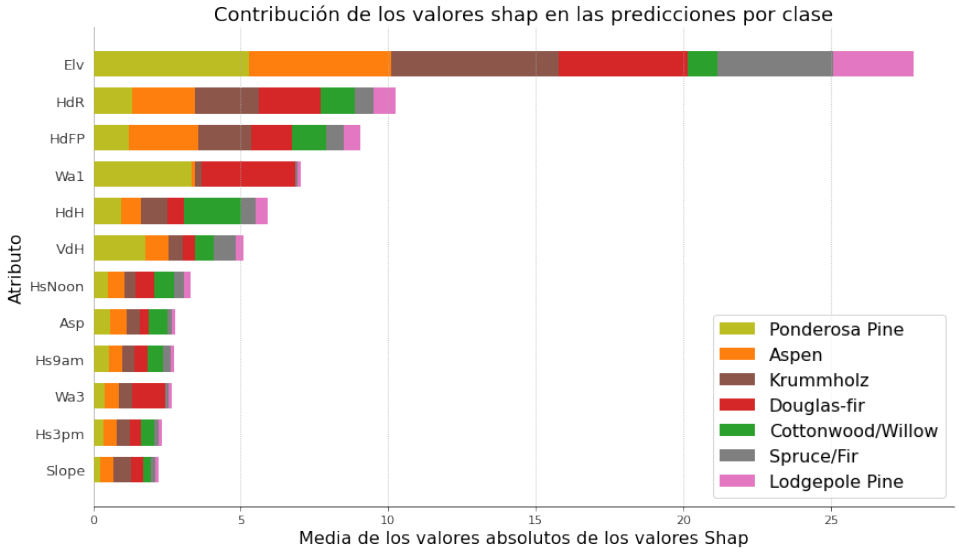

Figura 13: Barplot de las contribuciones de los valores shap al realizar las predicciones por atributo o columna. Se suman los valores por categoría y atributo, se promedian y se obtiene el valor absoluto.

Las contribuciones resultantes están ordenadas por importancia. Mientras mayor sea la media de los valores, mayor habrá sido la contribución de ese atributo en la predicción final de las siete categorías. De la Figura 13 se observa que el atributo que más contribuye en las predicciones es la *Elevación* (Elv), *Distancia horizontal a carreteras* (HdR) y *Distancia horizontal a puntos de incendio* (HdFP), resultados muy acordes a como crecen los árboles, el crecimiento de las distintas especies se basan en las condiciones del terreno, a más elevación se espera un terreno más seco, debido a que el agua escurre por la ladera; estar más cercano a carreteras podría ser un indicador de elevación ya que en general se encuentran a bajas alturas, o un indicador de influencia humana tal como lo podría ser la cercanía a puntos de incendios. Es interesante notar que muy pocas variables categóricas están en el top 10 de contribuciones, solo está *Área Silvestre 1* o Rawah (Wa1) y *Área silvestre 3* o Comanche Peak (Wa3), las dos categorías de área silvestre mayoritarias, aunque no se debe olvidar que el valores del gráfico consideran la media, por lo que no influye el tamaño de los datos, lo que si, a mayor cantidad es probable que el modelo pudiese aprender más de ellos. Por otro lado las categorías *Clase de suelo* (Stx) no están presentes, aunque no significa que no aporten en la predicción como veremos en los análisis siguientes.

Ahora, si se analiza por categoría en el gráfico general, la elevación no es predominante para la categoría Cottonwood/Willow, de hecho, la contribución de este atributo en la predicción final es

muy pequeña. Incluso en las categorías de *Área silvestre* anteriormente nombradas casi no aparece la categoría Cottonwood/Willow. Por otro lado, en cuanto a la *Distancia horizontal a cuerpos de agua* (HdH), esta categoría es mayoritaría, aportando incluso más que los otros atributos del modelo, esperable debido al gran consumo de agua de esta especie siendo un indicador importante de su presencia. Al igual que la categoría Cottonwood/Willow, Aspen es otra categoría minoritaria, aunque esta si aparece en el atributo *Elevación* (Elv), pero no se distingue aporte en el *Área silvestre 1* (Wa1).

Análisis por categoría

Para realizar las predicciones se debe partir de un valor base. Estos valores son los valores esperados para cada categoría, los que son calculados como el promedio de las probabilidades obtenidas del modelo para cada categoría y se puede entender como el valor predicho si no conociéramos ningún atributo de la instancia. Para nuestro modelo, estos valores son mostrados en la Tabla 11, donde las categorías más esperables son las dos mayoritarias.

Categoría	**Valor esperado**
Spruce/Fir	-3.62
Lodgepole Pine	-1.85
Ponderosa Pine	-23.50
Cottonwood/Willow	-25.35
Aspen	-24.72
Douglas-fir	-25.16
Krummholz	-23.69

Tabla 11: Valores esperados para cada categoría. En los gráficos se presenta como $E[f(X)]$.

A diferencia del modelo anterior, donde se sumaban las contribuciones a través de todas las predicciones, acá se realiza analiza como el modelo realiza la predicción de cada probabilidad, como son siete categorías, son siete probabilidades y los valores Shapley contribuyen distinto a cada categoría. Es por esto que se realizará un análisis de cada categoría utilizando beewarms plots, los cuales se pueden observar en las Figuras 14, 15, 16 y 17. Los beewarm plot nos muestra los valores Shapley para los 10 atributos más importantes a la hora de realizar la predicción. Cada punto es una instancia (o fila del dataset) y se apilan para mostrar la densidad de los datos; en el eje x se señala el valor Shapley y en el eje y los 10 atributos más importantes; el color nos indica si el valor para esa característica es alto o bajo.

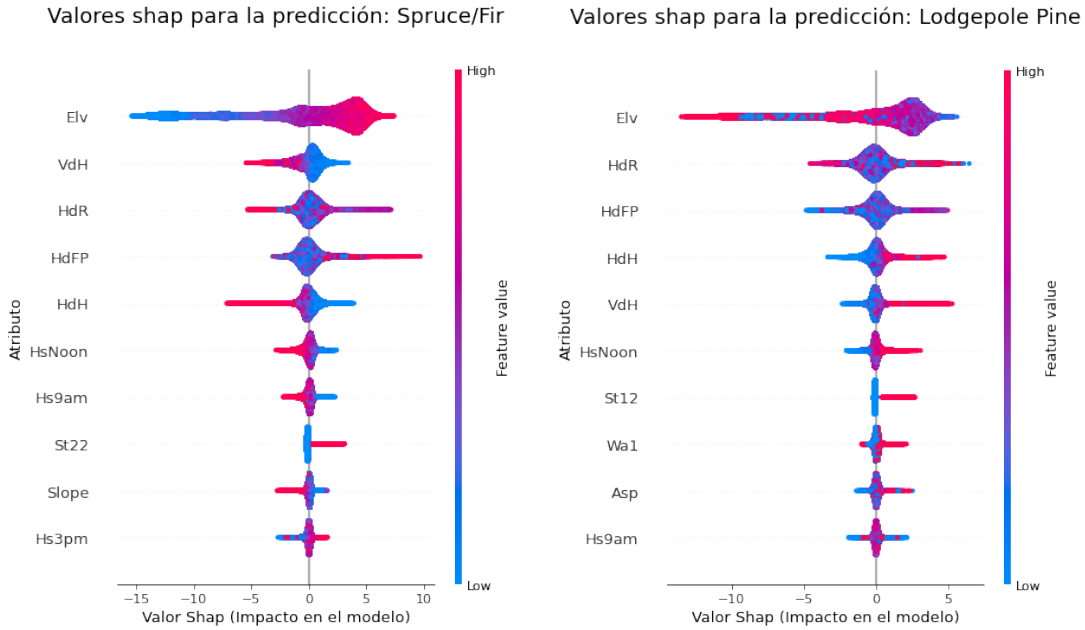

Figura 14: Beeplot para la predicción de las categorías Spruce/Fir y Lodgepole Pine

Tanto Spruce Fir y Lodgepole Pine son las dos categorías mayoritarias, aun así, las predicciones no se generan de la misma forma. Si analizamos el atributo Elevación, para la categoría Spruce/Fir la mayoría de los datos se encuentran a valores altos de *Elevación* (Elv) (alta densidad de color rojo), los que contribuyen positivamente a la predicción (al estar a la derecha del eje central), mientras que para la categoría Lodgepole pine, aunque tiene una distribución parecida a la de Spruce/Fir, los valores altos de elevación nos generan valores Shapley negativos (la cola más roja está a la izquierda del eje central), contribuyendo a que no sea esta categoría. Para los otros atributos la mayor densidad de datos está al centro, implicando poco aporte en cuanto a la predicción. Lo que si, aunque no aporten mucho a la predicción, son los atributos que terminan por decidir la categoría. Además, nos encontramos por ejemplo con el atributo *Distancia Vertical a Cuerpos de Agua* (VdH). Para la predicción de la categoría Spruce/Fir se encuentra como la segunda más importante, donde valores altos de este atributo genera valores Shapley negativos, aportando negativamente a la predicción de esta categoría, mientras que para la categoría Lodgepole Pine se encuentra en 5to lugar y a valores altos genera valores Shapley positivos y por lo tanto, ayudan a la predicción de estos datos. Finalmente los dos presentan un clase de suelo preferente, mientras que la categoría Spruce/Fir prefiere el *Tipo de Suelo 22*, la categoría Lodgepole Pine el *Tipo de Suelo 12*, lo que si, el no

pertenecer a estas categorías de suelo no aportan a la predicción, o no predicción, de estas categorías de cobertura de árbol, que en la figura se muestra como valores de bajo valor (cero o no perteneciente a esta categoría), están al centro del eje, no aportando a la predicción.

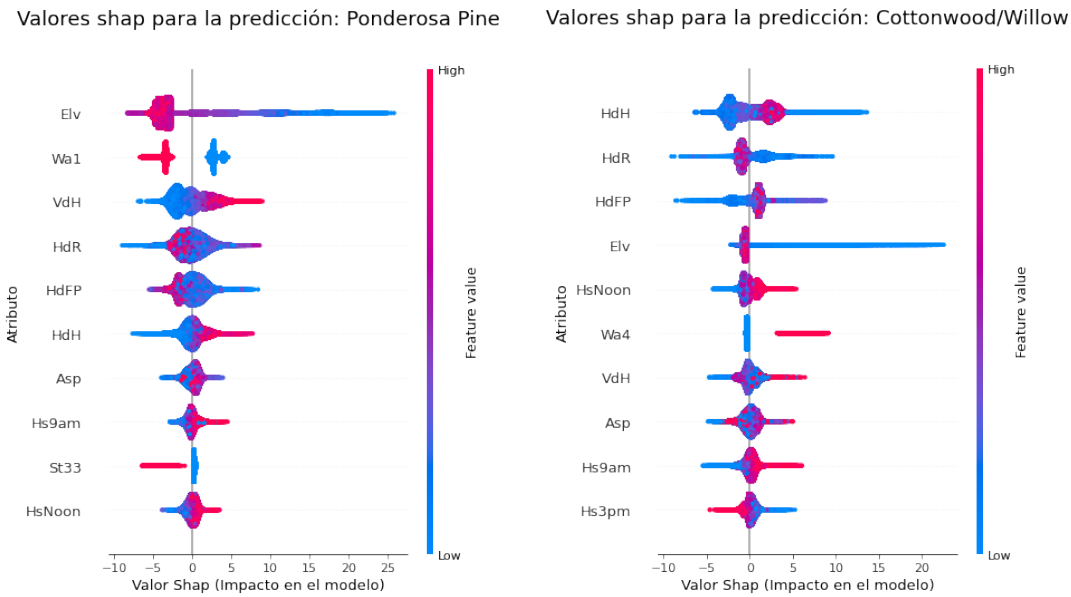

Figura 15: Beewarm plot para la predicción de las categorías Ponderosa Pine y Cottonwood/Willow

Otros comportamientos de los valores Shapley se pueden apreciar para la categoría Ponderosa Pine. Para la predicción, una de los atributos más importantes es el*Área silvestre 1* (Wa1), que claramente se separa en dos distribuciones separadas sin valores en el eje central, este comportamiento se produce cuando las variables categóricas son importantes al momento de tomar la decisión, indicando una predominancia de la cobertura de árboles en este tipo de *Área silvestre*, que si recordamos de la Figura 2, aun no siendo de las dos categorías mayoritarias, tenía una gran presencia en esta área. Un comportamiento similar con la mitad de importancia se puede apreciar en la categoría Cottonwood/Willow para el atributo *Área silvestre 4* (Wa4), donde aunque están separadas las distribuciones, solo una aporta en la predicción al tener valores Shapley lejanos a 0.

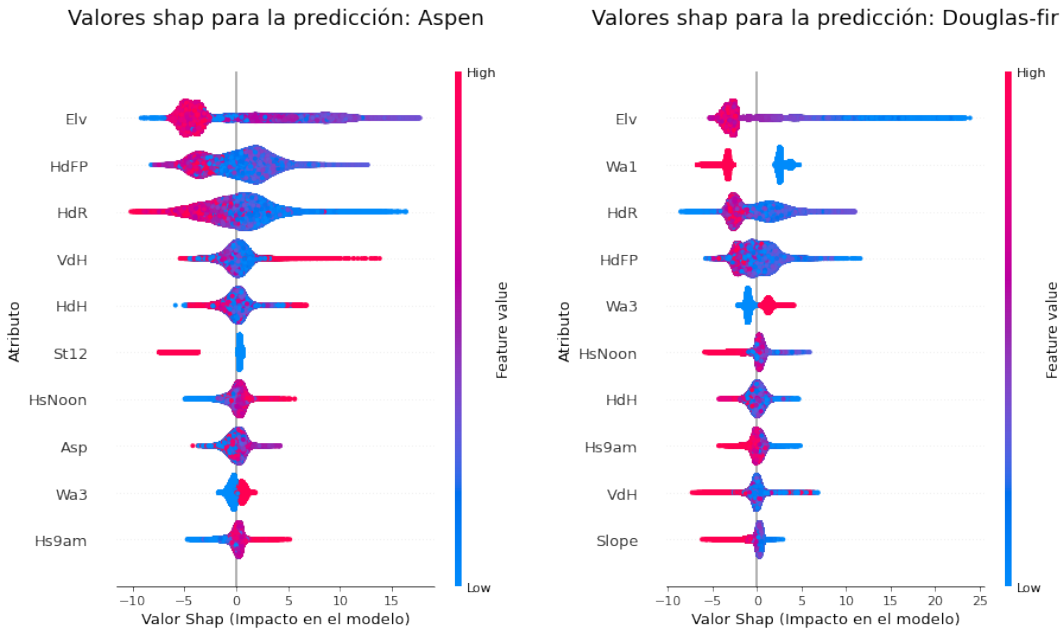

Figura 16: Beewarm plot para la predicción de las categorías Aspen y Douglas-fir

De las categorías ya analizadas se puede distinguir que la mayoría de los valores están centrados en el valor 0 de valores Shapley, no aportando en las predicciones finales, lo que indicaría que no todas las predicciones se calculan considerando la totalidad de los atributos, si la instancia está en el eje central, los valores Shapley son cercanos a 0, indicando su poco aporte a la hora de realizar la predicción. Además, la separación entre valores extremos es muy clara para la mayoría de los atributos al tener colas azules o rojas, pero no moradas, lo que indicaría que estos valores no existen para la categoría o están mimetizados en el eje central.

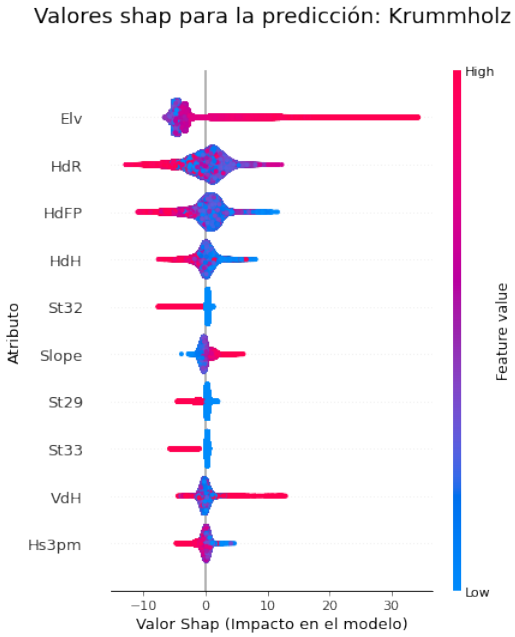

Figura 17: Beewarm plot para la predicción de la categoría Krummholz

Finalmente se puede apreciar que para la categoría Krummholz el atributo más importante es la *Elevación* (Elv) y casi no tomando en cuenta los otros atributos. A mayor elevación más probabilidad de que sea Krummholz, esperable para un árbol que solo crece en grandes alturas.

Análisis de predicciones específicas

El último análisis consistió en analizar las predicciones individualmente. Para realizar este análisis se seleccionaron cuatro instancias del dataset, dos Spruce/Fir, y dos Cottonwood/Willow, una predicha correctamente y la otra no. Cada instancia genera siete probabilidades, por lo cual se podría revisar cada una pero sería un análisis muy extenso, por esto, se analizará la dos predicciones que generaron una mayor probabilidad. Para realizar este análisis se utilizarán waterfall plots, donde se indica el aporte, positivo o negativo, de las instancias más importantes y la suma de las demás. Estos gráficos se pueden observar en las Figuras 18, 19, 20 y 21.

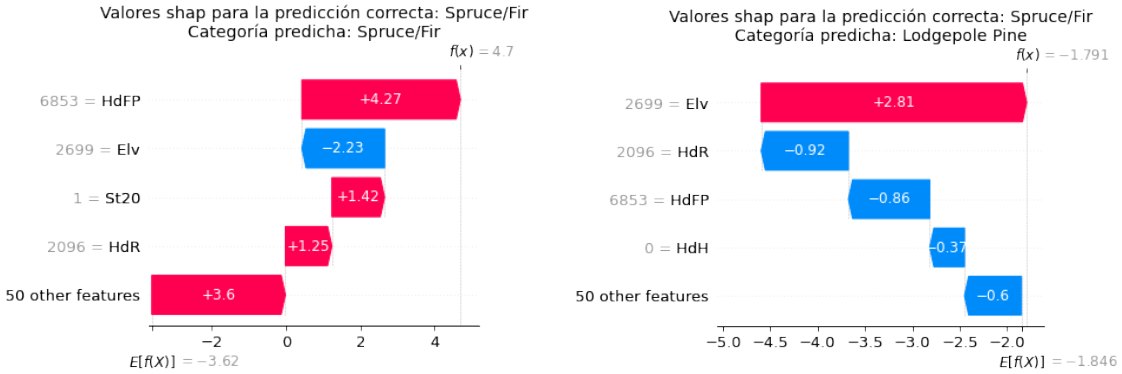

Figura 18: Waterfall plot para categoría Spruce/Fir de una instancia correctamente predicha. A la izquierda el gráfico relacionado a la probabilidad de ser la categoría Spruce/Fir y a la derecha Lodgepole Pine.

La primeras predicciones fueron realizadas sobre una instancia correctamente predicha de la categoría Spruce/Fir, se predijo con mayor probabilidad Spruce/Fir y la segunda categoría con mayor probabilidad es Lodgepole Pine, como se observa en la Figura 18 en los valores superiores como $f(x)$.

Para esta predicción la función de probabilidad fue de 4.7, comparado con la segunda categoría de -1.791. Si observamos, la mayoría de los atributos aportaron un valor positivo de valores Shapley, generando una predicción más segura de la categoría. Por otro lado, si vemos la siguiente categoría con mayor probabilidad, o Lodgepole Pine, la mayoría de los atributos sumaron valores negativos de valores Shapley, redujendo la probabilidad de que fuese esta categoría.

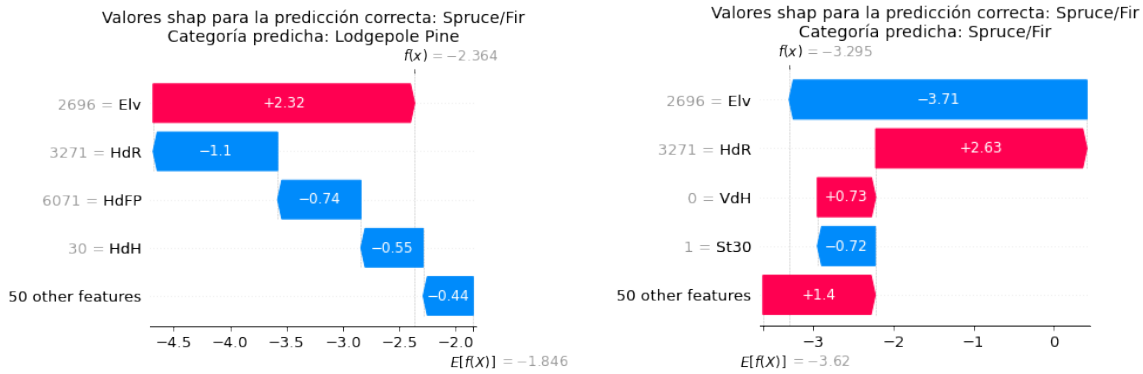

Figura 19: Waterfall plot para categoría Spruce/Fir de una instancia incorrectamente predicha como Lodgepole Pine. A la izquierda el gráfico relacionado a la probabilidad de ser la categoría Lodgepole Pine y a la derecha Spruce/Fir.

Al igual que la predicción anterior, esta categoría es Spruce/Fir, pero la predicción con mayor probabilidad fue Lodgepole Pine, lo que es incorrecto. Si nos fijamos en la probabilidad, los dos valores son muy cercanos, indicio de tener una separabilidad más compleja. Al analizar los atributos considerados al momento de la predicción, para la categoría Lodgepole Pine son muy similares a las de la Figura 18, mientras que para la categoría Spruce/Fir varían, tomando más importancia la *Elevación* (Elv) y menos importante la *Distancia horizontal a puntos de incendio* (HdFP), incluso no apareciendo entre el top 4 de atributos.

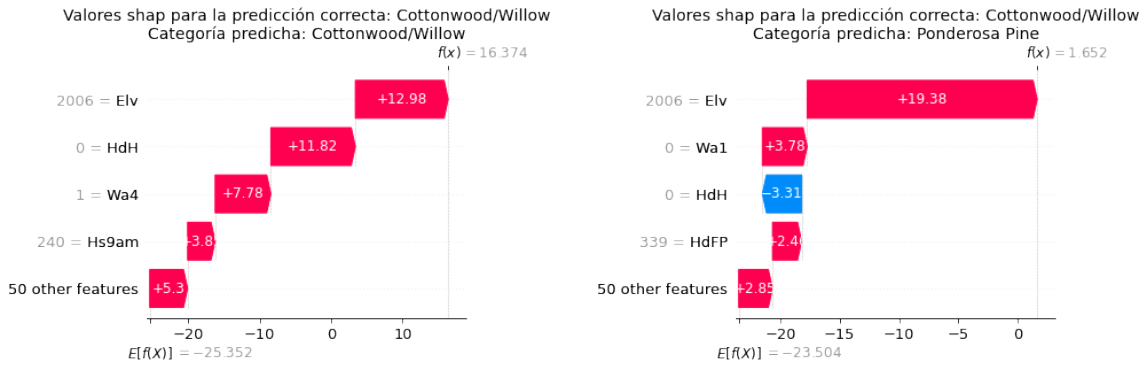

Figura 20: Waterfall plot para categoría Cottonwood/Willow de una instancia correctamente predicha. A la izquierda el gráfico relacionado a la probabilidad de ser la categoría Cottonwood/Willowr y a la derecha Ponderosa Pine.

Ahora, si se realiza el análisis de una categoría minoritaria como Cottonwood/Willow, obtenemos

las Figuras 20 y 21. Para la predicción correcta en la Figura 20, nos damos cuenta que la predicción fue muy segura comparado con la siguiente categoría con mayor probabilidad, 16.374 contra 1.652, además, las categorías en general aportaban valores positivos y altos de valores Shapley, comparado con la segunda categoría con mayor probabilidad que solo *Elevación* (Elv) aportó en gran cantidad.

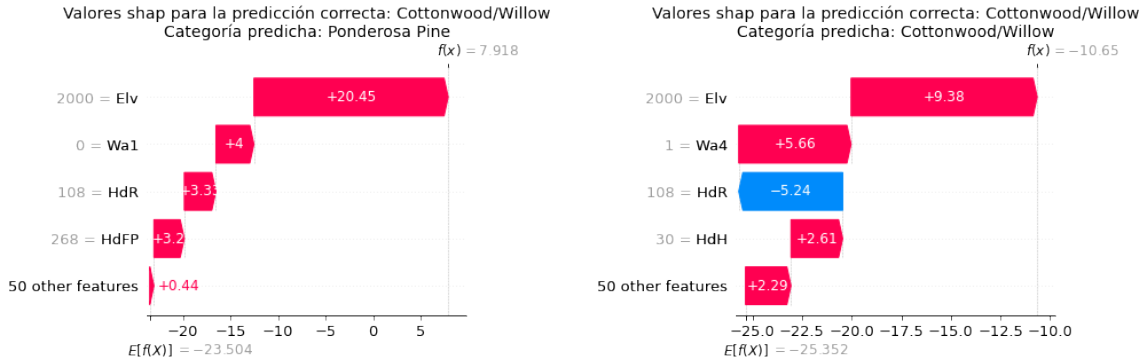

Figura 21: Waterfall plot para categoría Cottonwood/Willow de una instancia incorrectamente predicha como Lodgepole Pine. A la izquierda el gráfico relacionado a la probabilidad de ser la categoría Ponderosa Pine y a la derecha Cottonwood/Willow.

Finalmente, si consideramos la instancia que debería ser Cottonwood/Willow pero se predijo como Ponderosa Pine, nos damos cuenta de que sería muy complejo que el modelo predijera la categoría correcta, debido a que solo alcanzo -10.65 como probabilidad para la categoría correcta, mientras que para la categoría incorrecta obtuvo 7.918 siendo, al igual que el análisis anterior, la elevación un factor muy importante.

6.5. Discusión de los resultados

Al considerar los resultados anteriores, donde se alcanzaron métricas de accuracy promedio de 0.966 y F1 promedio de 0.943 para el modelo de LightGBM agregando datos sintéticos, se considera pertinente utilizar un modelo de LightGBM para realizar la predicciones de Tipo de Cobertura utilizando datos cartográficos. Además, al analizar la interpretabilidad del modelo, se estima pertinente utilizar las variables cartográficas como variables adicionales a otro tipo de metodologías de clasificación, estas variables son relevantes para estimar la especie de árbol que crecerá en determinada ubicación, siendo por ejemplo la Elevación (Elv) una variable muy importante para realizar las predicciones, variable que no es incluida por ejemplo, en modelos utilizando imágenes satelitales. La

comparación con otro tipo de metodologías y trabajos, además de señalar las ventajas y desventajas del modelo generado se discutirán a continuación. Finalmente se establecerán los pasos a seguir.

6.5.1. Comparativa de modelos

Actualmente son pocos los estudios que se basan en datos cartográficos para realizar este tipo de análisis, predominando en el último tiempo el uso de imágenes aéreas para realizar la clasificación de especies, principalmente debido a la facilidad de adquisición de datos.

Lindberg [17] por ejemplo, utiliza la tecnología Lidar, que genera imágenes de alta resolución utilizando un laser desde el cielo y capturando la luz reflejada. Al igual que este informe, se buscó la clasificación del árbol predominante, pero en un área de $15mx15m$ de cuatro categorías, *Pinus sylvestris, Picea abies*, especies de hoja caduca y terreno no reconocible. Para resolver el problema de clasificación se utilizó la metodología Mini raster cell. Con esta metología se logró un accuracy promedio de 75 %. Aunque este modelo cuente con la ventaja de clasificar directamente sobre el espacio las distintas categorías, no logra un gran rendimiento con la metodología utilizada.

Otro estudio es el de Deur [6], el cual combina técnicas de machine learning junto con imágenes satelitales de alta resolución del WorldView-3. En este estudio clasifican los pixeles como unidad base de medida en las predicciones, para llevar a cabo la clasificación de las especies *Quercus robur , Carpinus betulus y Alnus glutinosa*, pertenecientes a Croacia. Para el problema de clasificación se utilizaron técnicas de random forest y suport vector machine. Utilizando esta metodología se logró un acuraccy promedio de 85 % utilizando el modelo de random forest. Utilizando imágenes y modelos como random forest se lograron mejores resultado, pero aun bajo a lo logrado en este informe. Es importante mencionar que al utilizar imágenes satelitales, se hace necesario realizar técnicas de preprocesamiento, incluido la reducción de dimensionalidad y obtención de características, ya que solo una imagen 256x256 pixeles implicaría 65.536 variables para un modelo de random forest si es que solo se utilizara un espectro de la imagen, si ya se adicionan filtros y otros espectros esta información aumenta radicalmente.

Aunque con la información del set de datos se logró generar un modelo predictivo, no se puede generar una segmentación del terreno según la especie, esto debido a que no se cuenta con información espacial de las distintas instancias. Aun así, si se contara con el modelo DEM se podría realizar al igual que en los estudios anteriormente mencionados.

Modelo	**Trabajo** (F1-Acc)	**Gupta** (F1)	**Sjöqvist** (Acc)
LightGBM	0.943 - 0.966	-	-
R. Logistica	-	0.674	-
K-nn	-	0.854	-
LDA	-	0.641	-
Naive Bayes	-	0.574	0.661
Random Forest	0.912 - 0.946	0.732	0.950
Gradient Boosting trees	-	0.870	-
Neuronal network	-	0.841	-
Support vector machine	-	0.825	0.890

Tabla 12: Tabla comparativa de resultados. Mientras que Gruta solo reporta la métrica F1, precision y recall, Sjöqvist solo presenta el accuracy obtenido.

Si se analizan estudios similares utilizando el mismo set de datos se encuentran los estudios de Gupta [10] y Sjöqvist [30], el resumen de los resultados comparativos se pueden apreciar en la Tabla 12. El primero utiliza técnicas de preprocesado al escalar las variables y realiza reducción de dimensionalidad. Además, generó modelos de naive bajes, regresión logística, linear discriminant analysis, weighted knn, knn, neuronal networks, support vector machine, gradient boosting trees y random forest. Por otro lado Sjöqvist obtiene los mejores resultados utilizando todas las variables y utiliza modelos de support vector machine, naive bayes y random forest, llegando a métricas de accuracy del 95 % utilizando random forest, pero con métricas de accuracy de 0.779 y 0.770 para las categorías Cottonwood/Willow y Aspen. Comparativamente los dos estudios obtienen métricas menores a las logradas en este informe y aunque Sjöqvist lograra altas métricas de accuracy para el modelo de random forest, es en prejuicio de las categorías minoritarias.

6.5.2. Alcances del modelo

Como se explicó en la Sección 3: Descripción de datos, la mayor parte de los datos se obtuvo de manera remota, a excepción del etiquetado de datos que requiere un sobrevuelo en el terreno con alguien etiquetando la clase de árbol además de guardando la geolocalización. La tecnología actual nos permite generar modelos de terreno basado en la elevación de estos, incluso sin la necesidad de estar en el terreno [22]. Esto da una ventaja clara para el objetivo de este informe, realizar el estudio y obtención de información disminuyendo lo más posible la intervención humana. Además, al realizar analizar la interpretabilidad del modelo, observamos la importancia de estas variables en el crecimiento de las distintas especies, incluso abre la posibilidad de utilizar una menor cantidad de variables en el caso de querer realizar un modelo DEM: Solo con información de altura y distancia

a cuerpos de agua, carreteras y puntos de incendio se puede obtener la mayor parte del poder predictivo.

Esta misma ventaja lleva a la desventaja principal, no todos los territorios tienen un estudio detallado del terreno o las tecnologías necesarias para generar un modelo DEM. Sin este análisis basado en la altura del terreno es muy complejo obtener variables como la inclinación del terreno o sombra a las distintas horas del día, aunque aún se podría obtener distancias a cuerpos de agua, incendio y carreteras utilizando imágenes satelitales, pero se perdería la principal variable predictora: la Elevación.

Considerando que los incendios han ido aumentando con los años, tener un modelo que considere esta variable es muy importante. Dentro de nuestro análisis de interpretabilidad, se encontró que la tercera variable que más poder predictivo tiene es la distancia horizontal a puntos de incendios. Para algunas especies, como Aspen, es la segunda variable que mayor relevancia tiene como se puede observar en la Figura 16, mostrando que puede existir en zonas cercanas a incendios, como se describió en la Sección 3: *Descripción de datos*, al tener la posibilidad de repoblar rápidamente esos terrenos a pesar de quemarse [24]. Al igual que el análisis de incendio esto se puede aplicar a los cambios en los cuerpos de agua. Este tipo de modelo nos ayudaría a estimar la probabilidad de que exista una especie en zonas donde ocurran cambios debido a la sequía o cambios en el curso de agua, información muy valiosa en términos de manejo y conservación del ecosistema.

6.5.3. Pasos futuros

Con los resultados obtenidos en el trabajo, además de la interpretabilidad del modelo y los resultados de otros métdos, se estima pertinente utilizar las variables acá utilizadas para otras metodologías. Considerando que la variable altura puede ser compleja de obtener si no se tiene un modelo DEM, otras variables como distancias a cuerpos de agua, carreteras e incendios son variables de alta importancia para las predicciones. Aunque estas variables se puedan encontrar en las imágenes satelitales, los modelos utilizados para su análisis, como modelos basados en árboles de decisión, no generan relaciones espaciales entre los distintos pixeles, por lo cual es viable incluirlas en forma de las variables ya mencionadas.

Por otro lado, ya con el modelo entrenado, se podría sumar información al modelo DEM y generar un mapa de probabilidades de crecimiento de cada especie y así generar un censo forestal y determinación de la biomasa disponible. Además, considerando que siempre existirán cambios en la

geografía del terreno, introducción de nuevas especies o enfermedades, se debe medir la adaptabilidad del modelo frente a estos cambios. Finalmente se debe analizar la posibilidad de aplicar este tipo de estudio en otras reservas forestales.

7. Conclusión

A partir de la realización de este trabajo podemos concluir que usando datos cartográficos y algoritmos basados en árboles de decisión podemos predecir satisfactoriamente la cobertura de árboles. Se logró obtener un modelo basado en Random Forest que generó métricas de 0.946 para accuracy, un 0.912 de F1 y un 0.997 de AUC, mientras que para el mejor modelo de LightGBM se obtuvo un 0.969 de accuracy, un 0.945 de F1 y un 0.999 de AUC.

Para los modelos basados en LightGBM observamos que se genera sobre ajuste al elegir valores altos en los hiperparámetros, 6000 estimadores, 500 hojas y 50 en profundidad. Sin embargo, concluimos que este no perjudica negativamente en el resultado final.

Al analizar la métrica F1 para las distintas categorías utilizando el modelo de LightGBM, tanto Cottonwood/Willoww y Aspen obtuvieron bajos valores, cercanos a 0.9, mientras que las categorías mayoritarias, Spruce/Fir y Lodgepole Pine obtuvieron 0.969 y 0.974 respectivamente.

Al utilizar la función de pérdida Focal Loss en el modelo LightGBM no se logra aumentar las métricas de evaluiación, pero se obtuvo una ventaja al utilizar una menor cantidad de estimadores, aumentando la métrica F1 entre los 100 y 300 estimadores.

Generar información sintética de las dos categorías minoritarias mejoró la métrica promedio de F1 utilizando LightGBM como modelo, de 0.940 a 0.943, al pasar la categoría Cottonwood/Willow de 0.882 a 0.902 y Aspen de 0.896 a 0.899.

Al realizar el análisis de interpretabilidad, el atributo más importante para la toma de decisiones fue la *Elevación* seguido por la *Distancia horizontal a carreteras* y *Distancia horizontal a puntos de incendios*. Por otro las variables categóricas que más aportaron en la predicción fueron *Área silvestre 1 y 3*.

Finalmente, se estiman grandes ventajas al utilizar estas variables cartográficas, en especial para el manejo y estudio de poblaciones de árboles, debido a que se consideran variables como cercanía a incendios y cuerpos de agua. Aunque se presenta la principal desventaja que es obtener un modelo DEM para la obtención de las variables alturas, inclinación y sombra a través del día.

Comparativamente con otras metodologías utilizar datos cartográficos lleva a un aumento en las métricas de clasificación y da la posibilidad de generar la segmentación del terreno si es que se cuenta con el modelo DEM; por otro lado, al comparar con trabajos similares, se lograron mejores métricas de clasificación y un aumento en las métricas de las variables minoritarias Cottonwood/Willow y Aspen.

Referencias

[1] I. Ashrapov. Tabular gans for uneven distribution. *CoRR*, abs/2010.00638, 2020. URL `https://arxiv.org/abs/2010.00638`.

[2] J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, Dec. 1999. ISSN 01681699. doi: 10.1016/S0168-1699(99)00046-0. URL `https://linkinghub.elsevier.com/retrieve/pii/S0168169999000460`.

[3] L. Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. URL `http://www.jstor.org/stable/2676681`.

[4] L. Carniato. Multi-Class classification using Focal Loss and LightGBM. `https://bit.ly/3nT1E7G`, 2021. [Online; accessed 2-Nov-2021].

[5] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. *Ecology*, 88(11):2783–2792, Nov. 2007. ISSN 0012-9658. doi: 10.1890/07-0539.1. URL `http://doi.wiley.com/10.1890/07-0539.1`.

[6] M. Deur, M. Gasparovic, and I. Balenović. Tree species classification in mixed deciduous forests using very high spatial resolution satellite imagery and machine learning methods. *Remote Sensing*, 12:3926, 11 2020. doi: 10.3390/rs12233926.

[7] J. S. Evans, M. A. Murphy, Z. A. Holden, and S. A. Cushman. Modeling Species Distribution and Change Using Random Forest. In C. A. Drew, Y. F. Wiersma, and F. Huettmann, editors, *Predictive Species and Habitat Modeling in Landscape Ecology*, pages 139–159. Springer New York, New York, NY, 2011. ISBN 978-1-4419-7389-4 978-1-4419-7390-0. doi: 10.1007/978-1-4419-7390-0_8. URL `http://link.springer.com/10.1007/978-1-4419-7390-0_8`.

[8] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam, A. H. Vo, H. T. Dang, B. T. Nguyen, and V.-H. Pham. Vietnamese Herbal Plant Recognition Using Deep Convolutional Features. *International Journal of Machine Learning and Computing*, 9(3):363–367, June 2019. ISSN 20103700. doi: 10.18178/ijmlc.2019.9.3.811. URL `http://www.ijmlc.org/index.php?m=content&c=index&a=show&catid=85&id=937`.

[9] K. v. Gadow, C. Y. Zhang, C. Wehenkel, A. Pommerening, J. Corral-Rivas, M. Korol, S. Myklush, G. Y. Hui, A. Kiviste, and X. H. Zhao. Forest Structure and Diversity. In T. Pukkala and K. von Gadow, editors, *Continuous Cover Forestry*, volume 23, pages 29–83. Springer Netherlands, Dordrecht, 2012. ISBN 978-94-007-2201-9 978-94-007-2202-6. doi: 10.1007/978-94-007-2202-6_2. URL http://link.springer.com/10.1007/978-94-007-2202-6_2. Series Title: Managing Forest Ecosystems.

[10] A. Gupta, J. R, H. Sahwney, and A. Zalani. Classifying forest categories using cartographic variables, 04 2015.

[11] R. J. Hobbs, S. Arico, J. Aronson, J. S. Baron, P. Bridgewater, V. A. Cramer, P. R. Epstein, J. J. Ewel, C. A. Klink, A. E. Lugo, D. Norton, D. Ojima, D. M. Richardson, E. W. Sanderson, F. Valladares, M. Vilà, R. Zamora, and M. Zobel. Novel ecosystems: theoretical and management aspects of the new ecological world order: Novel ecosystems. *Global Ecology and Biogeography*, 15(1):1–7, Jan. 2006. ISSN 1466822X. doi: 10.1111/j.1466-822X.2006.00212.x. URL http://doi.wiley.com/10.1111/j.1466-822X.2006.00212.x.

[12] P. Holgate. The Angle-Count Method. *Biometrika*, 54(3/4):615, Dec. 1967. ISSN 00063444. doi: 10.2307/2335052. URL https://www.jstor.org/stable/2335052?origin=crossref.

[13] U. Irvine. Covertype Data Set Repository. https://archive.ics.uci.edu/ml/datasets/Covertype, ny. [Online; accessed 12-May-2021].

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[15] J. Lawler, J. Watson, and E. Game. Conservation in the face of climate change: Recent developments. *F1000Research*, 4, 10 2015. doi: 10.12688/f1000research.6490.1.

[16] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. URL http://arxiv.org/abs/1708.02002.

[17] E. Lindberg, J. Holmgren, and H. Olsson. Classification of tree species classes in a hemiboreal forest from multispectral airborne laser scanning data using a mini raster cell method.

International Journal of Applied Earth Observation and Geoinformation, 100:102334, 08 2021. doi: 10.1016/j.jag.2021.102334.

[18] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*, Mar. 2019. URL `http://arxiv.org/abs/1802.03888`. arXiv: 1802.03888.

[19] C. Messier, K. Puettmann, E. Filotas, and D. Coates. Dealing with non-linearity and uncertainty in forest management. *Current Forestry Reports*, 2, 06 2016. doi: 10.1007/s40725-016-0036-x.

[20] C. Molnar. *Interpretable Machine Learning*. 2019. URL `https://christophm.github.io/interpretable-ml-book/`.

[21] L. Naidoo, M. Cho, R. Mathieu, and G. Asner. Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 69:167–179, Apr. 2012. ISSN 09242716. doi: 10.1016/j.isprsjprs.2012.03.005. URL `https://linkinghub.elsevier.com/retrieve/pii/S0924271612000597`.

[22] R. Peckham and G. Jordan. *Digital Terrain Modelling: Development and Applications in a Policy Support Environment*. 01 2007. ISBN 978-3-540-36730-7. doi: 10.1007/978-3-540-36731-4.

[23] B. Quinto. *Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More*. Apress, Berkeley, CA, 2020. ISBN 978-1-4842-5668-8 978-1-4842-5669-5. doi: 10.1007/978-1-4842-5669-5. URL `http://link.springer.com/10.1007/978-1-4842-5669-5`.

[24] W. Ripple and L. E.J. The role of postfire coarse woody debris in aspen regeneration. *Western Journal of Applied Forestry*, 16:61–64, 04 2001. doi: 10.1093/wjaf/16.2.61.

[25] V. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. Atkinson, and C. Jeganathan. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, 121:93–107, June 2012. ISSN 00344257. doi: 10.1016/j.rse.2011.12.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0034425711004408`.

[26] V. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, Jan. 2012. ISSN 09242716. doi: 10.1016/j.isprsjprs.2011.11.002. URL `https://linkinghub.elsevier.com/retrieve/pii/S0924271611001304`.

[27] T. Ryan, T. Philippi, Y. Leiden, M. Dorcas, T. Wigley, and J. Gibbons. Monitoring herpetofauna in a managed forest landscape: Effects of habitat types and census techniques. *Forest Ecology and Management*, 167:83–90, 08 2002. doi: 10.1016/S0378-1127(01)00692-2.

[28] E. W. Sanderson, M. Jaiteh, M. A. Levy, K. H. Redford, A. V. Wannebo, and G. Woolmer. The Human Footprint and the Last of the Wild. *BioScience*, 52(10):891, 2002. ISSN 0006-3568. doi: 10.1641/0006-3568(2002)052[0891:THFATL]2.0.CO;2. URL `https://academic.oup.com/bioscience/article/52/10/891-904/354831`.

[29] H. Shi. Best-first Decision Tree Learning. *The University of Waikato, New Zealand*, 2007.

[30] H. Sjöqvist, M. Längkvist, and F. Javed. An analysis of fast learning methods for classifying forest cover types. *Applied Artificial Intelligence*, 34:1–19, 06 2020. doi: 10.1080/08839514.2020.1771523.

[31] L. Xu and K. Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *CoRR*, abs/1811.11264, 2018. URL `http://arxiv.org/abs/1811.11264`.

[32] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional GAN. *CoRR*, abs/1907.00503, 2019. URL `http://arxiv.org/abs/1907.00503`.