Pontificia Universidad Católica del Perú
Escuela de Posgrado

Análisis de la expansión vial en la Amazonía peruana y su impacto en el cambio climático

Tesis para optar por el grado de Magíster en Ingeniería Civil

Autor: *Gustavo Martín Larrea Gallegos*
Asesor: *Ian Vázquez Rowe*

14 de marzo de 2019

Resumen

La selva amazónica alberga alrededor del 60 % del bosque tropical del mundo y es un elemento fundamental en términos de biodiversidad, clima y secuestro de carbono del planeta. En este contexto, el Gobierno Peruano ratificó el año 2015 sus intenciones por reducir sus emisiones de Gases de Efecto Invernadero en un 20 % con respecto a un escenario habitual mediante reducciones en el sector de cambio de uso de suelos. La construcción de carreteras es una de las principales actividades asociadas a este sector e importante generador de deforestación. En los últimos años el Perú se ha atravesado un considerable incremento de su infraestructura vial, y se espera que esta expansión siga en aumento. En este sentido, la presente investigación tiene como principal objetivo contribuir al entendimiento de los efectos que la expansión vial puede generar en el cambio de uso de suelos, y posteriormente en el cambio climático en toda la Amazonía peruana. Para ello, se construyeron diferentes modelos de aprendizaje automático (random forest, regresión logística y redes neuronales) para predecir la potencial deforestación en un periodo de 15 años. Se utilizó información georreferenciada y herramientas computacionales del estado del arte. Los resultados indican que, evaluando solo un proyecto vial en particular, se podrían generar 73.2 Mt de CO_2eq. Este valor supera en demasía a las 60 Mt de CO_2eq estimadas por el Gobierno Peruano como meta de reducción. Por lo que se concluye que las estimaciones realizadas por el estado subestiman los efectos de la construcción de carreteras. Finalmente, el marco metodológico presentado es novedoso y útil para construir e implementar modelos de predicción de deforestación para el cálculo de emisiones de GEI y puede ser implementado para analizar otros casos de estudio.

A mis padres y hermanos...

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. La deforestación como fenómeno antrópico

La selva amazónica alberga alrededor del 60 % del bosque tropical del mundo y es un elemento fundamental en términos de biodiversidad, clima y secuestro de carbono del planeta (Houghton et al., 2000). La relevancia ambiental de este escenario hace contraste con la rampante disminución de bosque primario y la alarmante degradación de cobertura Amazónica (Laurance et al., 2002). En el caso de Brasil, nación que contiene el 70 % de la selva amazónica, desde el año 1992 se han deforestado alrededor de 2 millones de hectáreas anuales. Este fenómeno se ha incrementado sobre todo en países tropicales en vías de desarrollo. La pérdida de estos espacios altamente sensibles es parte de una dinámica muy compleja que está relacionada con la expansión urbana y agrícola, la pérdida de biodiversidad, minería aluvial, tala selectiva de bosques y la proliferación de enfermedades endémicas (Hall and Goodman, 1991). De hecho, la deforestación asociada con la minería aluvial genera importantes cambios ecológicos (i.e., destrucción de sistemas acuáticos y creación pozos de aguas estancadas, entre otros) que incrementan los nichos de cultivo de los mosquitos portadores del vector de la malaria (e.g., mosquito Anopheles) (Silbergeld et al., 2002). La Amazonía representa más del 30 % del área total del país y se ha perdido alrededor de 2 millones de hectáreas de bosque Amazónico en los últimos 15 años, lo que representa alrededor del 2 % del territorio Amazónico (WRI, 2016). Como muestra la Figura 1.1, esta pérdida de cobertura arbórea no ha sido repentina; por el contrario, es el resultado de una creciente tendencia en los últimos añós. Aun así, el Estado Peruano asumió el compromiso de reducir en 100 % la deforestación de bosque primario para el año 2030 (MINAM (Ministerio del Ambiente), 2016)

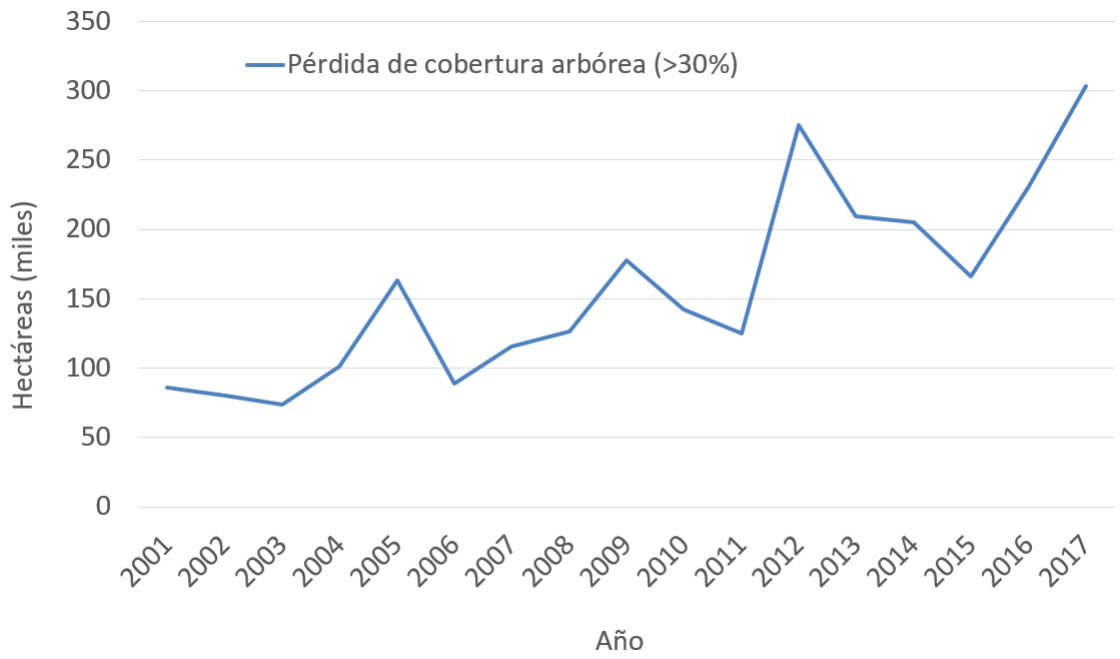

Figura 1.1: Variación anual de la pérdida de superficie arbórea en el Perú en los años 2001-2017. Fuente: Global Forest Watch (2019)

En el año 2016, el Estado Peruano ratificó las Contribuciones Nacionalmente Determinadas (NDC, por sus siglas en inglés) en la Conferencia de las Partes para el Cambio Climático (COP21) (UCFCCC, 2015). Estas detallan las intenciones del Gobierno por reducir sus emisiones de Gases de Efecto Invernadero (GEI) en un 20 % con respecto a un escenario habitual mediante reducciones en el sector de Uso de Suelos, Cambio en el Uso de Suelos y Silvicultura (USCUSS) (MINAM (Ministerio del Ambiente), 2016). En este sentido, el manejo de bosques y las actividades asociadas a estos son relevantes, particularmente, en los territorios de la cuenca Amazónica. Más de la mitad del territorio nacional está cubierto por bosque Amazónico. Sin embargo, pese a esta evidente abundancia de superficie forestal, la contribución económica que deriva de este sector al valor de producción bruto nacional es de 1 a 3 por ciento, considerando productos forestales maderables y no maderables (Held et al., 2015). Este aporte poco significativo del sector forestal a la economía nacional puede dar un alcance inicial de las razones por las cuales no ha existido una postura robusta hacia la protección de los servicios forestales y el manejo sostenible de sus recursos. De hecho, el principal recurso de los bosques es la madera. En el Perú, en el año 2012, se extrajeron 7.9 millones de metros cúbicos de este recurso. Alrededor del 89 % de la madera fue destinada como leña; 10 %, al sector industrial y comercial; y 1 % utilizado como carbón (MINAGRI, 2014). Estas estadísticas reflejan una mínima participación de los recursos forestales maderables en los sistemas económicos, principalmente, debido al poco valor agregado de estos productos a lo largo de la cadena de valor. En efecto, hasta el año 2013, solo el 11 % de la producción maderable pasaba por un proceso de transformación. Esta información se refleja en su poco aporte al PBI (1.1 %) y su poca contribución en la generación de empleo (0.3 %) (MINAGRI, 2014). Sin embargo, estas cifras, aunque pequeñas, van en aumento debido al crecimiento de la población y al incremento de la demanda interna, sobre todo en aquellos sectores industriales que consumen la madera como insumo o materia prima. Como puede resultar evidente, la principal amenaza para el sector forestal es la pérdida de su principal recurso: los bosques. En este sentido, la deforestación es un fenómeno que está relacionado con las distintas actividades antrópicas en las que se busca obtener beneficios económicos del uso de los recursos forestales. Diferentes manifestaciones de esta dinámica económica han podido observarse en el último siglo. De hecho, desde la primera intrusión del sistema mercantilista extractivo en la Amazonía desatado por la fiebre del caucho, la selva Amazónica ha sido escenario de conflictos que giraron en torno a sus recursos (Reyna, 1942).

1.2. La expansión vial en el Perú

La infraestructura es un elemento fundamental en el desarrollo económico de un país y es imprescindible por las sociedades modernas. Más específicamente, se ha demostrado que la infraestructura vial tiene un importante efecto sobre el crecimiento de la economía de distintos países debido a que incrementa la productividad (Aschauer, 1989; Canning and Fay, 1993). En el caso del Perú, la influencia de las carreteras en la economía no es muy diferente a lo observado en distintos países. En efecto, Vásquez y Bendezú [2008] analizaron la influencia de la inversión en infraestructura vial sobre el crecimiento económico del Perú en el periodo 1970 - 2003. Los autores determinaron que la infraestructura vial disminuye el tiempo de adaptación de los precios ante algún determinado shock, posibilita la existencia de mercados eficientes y eleva los éstandares de calidad de vida [Banco Mundial, 1994; Vásquez y Bendezú, 2008]. Sin embargo, la relación entre el incremento de la infraestructura vial y su efecto sobre la economía no es lineal. Esto se debe a que a medida que la inversión en carreteras se incrementa, su efecto marginal en la economía disminuye [Vásquez y Bendezú, 2008; Aschauer, 1989]. Esta no-linearidad permite inferir que los beneficios económicos de invertir en carreteras serán más significativos en países en vías de desarrollo que en países desarrollados.

En el Perú se ha observado un considerable incremento de la infraestructura vial durante los últimos 10 años. De hecho, como muestra la figura 1.2, a partir del año 2007 se ha presenciado un importante aumento de la cantidad de kilómetros construidos que forman parte de la Red Vial Nacional (RVN). Este considerable desarrollo de infraestructura puede estar ligado al crecimiento económico por el que atravesó el Perú durante la primera década del último milenio. Sin embargo, aunque el crecimiento de RVN es notable, se considera que existe aún una importante brecha de infraestructura que requiere ser cubierta en referencia a la construcción de nuevas vías y al mantenimiento de las vías existentes (Coronado, 203; **?**). En efecto, como puede observarse en la figura 1.3, aunque la longitud de vías aumentó, hasta el año 2016, existe aún una considerable deficiencia de infraestructura vial, sobretodo en la zona de la Amazonía peruana. En este sentido, se puede esperar que en los próximos años las regiones de la costa y sierra inviertan en mantenimiento y mejoramiento de vías; mientras que las regiones de la selva lo hagan en construcción de nuevas carreteras.

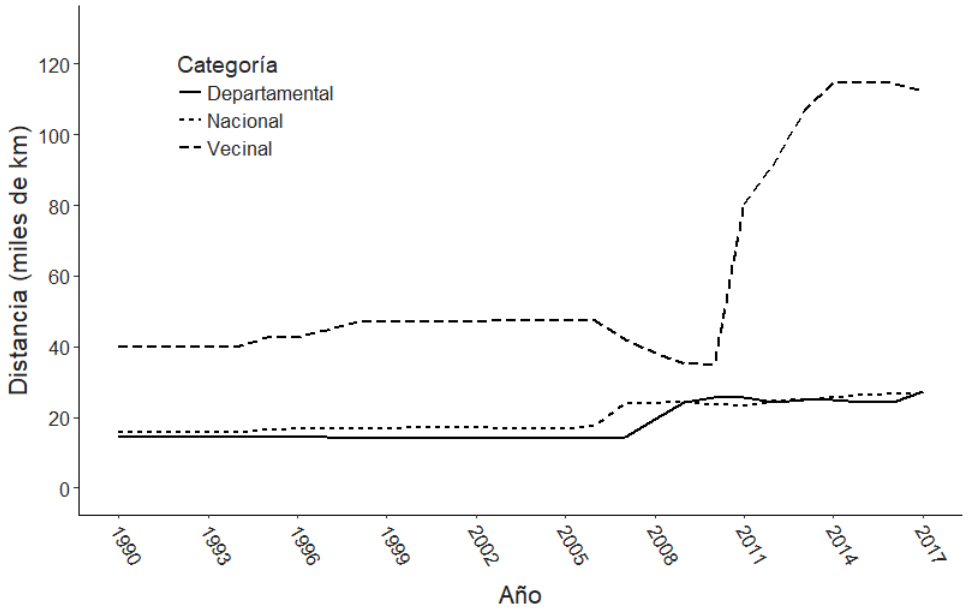

Figura 1.2: Desarrollo de la Red Vial Nacional entre los años 1990 y 2017 por cada tipo de categoría de vía. Fuente: INEI (2019)

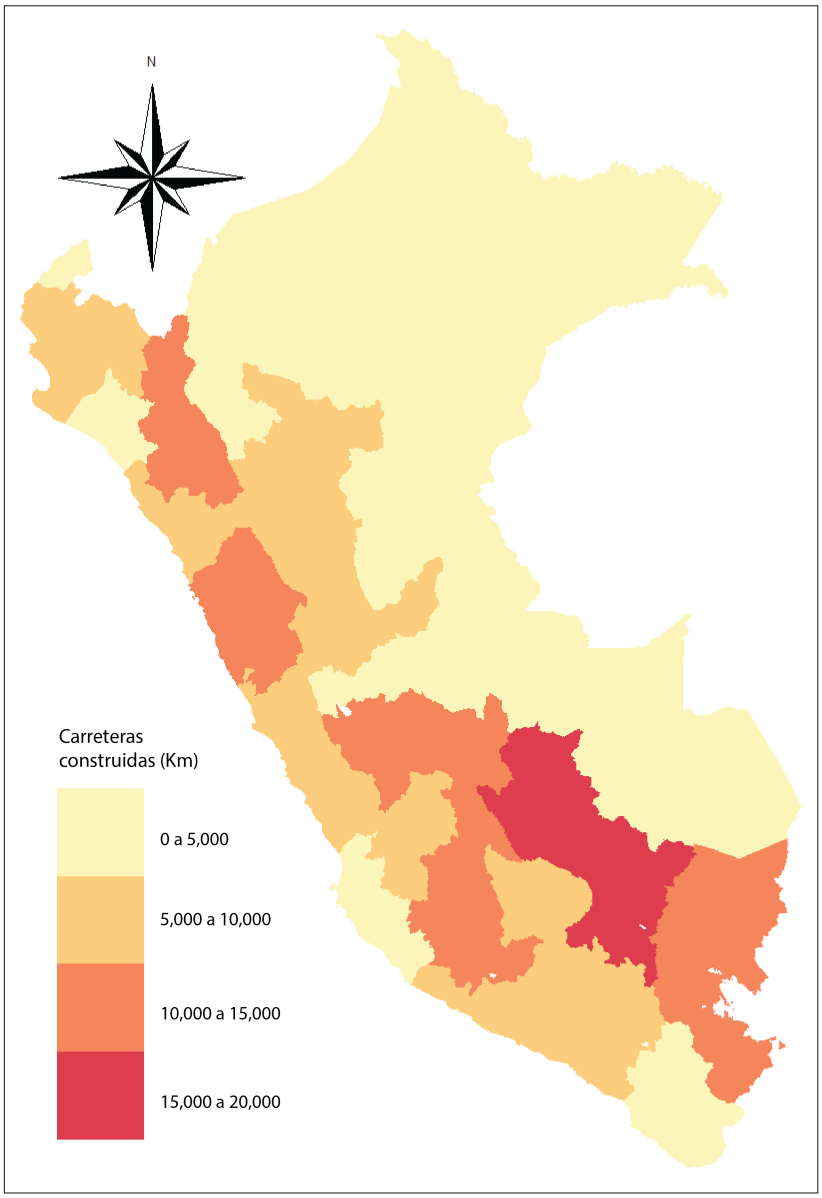

Figura 1.3: Mapa de carreteras construidas por departamento hasta el año 2016. Fuente: INEI (2019)

La expansión vial y el incremento de la accesibilidad no solo están asociados a los efectos positivos del crecimiento económico, sino que existen también una serie de efectos socio-ambientales vinculados a esta actividad. Por su naturaleza intrusiva, toda nueva carretera incrementa considerablemente el riesgo de deforestación. Esto último se debe a que la pérdida de bosque suele ser espacialmente rampante (Boakes et al., 2010) y a que las carreteras generan, a su vez, la aparición de más carreteras de menor jerarquía (Laurance et al., 2002, 2015). Como se puede esperar, estos efectos no dependen solo de la construcción de nuevas vías, sino también de las modificaciones en las características de las carreteras existentes. En el caso de la Amazonía, las carreteras no-pavimentadas suelen ser inutilizadas durante los periodos de lluvia, por lo que el solo hecho de pavimentarlas permite incrementar su tiempo de uso y los efectos asociados a este (Laurance et al., 2009, 2015). Un claro ejemplo de este este fenómeno se aprecia en el caso de la carretera IIRSA SUR, que conecta los países de Perú y Brasil. El asfaltado de esta vía condujo a que se incremente el flujo vehicular durante todo el año y a que, al mismo tiempo, se de una considerable reducción en los tiempos de viaje. Debido a estos efectos, la deforestación en los alrededores se incrementó, así como la expansión urbana, agrícola y la minería aluvial (Asner et al., 2013; Laurance et al., 2015; Delgado, 2008).

En este sentido, medidas como limitar directamente la construcción de nuevas carreteras, gestionar la extracción maderera, mejorar las herramientas de medición de impactos ambientales, mejorar el diseño vial, entre otras, han sido propuestas tanto por autores de la literatura (Laurance et al., 2009) como por autoridades competentes (MINAM (Ministerio del Ambiente), 2016). Desde un punto de vista taxonómico, Laurance et al. (2009) clasifica estas diversas estrategias de mitigación de dos formas. La primera clasificación engloba a los esfuerzos a escala local enfocados en reducir los impactos de las nuevas carreteras y de las ya existentes. La segunda clasificación comprende a todos los esfuerzos a escala regional que buscan limitar la expansión e intrusión de las carreteras sobre áreas ecológicamente sensibles. Como puede verse, la problemática generada por la expansión vial puede ser abordada de multiples maneras. Sin embargo, en esta investigación se buscó generar herramientas que contribuyan a las medidas encasilladas en la segunda clasificación, es decir, a aquellas de escala regional. En específico, el presente manuscrito desarrollará el proceso seguido en el diseño y elaboración de una herramienta que permita mejorar las actuales estrategias de diseño vial.

1.3. Objetivos y justificación

Esta investigación tiene como principal objetivo contribuir al entendimiento de los efectos que la expansión vial puede generar en el cambio de uso de suelos, y posteriormente en el cambio climático. La secuencia lógica que motiva el planteamiento de este objetivo considera a la acción de generar accesibilidad (i.e., construir una carretera) como variable que incrementa el riesgo de cambio de uso de suelo (e.g, deforestación). Aunque el enfoque de cambio de uso de suelos pued e ser amplio y complejo en su definición, esta investigación se centrará principalmente en la deforestación. En este sentido, debido a que la zona de estudio es la Amazonía, toda mención sobre el cambio de uso de suelos en este documento será una referencia del cambio de una zona forestal a otra zona de cualquier tipo. La motivación en la selección de la Amazonía como zona de estudio y la justificación del proyecto deriva de tres principales argumentos:

Argumento 1: La selva Amazónica alberga zonas con las reservas más grandes de biodiversidad y de carbono del mundo (Asner, 2014)

Argumento 2: El Estado Peruano asumió el compromiso de reducir sus emisiones del sector de Uso de Suelos y Silvicultura en 60.57 Mt CO_2eq al 2030 (MINAM (Ministerio del Ambiente), 2016).

Argumento 3: Muy poca superficie de la selva Amazónica peruana que se encuentra conectada por vía terrestre y existe una serie de proyectos viales de gran embergadura en planificación (MTC, 2015).

La combinación de estos tres argumentos conlleva a considerar justificada la necesidad de estudiar el riesgo de deforestación que estos proyectos viales pueden incentivar. Además, de contar con un modelo, se podría simular los efectos de distintas alternativas de diseño, lo cual permitiría tomar mejores decisiones respecto al diseño de estas vías o anticipar planes de mitigación. Respecto a la estrategia metodológica adoptada en este estudio (ver Capítulo 3) existen diferentes enfoques para abordar los problemas de predicción de deforestación (i.e., promedio histórico, funciones de tiempo y modelado en función de variables)(VCS, 2012). Se buscó que el flujo metodológico a implementar tuviese un enfoque de modelado en función de variables y que satisfaciese los siguientes requerimientos:

Requerimiento 1: Los modelos generados deben estar explícitamente espacializados y restringidos a tener como variables predictoras solo aquellas que deriven de actividades en las que el Estado tenga plena capacidad de decisión (i.e., creación de ANP, construcción de carreteras, Zonificación Económica Ecológica, entre otros).

Requerimiento 2: Se debe poder experimentar con modelos estadísticos tradicionales y con

modelos de aprendizaje de máquina novedosos encontrados en el estado del arte.

Requerimiento 3: La metodología debe ser completamente replicable y escalable a todo el territorio nacional.

Estos tres requerimientos fueron tomados como lineamientos durante el desarrollo de la investigación. En cada etapa del estudio, los métodos, datos y plataformas de trabajo fueron elegidos o moldeados a fin de satisfacer esta necesidad. Teniendo en cuenta lo antes mencionado, se propusieron 5 objetivos específicos:

- Profundizar el entendimiento de las dinámicas económicas, sociales y de carbono en la Amazonía mediante una exhaustiva revisión de la literatura
- Diseñar un sistema de análisis de datos que satisfaga los requerimientos del objetivo general
- Estudiar la relación que existe entre las diferentes variables de la base de datos construida y su influencia en las tasas de deforestación en la Amazonía
- Proponer y validar modelos de predicción de riesgo de deforestación
- Estimar las emisiones de gases de efecto invernadero de los principales proyectos viales

El cumplimiento de estos objetivos específicos es imperante para satisfacer el objetivo general. Sin embargo, es necesario comenzar proponiendo tres hipótesis fundamentales. Se espera que estas hipótesis sean validadas or rechazadas a partir de los resultados de la investigación. La primera hipótesis H1 está relacionada con la importancia que tendrá el Plan Nacional Vial en el cambio climático y los compromisos nacionales. La segunda hipótesis H2 busca responder la interrogante recurrente que surge a partir de la popularidad de los modelos aprendizaje profundo (i.e., Deep Learning) y su potencial superioridad frente a los modelos estadísticos tradicionales. La última hipótesis H3 está vinculada a la necesidad de implementar los modelos propuestos en la toma de decisiones, sobre todo por la complejidad de su elaboración y el alto costo computacional de estos. Adicionalmente, a lo largo del desarrollo de este manuscrito, se propondrán distintas sub-hipótesis que se originan a partir de interrogantes específicas y enfocadas en los métodos y supuestos con los que se inicia la investigación. Las principales interrogantes de este proyecto son las siguientes:

Hipótesis 1 (H1): *Los compromisos ambientales del país subestiman las emisiones generadas por los cambios de uso de suelo.*

Hipótesis 2 (H2): *Los modelos de ensamblado y de aprendizaje profundo superiores que los modelos estadísticos tradicionales*

Hipótesis 3 (H3): *Es posible replicar y escalar los modelos de predicción para su utilización en la toma de decisiones*

Capítulo 2

Estado del arte

2.1. Cambio de uso de suelos: alcances y *statu quo*

El cambio de uso de suelo está definido como el cambio claro y permanente en el uso de suelo que se asocia con modificaciones en la cobertura de la superficie y en las reservas de carbono (Watson et al., 2001). Estos cambios son una fuente considerable de emisiones de gases de efecto invernadero; de hecho, representan alrededor del 9 % del total de emisiones globales (Le Quéré et al., 2013). En este sentido, esta relevancia global ha ocasionado que este sector sea sido considerado como prioritario por muchas instituciones (Watson et al., 2001; Van Stappen et al., 2011). Estos cambios se dividen en cambios directos del uso de suelos (LUC) y cambios indirectos del uso de suelos (iLUC). Por un lado, el primero corresponde a un cálculo simple de la superficie transformada en el mismo lugar e instante donde se realiza la actividad o proceso que se desea estudiar. En lo que refiere al estudio de carreteras, (Larrea-Gallegos et al., 2017) incluyeron los cambios directos ocasionados por la transformación de terreno forestal a superficie de rodadura y derecho de vía en un estudio de Análisis de Ciclo de Vida (ACV) realizado a la construcción de un proyecto vial en el departamento de Madre de Dios. En este caso, en el cálculo de las emisiones se requirió cuantificar las hectáreas, por metro de carretera, que dejaron de ser bosque para luego ser convertidas a unidades de CO_2eq. Este cálculo es computacionalmente trivial pero, debido a que el proyecto se ubicó en la Amazonía, se requirió del uso de otros métodos y modelos de descomposición de biomasa para determinar el valor de emisión final (Larrea-Gallegos et al., 2017). Por otro lado, los iLUC son todos aquellos cambios que se generan en otras áreas y en diferentes periodos temporales, distintos a los de la actividad estudiada pero que no existirían de no realizarse dicha actividad. Si se toma a la carretera como ejemplo, los iLUCs corresponderían a toda la deforestación ocurrida fuera de la superficie de rodadura y de derecho de vía que ocurriese después de terminada la construcción. Como se puede suponer, realizar un cálculo de este efecto en cadena es sumamente complicado, sobre todo porque no es posible distinguir si la deforestación en ciertas partes se debe completa o parcialmente a la construcción de una carretera. En este sentido, distintos métodos han sido desarrollados para estimar los iLUCS y las emisiones de GHG asociadas. De manera general, es posible distinguir tres tipos de modelos de iLUCs en la literatura: biofísicos, económicos, y basados en reglas (Schmidt et al., 2015).

Los modelos biofísicos buscan relacionar la demanda de terreno y de cultivos con información física de rendimiento y datos estadísticos de deforestación (Schmidt et al., 2015). Los modelos económicos suelen basar su estructura en modelos de equilibrio general (GEM) o equilibro parcial (PEM) que incluyen información de la producción agrícola global y tablas de insumo-producto. Por último, los modelos basados en reglas son lineamientos que incluyen criterios de otras guías (i.e., PAS2050, GHG-protocol y PEF-guide) que toman en cuenta la ocupación del suelo en un periodo previo de 20 años y amortizan el valor de las emisiones de manera anual. Una amplia descripción de los modelos mencionados puede encontrarse en Schmidt et al. (2015). Estos métodos, utilizados fundamentalmente en el campo del ACV, tienen un enfoque en el que se busca entender el efecto del desplazamiento de cultivos a zonas forestales. Sin embargo, esta visión puede ser miope cuando se desea estudiar otros procesos o actividades que también están ligados a la deforestación y al cambio de uso de suelos. Adicionalmente,

su alto nivel de generalización puede conllevar a obtener resultados con alta incertidumbre si es que se desea estudiar procesos de poca escala o con características regionales establecidas (De Rosa, 2018).

2.2. Avances en teledección, aprendizaje automatizado y análisis basado en la nube

A finales de los años 80, diversos autores incursionaron en el uso de tecnologías de teledetección mediante el análisis de imágenes satelitales de alta resolución con la finalidad de estudiar la deforestación (Fearnside, 2003; Nelson and Hellerstein, 1997; Pfaff, 1999; Angelsen and Kaimowitz, 1999). Este fenómeno fue estudiado utilizando diferentes enfoques. Por ejemplo, desde un punto de vista económico, Nelson and Hellerstein (1997) proponía que las carreteras incentivaban deforestación debido a que disminuían el costo de acceso a las zonas forestales. Los modelos económicos que utilizó eran probados y refinados con datos empíricos obtenidos de distintas tomas de satélites y algoritmos de clasificación no supervisada. Sin embargo, el nivel de precisión de estos algoritmos estaba limitado a las capacidades de procesamiento de los ordenadores de aquella época (Congalton, 1991). Aunque considerar a la economía como principal factor de la deforestación era intuitivo y razonable, se demostró que este único factor no es suficiente para entender las dinámicas de deforestación a escalas menores a la nacional (Leblois et al., 2017).

En Leblois et al. (2017), se utilizó información mundial de deforestación en alta resolución, para actualizar y validar los modelos y resultados que se obtuvieron a lo largo de investigaciones realizadas durante los años 1990 y 2000. Leblois et al. (2017) procesó la deforestación anual por país durante los años de estudio para generar modelos de regresión utilizando variables independientes como exportación agrícola, terreno cultivado, densidad poblacional, entre otros. Los resultados de estas regresiones sirvieron para validar los modelos propuestos décadas atrás. Los autores concluyen que los determinantes de deforestación pronosticados en los años 90 siguen siendo válidos en la actualidad. Esto quiere decir que los modelos de predicción estimados tienen razonable certeza pese a que estos análisis son de escala global. La agricultura, como era de esperar, es una variable vinculada al crecimiento de carreteras y desarrollo urbano, aspectos de carácter nacional que no son evaluados en ninguno de los estudios analizados por Leblois et al. (2017). Finalmente, estos autores recomiendan realizar investigaciones relacionadas a la calidad de los bosques ya que tienen influencia sobre las políticas REDD+ y los incentivos económicos vinculados a este último. En las últimas décadas, tanto la academia como entidades gubernamentales han volcado esfuerzos para estudiar este fenómeno e implementar políticas de control. Las investigaciones más recientes incluyen variables características de la infraestructura, la geomorfología y el clima. Estas variables se determinan dependiendo del tipo factor que se desea profundizar (Perz et al., 2013; Baraloto et al., 2015; Barber et al., 2014; Miranda et al., 2014). Perz et al. (2013), por ejemplo, publicó una investigación en la cual analiza el cambio de la cobertura terrestre, asfaltado de carreteras, y la deforestación a nivel de comunidades a lo largo del trazo de la Carretera Interoceánica Sur (IIRSA por sus siglas en inglés), en los países de Perú, Bolivia, y Brasil. En este proyecto se realizó un análisis multivariado del cambio de la cobertura terrestre a través de las regiones que la IIRSA recorre a lo largo de los años 2005 y 2010. Esta investigación utiliza variables biofísicas, socioeconómicas, y de cobertura terrestre para realizar el análisis; así también, datos de las carreteras asfaltadas y no asfaltadas dentro de los países estudiados. Se tomaron alrededor de 200 muestras que consistían en visitas de campo, y recolección de testimonios y encuestas. El clima, elevación, distancia a mercados cercanos, estado de la carretera, entre otros, fueron utilizados como variables. Tres modelos de regresión lineal fueron generados para cada año y uno para la variación. Aunque se obtuvieron coeficientes de determinación (R^2) bastante altos, estos solo se enfocan en la deforestación de comunidades cercanas a la carretera; además, variables intrínsecas a la geometría de la vía son ignoradas.

Baraloto et al. (2015) proponen el estudio de la relación entre la degradación de bosques, la deforestación y las carreteras. Esta publicación es de relevancia ya que estudia la variable relacionada a la calidad de bosques, aquella que se menciona como relevante en las políticas REDD+. En esta

investigación se toma también como caso de estudio la región tri-fronteriza de Perú, Brasil y Bolivia. Se tomaron muestras a lo largo de la carretera utilizando "cuadrantes de vegetación" (vegetation plots en inglés). Estos últimos corresponden a un método en el que se delimitan cuadrantes dentro de los cuales se midió la biomasa superficial y subterránea a lo largo de los años 2008 y 2010. Se utilizaron imágenes satelitales para contabilizar el cambio de terreno deforestado y se digitalizaron los mapas de carreteras principales, secundarias y terciarias. En este caso, se utilizaron variables como la distancia a la carretera, distancia al centro urbano, distancia a los andes, tiempo que la IIRSA lleva pavimentada, entre otros. Finalmente, Baraloto et al. (2015) generaron un modelo de regresión lineal en el que se concluyó que, pese a que existe una alta correlación entre la distancia de la vía y la deforestación, este fenómeno no se replica cuando se analiza la distancia de la vía y la degradación del bosque. Los autores manejan posibles explicaciones entre las que se incluyen deficiencias de muestreo y alta heterogeneidad de la zona.

De forma casi paralela, Barber et al. (2014) publicaron los resultados de su investigación en la que se vincula por primera vez el efecto mitigador de las Áreas Naturales Protegidas (ANP), la deforestación y las carreteras. Este estudio recurre al uso de un modelo de regresión lineal y es la primera investigación que analiza futuros escenarios plausibles. En este caso se realizó un análisis empírico espacial en el que se calculó toda el área de deforestación adyacente a todas las carreteras legales e ilegales de la Amazonía brasilera. La importancia de considerar las carreteras ilegales en el análisis se debido a que estas eran construidas sin autorización y están, comunmente, ligadas a actividades altamente generadoras de deforestación. Las carreteras estatales fueron recopiladas a partir de información gubernamental; y aquellas ilegales, a partir de clasificación utilizando imágenes satelitales. Este último método es de suma relevancia ya que muestra un inicio importante en el uso de clasificadores para el mapeo de carreteras a escala nacional. Los resultados señalan que el 95 % de la deforestación adyacente a todas las vías brasileras, en promedio, ocurre dentro de los 5.5 km más cercanos. De forma similar, Miranda et al. (2014) estudiaron la relación que existe entre las ANP, las comunidades nativas y la deforestación. Este estudio construyó un modelo de regresión y determinó una correlación entre la creación de ANP y la disminución de deforestación.

Aunque se reconoce la importancia de las ANP como elementos mitigadores, tal y como señala Weisse and Naughton-Treves (2016), poco se ha debatido respecto a las zonas de amortiguamiento (ZA). Estos autores señalan que es importante incrementar la integración de entes fiscalizadores en estas zonas, debido a que estos espacios tienen efectos mitigadores sobre la deforestación y las actividades mineras. Una explicación a este efecto es que las ZA estan destinadas como zonas de transición entre actividades restringidas y cotidianas, por lo que suelen ser consideradas como herramientas "teóricas" de conservación.

Para entender las limitaciones y aportes de los estudios encontrados en la literatura es necesario explorar la taxonomía del análisis de datos, disciplina en la cual la mayor parte de estos estudios recae. Estos trabajos pueden clasificarse como estudios dirigidos a la predicción y estudios con enfoque de análisis. Esta división corresponde al paradigma fundamental del análisis de datos en el que se busca representar algún fénomeno natural de manera simplificada. En el primer tipo de estudio se espera pronosticar fenómenos en escenarios futuros, mientras que el segundo tipo busca identificar patrones, caracaterísticas y extraer información del fenómeno que pueda resultar de utilidad. Más allá de esta distinción primaria, un estudio puede seguir dos filosofías metodológicas fundamentales: la filosofía de modelado de datos y la filosofía de modelado algorítmico. Por un lado, el primero utiliza modelos estocásticos establecidos con supuestos fuertes. La validación de estos modelos suele requiere pruebas de hipótesis, pruebas de bondad de ajustes, análisis de residuales, entre otros. Por otro lado, el segundo, asume que la estructura del modelo es desconocida y lo que se busca es una función $\boldsymbol{f(x)}$ que permita predecir $\boldsymbol{y}$ a partir de $\boldsymbol{x}$ (Breiman, 2001). Una interpretación tangible de esta distinción sería la que separa a los modelos estadísticos tradicionales (e.g., modelos lineales, regresiones logísticas, entre otros) de los modelos de aprendizaje de máquina (e.g., Árboles de decisión, Redes Neuronales, entre otros) (ver figura 2.1 y figura 2.2). Finalmente, la selección de los modelos y de la metodología responde

a las necesidades y preguntas particulares de cada proyecto. Igualmente, los recursos (i.e., poder computacional y presupuesto) y la disponibilidad de datos son condicionantes de relevancia.

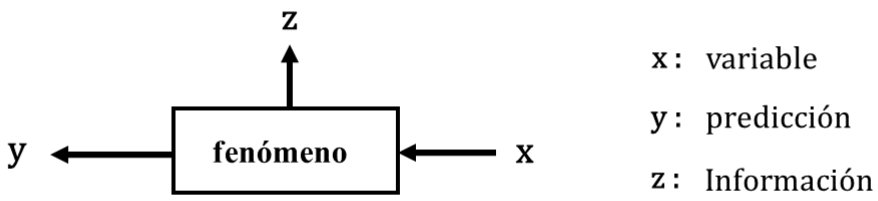

Figura 2.1: Representación gráfica del paradigma del análisis de datos. Adaptado de Breiman (2001)

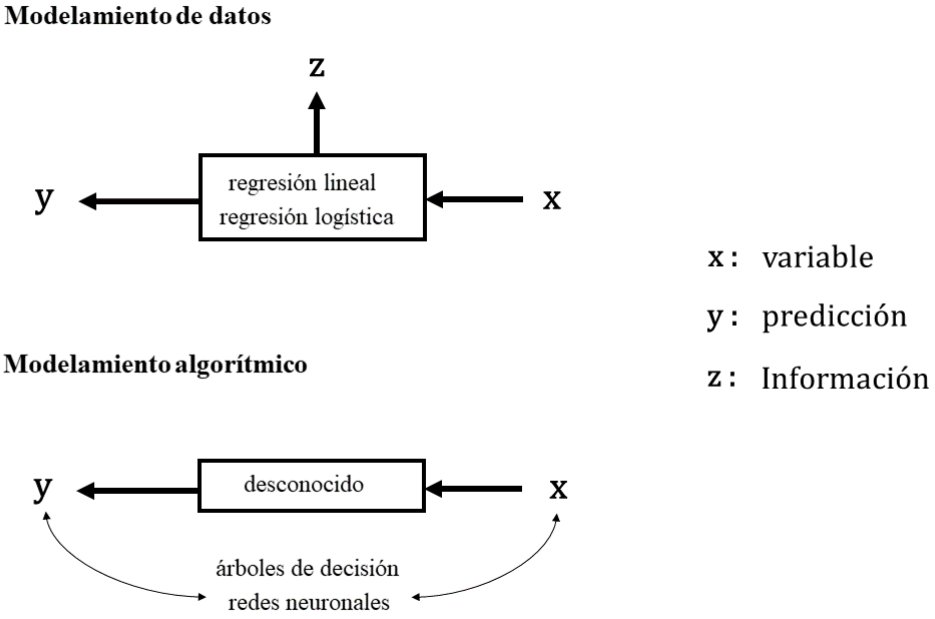

Figura 2.2: Representación gráfica de las dos filosofías de construcción de modelos en el análisis de datos. Adaptado de Breinman (2001)

Los estudios antes mencionados se caracterizan por el uso de modelos estadísticos convencionales. De hecho, este enfoque estadístico en el análisis y predicción del riesgo de deforestación ha predominado en la literatura. Sin embargo, en 2004, Mas et al. (2004) publicaron un estudio de predicción de deforestación que propuso por primera vez el uso de modelos de redes neuronales (NN), denominados perceptrones multicapa por su simpleza. Las NN son modelos no lineales que pueden adaptarse, en teoría, a cualquier distribución de datos y son capaces de aproximar cualquier fenómeno (Yadav and Sood, 2013). Mas et al. (2004) propusieron una arquitectura de red con 3 a 8 capas conectadas (i.e., una capa de entrada, varias capas ocultas y una capa de salida). El modelo entrenó sus parámetros a partir de los datos ingestados y mediante multiples etapas demominadas épocas de entrenamiento. Se utilizaron 6 variables predictoras y 2 variables dependientes (i.e., deforestación y regeneración de bosque). Los resultados mostraron que estos modelos tienden a sobre entrenarse y no ser exitosamente

generalizables. Sin embargo, Mas et al. (2004) no realizó comparación alguna frente a otro tipo de modelo con la misma base de datos. En contraste con lo antes mencionado, Mayfield et al. (2017) realizaron una profunda comparación del desempeño de los distintos métodos estadísticos y de aprendizaje de máquina. Este estudio consideró a la deforestación como un fenómeno de rango binario (i.e., 1 si ocurre, 0 si no ocurre) y utilizando 18 variables predictoras. Se evaluaron modelos lineales generalizados (GLM) y generalizados mixtos (GLMM), NN, redes bayesianas (BN) y procesos gaussianos (GP). El flujo metodológico consistió en extraer los datos de fuentes de libre acceso y procesarlos en softwares dedicados al manejo de información georreferenciada, entrenar los modelos en distintas plataformas, y finalmente, analizar los datos de manera independiente utilizando distintos softwares. Aunque este sistema parece un protocolo de investigación razonable, algunos modelos fueron implementados de manera limitada debido a la falta de poder computacional (i.e., NN). En el caso de las redes neuronales implementadas, Mayfield et al. (2017) consideraron de 1 a 2 capas ocultas con 30 a 60 neuronas, arquitectura relativamente limitada si se compara con los modelos del estado del arte. Esta limitación se debe a que la información geoespacial, así como los modelos seleccionados, son complejos de implementar y, sobretodo, hacen que la iteración por la búsqueda de los mejores parámetros sea tediosa.

El sistema utilizado por Mayfield et al. (2017) muestra que la complejidad de un estudio puede escalar debido a la necesidad de utilizar distintas plataformas de trabajo y a los largos tiempos de computo, un problema que viene siendo solucionado por la computación en la nube. De acuerdo al Instituto Nacional de Estándares y Tecnología (NIST), la computación en la nube (*cloud computing* en inglés) se define como un modelo diseñado para permitir el acceso *on-demand* a una gama de recursos computacionales configurables (e.g., redes, servidores, almacenamiento, aplicaciones y servicios) que pueden ser rápidamente provisionados y distribuidos con el mínimo esfuerzo de gestión o interacción con el distribuidor del servicio (Ahmad Bhat et al., 2011). De esta forma, en la última década se han venido implementado diversas plataformas de procesamiento en la nube (i.e., Google Cloud Plataform, Azure, Amazon Web Service) que han permitido a la comunidad científica facilitar la expansión de las barreras computacionales de la investigación. En lo referido al GIS, la novedosa plataforma Google Earth Engine (Gorelick et al., 2017) utiliza los servicios de computación en la nube para realizar computación paralelizada dedicada a operaciones con información georeferenciada. Esta plataforma cuenta con su propio interfaz de programación de aplicaciones (*API* por sus siglas en inglés) y tiene almacenado petabytes de imágenes satelitales de diversas fuentes públicas. Desde su aparición en 2015, diversas contribuciones de relevancia fueron realizadas. Por ejemplo, mapas de deforestación de alta resolución a escala mundial (Hansen et al., 2013), detección de áreas incendiadas en toda América Latina (Bastarrika et al., 2018), mapeos de la superficie urbana mundial (Liu et al., 2018), entre otros.

2.3. Métodos de cálculo de emisiones de gases de efecto invernadero

Existen diferentes métodos propuestos por el IPCC para la estimación del contenido de carbono superficial en los diferentes tipos de superficie (IPCC, 2006). Estos métodos clasifican los distintos tipos de suelos y otorgan un contenido de carbono de acuerdo a determinadas características del suelo. Aunque estos métodos son utilizados mundialmente para la construcción de los Inventarios Nacionales de Gases de Efecto Invernadero, carecen de resolución espacial y utilizan información promedio. En contraste, Asner et al. (2014) presentaron mapas de densidad de carbono superficial de 1 hectárea de resolución del territorio peruano. Estos mapas fueron construidos utilizando información de sensores LiDAR y sobrevolando toda la Amazonía. Para ello, los autores tomaron muestras del contenido de carbono en 1 hectárea y del espectro de la muestra para ingestarlos en un modelo de aprendizaje automatizado (i.e., random forest) (Mascaro et al., 2014). Este mapa representa la aproximación más precisa del contenido de carbono superficial del territorio nacional.

En lo que respecta a la estimación de gases de efecto invernadero, el criterio aceptado y utilizado por la comunidad científica es el propuesto por el IPCC (IPCC, 2006). Estos lineamientos sugieren el uso de factores de caracterización para expresar las emisiones en una unidad única de medición

denominada CO_2eq. Esta unidad expresa el potencial de cambio climático tomando como referencia 1 kg de CO_2 gaseoso. El desafío en el cálculo de las emisiones de gases de efecto invernadero está en determinar las cantidades y los tipos de gases que son emitidos en los procesos estudiados. Larrea-Gallegos et al. (2017) midió los impactos del CO_2 y el CH_4 de cada hectárea deforestada. Sin embargo, el fin de vida de cada árbol deforestado es incierto ya que este puede ser utilizado como madera de mueble, quemado como combustible, o simplemente dejado de lado para su descomposición natural. En el estudio de Larrea-Gallegos et al. (2017) se determinó que los residuos de desbroce eran dejados de lado para su descomposición. En ese caso, se consideró que el 97.03 % del carbono se transformaba en CO_2 y el resto en CH_4. En lo que refiere al cálculo del contenido de carbono en el suelo, este fue estimado a partir del contenido de carbono superficial, siguiendo las recomendaciones y el modelo propuesto por Saatchi et al. (2011). Finalmente, debido a que se consideró todo el ciclo de vida, los autores asumieron que, eventualmente, el carbono contenido en la biomasa terminaría siendo emitido al medio ambiente. Este último criterio obedece al periodo de análisis seleccionado como parte de la metodología de propuesta por el IPCC (i.e., 100 años).

Capítulo 3

Materiales y métodos

3.1. Construcción de modelos de predicción de deforestación

3.1.1. Selección de las zonas y sub-zonas de análisis

La región de interés (denominada ROI de ahora en adelante) está delimitada por toda el área del Bioma Amazónico comprendida dentro del territorio peruano. La definición de este Bioma obedece a la delimitación determinada en la clasificación de ecorregiones realizada por Olson y Dinerstein (2002). La ROI tiene una extensión aproximada de 700 mil km^2 e incluye porciones de los territorios de los departamentos de Amazonas, Loreto, San Martín, Huánuco, Ucayali, Pasco, Junín, Cusco, Madre de Dios y Puno (ver Figura 3.1). De igual forma, la ROI elegida es de particular importancia debido a que esta incluye a los más importantes Parques Nacionales, como el Parque Nacional del Manu o el Parque Nacional Sierra del Divisor. Alrededor del 63 % del territorio nacional es considerado como superficie arbórea (OECD, 2015); no obstante, se espera que importantes proyectos de infraestructura e inversión se ejecuten en la zona de estudio en los próximos 10 años (e.g., el tren Bioceánico y la carretera nacional PE-4S) (Gestión, 2015;2018). Adicionalmente, en los últimos años se ha detectado un incremento considerable de producción agrícola. Mucho de esto está directamente vinculado con la expansión de plantaciones de palma aceitera (Vijay et al., 2018), así como el aumento de la agricultura de pequeña escala (Ravikumar et al., 2017). Esta ROI incluye también las zonas de afectación minera localizadas en el departamento de Madre de Dios. Estas areas son de particular interés debido a que la deforestación de la zona está también asociada a emisiones de contaminantes altamente tóxicos como el mercurio (Asner et al., 2013; Kahhat et al., 2019).

Aunque la extensión del ROI es vasta, las zonas en donde la tasa de deforestación se ha incrementado se encuentran focalizadas e idenficadas (Finer et al., 2018) De igual modo, existe una clara diferencia entre los factores de deforestación dependiendo de la ubicación geográfica que se observe, por lo que existe una motivación intrínseca hacía sub-dividir el ROI. Este enfoque que busca analizar la problemática tomando en cuenta la regionalización ha sido aplicado por otros autores, aunque de diferente manera (Delgado, 2008). En este sentido, en esta investigación se partirá del supuesto *a priori* de que existe una clara heterogeneidad entre las distintas regiones en las que ocurre deforestación, por lo que se propuso la siguiente hipótesis SH4.

Hipótesis 4 (SH4): *No existe un modelo generalizable para toda la Amazonía peruana que considere todas las variables asociadas a la deforestación y tenga un alto nivel de predicción.*

La premisa de la inexistencia de un modelo completamente generalizable para toda la Amazonía motiva la fragmentación del ROI en zonas (denominadas *clusters* de ahora en adelante) cuya cantidad y ubicación son inicialmente desconocidas. Debido a que el fénomeno de interés es la deforestación, se comenzó proponiendo una cantidad arbitraria k de *clusters* de deforestación utilizando la información proveniente del Mapa de Cambio de Covertura Arbórea de Hansen Hansen et al. (2013). Este planteamiento motiva a agrupar el territorio de estudio en una cantidad de *clusters* lo suficientemente grande para que las características sean similares y se maximice la predictibilidad. Sin embargo, el número de

Figura 3.1: Región de Interés seleccionada del Bioma Amazónico

cluster debe ser el menor posible, de modo que los modelos tengan un mínimo nivel de generalización y sean prácticos en su implementación. Para esto, se utilizó el algoritmo de *K-medios* para clasificar los píxeles de deforestación en 8 grupos o clases. *K-medios* es un algoritmo de clasificación no supervisada que agrupa los datos en K clases. En este caso, la distancia estadística (i.e., distancia euclidiana) del píxel clasificado a la media del resto de píxeles de su clase es menor que la distancia al resto de medias de otras clases (Witten et al., 2017). El cálculo de las medias toma en cuenta las variables latitud, longtitud y altura de los casos de cambio de cobertura arbórea en el periodo 2010-2017 Hansen et al. (2013). La Figura 3.2 muestra un ejemplo gráfico de la secuencia iterativa del algoritmo de *kmeans*. Se puede observar que la posición de los puntos iniciales varía durante cada iteración hasta converger a un punto que representa la media del grupo correspondiente.

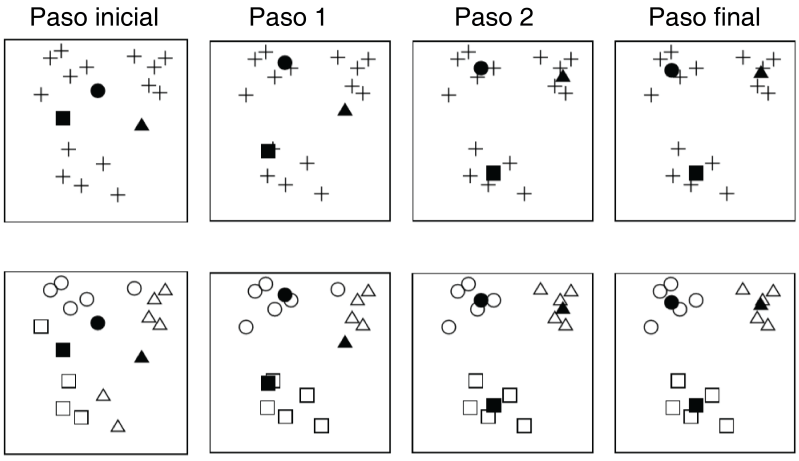

Figura 3.2: Representación gráfica del algoritmo de *K-medios*. Adaptado de Witten et al. (2017)

La Figura 3.3 permite apreciar que las zonas resultantes son similares en ubicación y altitud. Esta selección permitió definir zonas de estudio que compartieron un criterio homogéneo de selección. Este agrupamiento responde no solo a la necesidad de analizar previamente la distribución de los datos, sino a que es computacionalmente conveniente trabajar con modelos de tamaño reducido en las primeras etapas de entrenamiento. El valor de k fue aumentando desde 4 hasta 8, siendo este último el número final de *clusters* elegido. Esta selección fue arbitraria pero estuvo condicionada por una revisión de las carácterísticas de las zonas deforestadas. Se descartaron 3 de las 8 regiones agrupadas debido a que estas contenían a muy poca cantidad de píxeles y se asumió que no eran relevantes para el análisis. En la tabla 3.1 se indican los *clusters* elegidos y se incluye una breve descripción de estos. Finalmente, es importante señalar que, debido a la naturaleza del algoritmo de *K-medios*, la pertenencia de cada píxel a un determinado *cluster* puede ser diferente en cada iteración. Esto quiere decir que la repetición del algoritmo no genera, necesariamente, los mismos resultados. Sin embargo, se observó que en distintas iteraciones, los *clusters* tienden a converger a determinadas zonas y son solo pocos píxeles los que alternan de zonas en cada iteración. En este sentido, se consideró que la replicabilidad del estudio no se ve afectada por el uso de este algoritmo de agrupamiento.

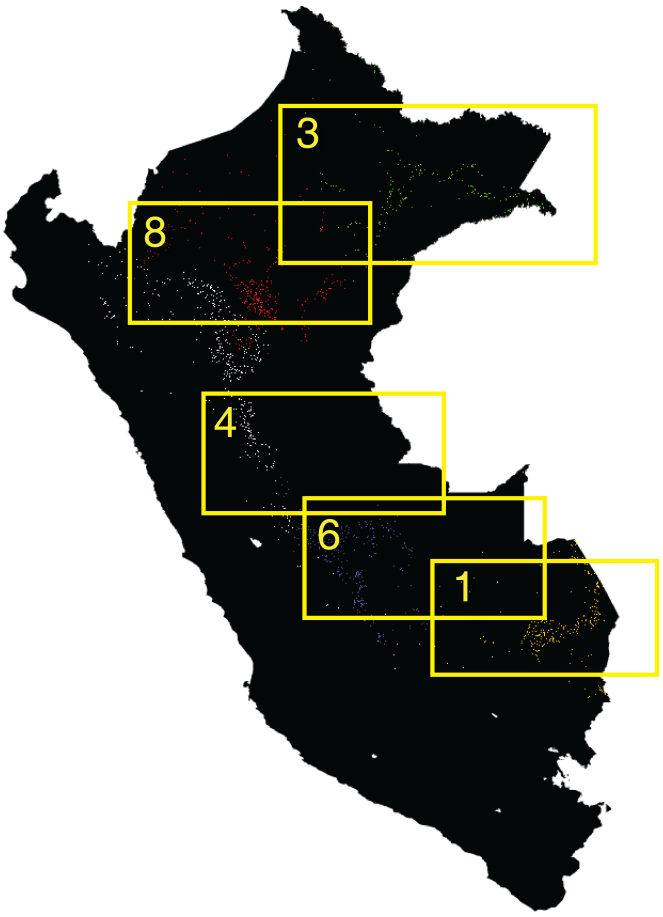

Figura 3.3: Ubicación de los clusters obtenidos con el algoritmo de *K-medios*. Los distintos colores indican diferentes clusters

.

3.1.2. Recolección y procesamiento de datos

Para construir la base de datos del modelo se utilizó un Sistema de Información Geográfica (GIS, por sus siglas en inglés) con información georreferenciada de libre acceso. Este tipo de información se caracteriza por poseer una estructura en la que los datos pueden ser atribuidos a una o múltiples ubicaciones geográficas específicas. Dicho de otro modo, las dos dimensiones correspondientes a cada coordenada (i.e., latitud y longitud) pueden ser expandidas si se adicionan más dimensiones (e.g., altu-

Tabla 3.1: Descripción de clústers analizados en la investigación

Clúster	Código	Descripción
1	C1	Comprende principalmente la región de Madre de Dios. Incluye deforestación por zonas mineras.
3	C3	Agrupa la deforestación ocurrida en los alrededores de las ciudades de Iquitos y Nauta. Incluye deforestación por inundaciones.
4	C4	Agrupa la deforestación ocurrida en los alrededores de Pucallpa. Se considera la deforestación por aumento de plantaciones de palma aceitera.
6	C6	Este cluster contiene a la deforestación ocurrida en la zona sur de Ucayali, al noreste de la región Cusco.
8	C8	Resulta de la combinación de la deforestación ocurrida en Yurimaguas y la deforestación en la fronte de Amazonas y Loreto.
Perú	CT	Este clúster representa el agrupamiento de la deforestación de los 5 clústers antes descritos.

ra, temperatura, entre otros) (Gold, 2016). Para manejar y visualizar estos datos multidimensionales se utilizaron las herramientas QGIS v2.18 (QGIS, Development Team, 2009) y Google Earth Engine (Norelick et al., 2017). El primero es un software GIS de escritorio de código abierto; y el segundo, una plataforma basada en la nube destinada al análisis de información georreferenciada. Dependiendo del tipo de análisis que se desee realizar, la información espacial se representa, principalmente, mediante dos tipos de modelos: los modelos de grillas ráster y los modelos vectoriales. El primero puede describirse como un conjunto de valores ordenados en filas y columnas en un plano bidimensional donde cada celda, también denominado pixel, posee un par de coordenadas y puede almacenar más valores (Mclnerney y Kempeneers, 2014). Una fotografía digital convencional, por ejemplo, está representada a través de un modelo de grillas ráster debido a que cada píxel almacena 3 valores, cada uno correspondiente a la reflectancia de las bandas rojo, verde y azul. En el caso del segundo tipo de modelo, los datos espaciales se representan como vectores que cuentan con una coordenada de inicio y otra de final. Estos objetos pueden ser puntos, líneas o polígonos y sus características son almacenadas en una tabla de atributos (Mclnerney y Kempeneers, 2014). La Figura 3.4 muestra la diferencia entre estos dos modelos.

En lo que se refiere a los datos utilizados, estos provienen de distintas fuentes de libre acceso y se encuentran en formatos ráster y vectorial. La finalidad del uso de estos datos espaciales es poder extraer distintas variables predictoras para realizar una estimación atribuida a una sola coordenada; sin embargo, como se puede intuir, los datos pueden tener formatos incompatibles entre ellos debido a que utilizan diferentes modelos de representación. En este sentido, se decidió transformar los datos vectoriales a datos ráster a través de un proceso denominado rasterización. Por más que la nueva base de datos posea un formato homogéneo, la determinación de variables puede requerir un procesamiento adicional. Por ejemplo, en ciertos casos, la variable predictora podría corresponder directamente al tipo de dato obtenido de la fuente; en otros, puede ser necesario implementar algún algoritmo adicional que permita obtener la variable de interés a partir de los datos ráster originales, como sucede con los mapas viales. En este caso, los datos de ubicación de carreteras son inútiles si no se interpreta que la cercanía a estas vías puede ser un factor determinante en la deforestación; ergo, se debe generar la variable çercanía a carretera". Esta heurística que busca transformar los datos para obtener más información se denomina ingeniería de características, y es en algunos casos, el factor determinante en el éxito de la construcción de modelos. La Tabla 3.2 indica los nombres de las bases de datos consultadas y sus metadatos (i.e., resolución, temporalidad y tipo de modelo).

Aunque muchos de los fenómenos asociados a la deforestación han sido identificados, la tarea de seleccionar las mejores variables en la elaboración de un modelo no resulta trivial ya que en muchos casos el nivel de predictibilidad puede depender de una buena selección de estas (Reid Turner et al.,

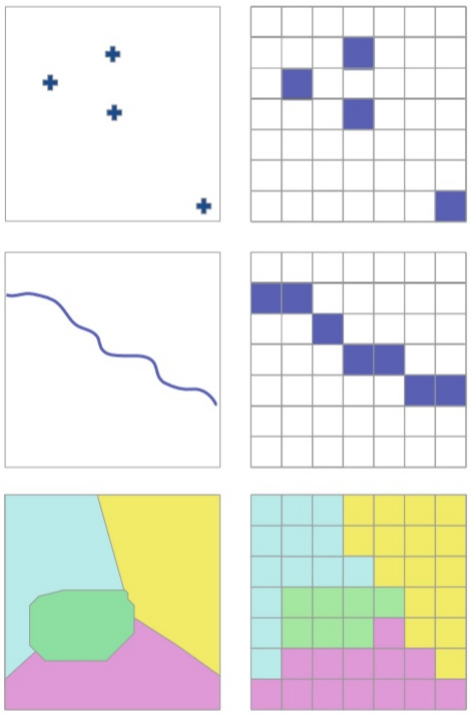

Figura 3.4: Representaciones de puntos, líneas, y polígonos utilizando un modelo ráster (derecha) y un modelo de vectores (izquierda) (extraído de Mclnerney y Kempeneers (2014))

Tabla 3.2: Metadatos de la información georreferenciada utilizada

Nombre	Tipo	Fuente	Resolución(m/px)	Temporalida
Cambio de cobertura arbórea	Ráster	Hansen et al. (2012)	30	2010-2017
Mapa vial del Perú	Shape	MTC(2016)	NaN	2017
Áreas Naturales Protegidas	Shape	MINAM (Ministerio del Ambiente)	NaN	2010-2017
Pueblos y ciudades del Perú	Shape	MTC(2016)	NaN	2017
Digital Elevation Model	Ráster	USGS	30	NaN

1999; Guyon et al., 2003). De hecho, a lo largo de la última década, los estudios de deforestación han ido adicionando nuevas variables, algunas de estas resultantes de la transformación de variables plenamente conocidas a partir de la transformación de datos ya existentes. La Tabla 3.3 muestra un resumen de las variables más utilizadas en modelos de predicción y análisis, tanto estadísticos como de aprendizaje automatizado, encontrados en la literatura. En la mayoría de los casos, el descarte, selección o transformación de variables no responde a criterios únicamente arbitrarios, sino que depende de las restricciones del tipo de modelo seleccionado y de la precisión que se alcance en las etapas de entrenamiento y validación. La selección del tipo de modelo depende de la cantidad de datos y variables disponibles, y en la mayoría de casos, mientras más se conozca el fenómeno que se desea predecir, más sencilla será la selección de un modelo adecuado. Esta lógica corresponde al adagio: *there ain't no such thing as a free lunch*, que dice, en resumen, que no existe un solo modelo que funcione o sea generalizable para todos los casos; lo que implica que cada fenómeno requiera un modelo particular (Wolpert, 1996; Domingos and Pedro, 2012).

Tabla 3.3: Variables utilizadas en la construcción de modelos

Variables	Rango	Resolución	Fuente	Utilización previa
Densidad de carbono (MgC)	[0,135]	100m/px	Asner et al. (2014)	NaN
Cambio de cobertura arbórea	[0,1]	30m/px	Hansen et al. (2015)	NaN
Altura (m)	[0,6000]	30m/px	USGS	Mayfield et al. (2015), Bax et al. (2016)
Distancia a centro urbano (km)	[0,15]	30m/px	MTC	Mayfield et al. (2015), Bax et al. (2016), Mas et al. (2004)
Distancia a ANP	[0,15]	30m/px	MINAM (Ministerio del Ambiente)	Mayfield et al. (2015), Barber et al. (2014)
Distancia a ZA	[0,15]	30m/px	MINAM (Ministerio del Ambiente)	Barber et al. (2014)
Distancia a carretera nacional	[0,15]	30m/px	MTC	Mayfield et al. (2015), Barber et al. (2014), Mas et al. (2004)
Distancia a carretera departamental	[0,15]	30m/px	MTC	Mayfield et al. (2015), Barber et al. (2014), Mas et al. (2004)
Distancia a carretera vecinal	[0,15]	30m/px	MTC	Mayfield et al. (2015), Barber et al. (2014), Mas et al. (2004)
Distancia a cualquier vía	[0,15]	30m/px	MTC	Barber et al. (2014)
Latitud	[-70,-71]	100m/px	Earth Engine	Mayfield et al. (2015), Mas et al. (2004)
Longitud	[-8,-12]	100m/px	Earth Engine	Mayfield et al. (2015), Mas et al. (2004)

Las variables propuestas buscan contener la mayor información de cada observación y que a su vez contribuyan a la capacidad predictiva del modelo. A continuación se describirá brevemente las características de las variables utilizadas:

Densidad de carbono: Esta variable corresponde a los datos obtenidos directamente del Mapa de Densidad de Carbono elaborado por Asner et al. (2014). Este mapa ráster contiene en cada píxel el valor del contenido de carbono superficial correspondiente. Este valor es, hasta el momento, la mejor aproximación a la densidad de carbono real del territorio peruano encontrado en la literatura. El mapa fue elaborado utilizando información LiDAR (i.e., detección de distancias mediante laser) obtenida a partir de sobrevuelos realizados al territorio peruano. El modelo utilizado (i.e., random forest) utilizó datos reales tomados de distintas zonas de estudio a lo largo de toda la Amazonía para ser entrenados (Mascaro et al., 2014). Finalmente, el mapa resultante contiene información a una resolución de 100 metros y viene siendo utilizada por el Gobierno y la academia.

Cambio de cobertura arbórea: Esta variable es obtenida a partir del Mapa Global de Cambio de Bosques, producido por Hansen et al.(2014), que se encuentra disponible en la plataforma de GEE. Este mapa ráster fue construido utilizando modelos de aprendizaje automatizado a partir de muestras de zonas deforestadas tomadas en campo. Interesantemente, este mapa fue la primera implementación de GEE en un proyecto de investigación y contiene información global de cambio de cobertura arbórea. Para esta investigación se utilizó la versión 1.5 del mapa que incluye estimaciones de desforestación

producidas en el periodo 2000-2017. La imagen multibanda contiene datos espectrales de cada píxel para los años 2000 y 2017. De igual forma, la información indica el año en el que el píxel fue deforestado. Se seleccionaron todos los píxeles deforestados en el periodo 2000-2017 y sus características expresadas de forma binaria (i.e., 0: no deforestado y 1: deforestado).

Altura: Esta variable corresponde a un mapa global de elevación digital (DEM, por sus siglas en inglés). Este DEM fue elaborado por Farr et al.(2007) utilizando los datos obtenidos por el Shuttle Radar Topography Mission (SRTM, por sus siglas en inglés). El mapa ráster muestra la topografía del planeta a un resolución de 30 metros tomada el mes de febrero del año 2000. Los valores contenidos en cada píxel se encuentran en el rango de 0 a 8700 metros y, para esta investigación, solo se enmascaró la zona de estudio delimitada por el territorio peruano.

Distancia a centro poblado: Esta variable indica la distancia euclidiana, en metros, hacia el centro poblado más cercano. Esta distancia se encuentra en el rango de 0 (si el píxel se ubica exactamente en el pueblo) y 300000 metros. Los centros poblados considerados son todos aquellos registrados por el MTC (2018), incluyendo centros poblados urbanos y rurales.

Distancia a ANP: Esta variable indica la distancia euclidiana, en metros, hacia el Área Natural Protegida más cercana. Estas áreas se definen como espacios destinados a conservar la diversidad biológica y demás valores asociados de interés cultural, paisajístico y científico (Resolución Presidencial 57-2014-SERNANP). En este sentido, se consideraron todas las ANP incluidas en MINAM (Ministerio del Ambiente). Se decidió considerar el efecto de las ANP en el modelo debido a la existencia de evidencia científica que señala que estas zonas generan un efecto de mitigador de deforestación (Cropper et al., 2001; Miranda et al., 2014; Barber et al., 2014). Esta variable se encuentra en un rango de 0 (si el píxel se encuentra en el ANP) y 100000 metros.

Distancia a ZA: La distancia a una Zona de Amortiguamiento indica la distancia euclidiana, en metros, a la zona que se encuentra adyacente a una ANP y está destinada a garantizar su protección (Resolución Presidencial 57-2014-SERNANP). Esta zona es un espacio de transición entre las ANP y las zonas no protegidas en el que las actividades son controladas (e.g., agricultura, urbanización, entre otros). Esta variable es considerada de relevancia debido a que en los últimos años se ha puesto en debate la utilidad de estas áreas como elementos garantizadores de protección de las ANP (Weisse and Naughton-Treves, 2016).

Distancia a carretera nacional: Esta variable indica la distancia euclidiana hacia la carretera nacional más cercana. Esta denominación es otorgada por el MTC y corresponde a la clasificación establecida en el Mapa Vial Nacional (DECRETO SUPREMO 011-2016-MTC). Las carreteras nacionales se dividen en longitudinales y transversales. Las primeras conectan las fronteras norte y sur del país; mientras que las segundas, la costa y la selva. Estas infraestructuras son diseñadas teniendo en cuenta altos valores de Índice Medio Diario Anual (unidad de medición de tráfico) y son siempre asfaltadas. Estas vías son relevantes desde la perspectiva tomada por esta investigación debido a que se tiene registro de los efectos que la carretera transversal PE - 30C ha generado en la tasa de deforestación de las zonas adyacente a su trayectoria (i.e., carretera interoceánica Perú y Brasil) (Delgado, 2008). Adicionalmente, se cuenta con información de los trazos de futuras carreteras nacionales que serán construidas en los próximos años.

Distancia a carretera departamental: Esta variable tiene las mismas características que **Distancia a cerretera nacional** pero considerando que la distancia es calculada a una carretera clasificada como departamental. Estas carreteras se encuentran bajo jurisdicción de cada Gobierno Regional y complementan la función de las carreteras nacionales. El objetivo de estas carreteras es garantizar la continuidad en la comunicación de los departamentos colindantes (DECRETO SUPREMO 011-2016-MTC).

Distancia a carretera vecinal: Al igual que los casos anteriores, esta variable resulta del cálculo de la distancia a una carretera clasificada como vecinal. Estas carreteras tienen como función unir los principales centros poblados y centros de producción entre ellos, y con el resto del país. Estas vías son responsabilidad de los Gobiernos Locales e pueden indicar el último punto de una ruta. En este caso, la relevancia de esta variable se debe a la naturaleza expansiva de este tipo de vías. Dicho de otro modo, las carreteras vecinales suelen construirse constantemente, en muchos de los casos, al margen de la ley y, dependiendo de la zona, pueden estar destinadas exclusivamente a actividades ilícitas (Gallice et al., 2017). Estas infraestructuras no están diseñadas para tolerar un alto IMDA y, dependiendo de su ubicación, pueden ser pavimentadas o no. Existe evidencia científica que la construcción de estas carreteras puede estar ligada a grandes emisiones de gases de efecto invernadero, sobre todo si estas se localizan en la Amazonía (Larrea-Gallegos et al., 2017). A diferencia de las carreteras nacionales y departamentales, el Plan Vial Nacional no considera los futuros proyectos de construcción. En este sentido, la expansión víal, vista desde este nivel, responde a motivaciones de los agentes interesados que se expresa a través de los presupuestos participativos y la intervención de la sociedad.

Distancia a cualquier vía: Esta variable resulta del calculo de la distancia a cualquiera de los tres tipos de carreteras antes mencionadas. La presente investigación incluyó esta variable para determinar si su participación tendría alguna influencia en los resultados.

Latitud y Longitud: Estas variables indican la ubicación geográfica de cada observación y fueron propuestas con la finalidad de considerar las características locales en cada uno de los modelos a implementar.

Finalmente, con este conjunto de rásters de una sola banda fue agrupado para construir un ráster final multibanda. Este último se puede interpretar como una base de datos estructurada en la que cada píxel es una observación y cada banda una variable. En este sentido, el entrenamiento y validación de los distintos modelos fue realizado utilizando una muestra de esta gran base de datos. Para ello, se utilizó una técnica denominada muestreo estratificado. La característica estratificada de esta operación se debe a que se dividió la capa en 2 estratos correspondientes a las clases que se desean predecir (i.e., deforestado y no deforestado) y se seleccionó una muestra aleatoria de cada estrato con la misma cantidad de puntos. Este enfoque estadístico ha sido ampliamente utilizado en la estimación del cambio de uso de suelos, sobretodo debido a la existencia de datos desbalanceados (Stehman, 2012; Olofsson et al., 2014). Esta estrategia permite equilibrar la distribución de los datos de entrenamiento, maximizando la cantidad de observaciones de la clase de interés y equiparando esta con las observaciones de la clase complementaria.

El fenómeno de deforestación es un ejemplo en el que una de las clases binarias (i.e., deforestado) tiene una muy escasa ocurrencia en comparación con la clase complementaria (i.e., no deforestado). Esta desproporción de los datos ocasiona que el modelo se encuentre sesgado a predecir con mayor precisión la clase de mayor ocurrencia (i.e., no deforestado) y no la clase de interés (i.e., deforestado) (Chawla, 2009; Haibo He and Garcia, 2009; Mayfield et al., 2017). Finalmente, se seleccionó el 80 % de los datos para ser utilizados en el entrenamiento y la validación, y el 20 % restante fue designado como grupo de prueba. Este grupo de entrenamiento y validación fue a su vez dividido en 2 partes iguales, esto con el fin de ser utilizado en el proceso de validación propio del entrenamiento. El flujo metodológico de los procedimientos seguidos en el procesamiento y selección de datos está representado graficamente en la Figura 3.5.

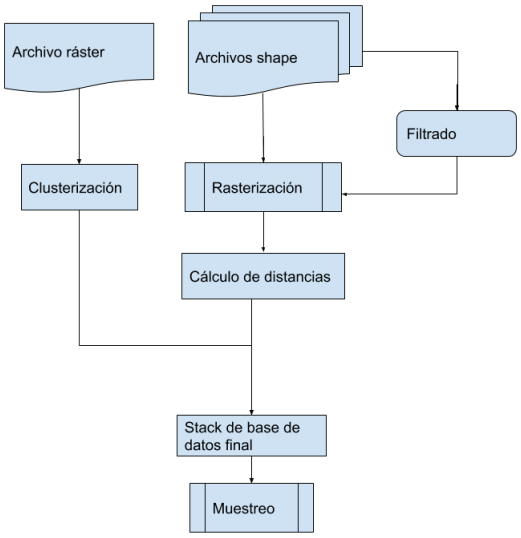

Figura 3.5: Flujo metodológico del procesamiento y la preparación de los datos espaciales

Con el uso de las variables antes descritas, se construyeron los modelos y se consideró a la deforestación como una variable binaria a predecir de clases 0 y 1, en donde la clase 1 indica la ocurrencia de deforestación. La principal ventaja de contar con los datos en forma de imagen de 11 bandas es que es posible realizar remuestreos o modificar el mecanismo de muestreo de puntos. En otras palabras, la máxima cantidad de datos de entrenamiento está determinada por los píxeles contenidos en la imagen, en este caso, todo el territorio peruano. De igual forma, la simulación de nuevos escenarios requiere únicamente de la incorporación de la información espacializada que se desee simular (i.e., nuevas carreteras) para generar un nuevo conjunto de datos de predicción. Esta representación geográfica de los datos permite visualizar la distribución de las predicciones y explicar el comportamiento de la potencial deforestación incluso para un público no especializado en el tema. Finalmente, la representación gráfica del muestreo estratificado y de la designación de datos para el entrenamiento y validación se aprecia en la Figura 3.6.

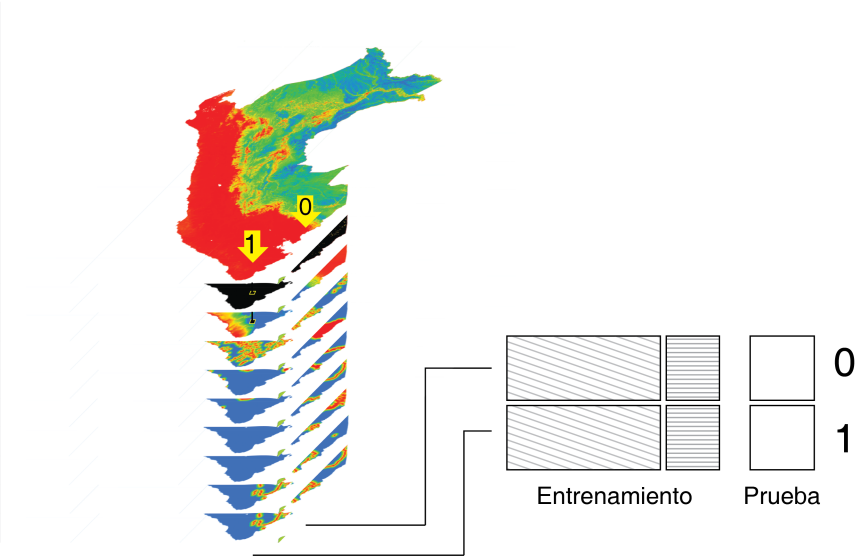

Figura 3.6: Esquematización del proceso de muestreo estratificado y designación de los grupos de entrenamiento y prueba. Se extrae la misma cantidad de muestras deforestadas y no deforestadas de una imagen multibanda.

Para cada cluster se construyó una zona de influencia de aproximadamente 50 km alrededor de todas las carreteras. Esto con la finalidad de reducir la computación innecesaria en zonas dónde no existe actividad ni influencia de las variables de interés. Finalmente, se muestrearon veinte mil puntos de la zona de influencia de cada cluster.

3.2. Construcción y validación de modelos de predicción

El desempeño de un modelo se mide a través de su error de clasificación. Este error tiene dos componentes principales que se pueden controlar durante la construcción del modelo: el sesgo y la varianza. El sesgo determina la precisión, o cuan cerca se encuentran las predicciones del valor real; y la varianza, la precisión del modelo cuando es probado con diferentes conjuntos de datos. Estos dos componentes suelen ser inversamente proporcionales, por lo que se espera que el modelo elegido tenga un adecuado equilibrio sesgo-varianza (Zhang and Ma, 2012). Alcanzar este equilibrio permite tener un modelo con una precisión adecuada y que al mismo tiempo sea lo suficientemente generalizable. En este capítulo se describirán las características y el proceso seguido en el entrenamiento y validación de los modelos, en los que se tomará la búsqueda del equilibrio sesgo-varianza como principal criterio de aceptación de un modelo. La heurística detrás de la búsqueda de los mejores parámetros será la misma en cada uno de los clústers seleccionados. A continuación se describirá brevemente las características fundamentales de los 4 modelos utilizados en esta investigación.

3.2.1. Regresión logística

El modelo de regresión logística es un modelo de clasificación que pertenece a la familia de los modelos lineales generalizados (GLM). Este tipo de modelo permite realizar clasificación de variables binarias o dicotómicas y es una extensión del modelo lineal convencional. La mencionada generalización se produce cuando el predictor lineal $\sum_{j=1}^{p} x_j\beta_j$, con p covariables x y parámetros β, es introducido en una función de enlace $g(\cdot)$. Esta función de enlace relaciona el predictor lineal con el valor esperado μ de la variable aleatoria a predecir Y (McCullagh and Nelder, 1989). En este sentido, el modelo generalizado tiene la forma mostrada en la ecuación 3.1

$$\mu = g(\sum_{j=1}^{p} x_j\beta_j) \tag{3.1}$$

En un modelo lineal convencional, la función de enlace corresponde a la función identidad, por lo que μ es igual al predictor lineal. Sin embargo, de acuerdo a la naturaleza de la variable de respuesta, esta función de enlace puede ser diferente (e.g., exponencial, sigmoide, entre otros) (Hosmer et al.). En el caso particular de los problemas de clasificación binaria se espera que μ sea entero y se encuentre en el rango de $[0, 1]$. En este caso, la función de enlace utilizada es la sigmoide, que permite tener un rango de $[0, 1]$ en un dominio entre $[-\infty, \infty]$. La ecuación 3.2 muestra la expresión análitica de la regresión logística.

$$\mu = \frac{1}{1 + e^{-\sum_{j=1}^{p} x_j\beta_j}} \tag{3.2}$$

Como se puede observar en el plano cartesiano (ver figura 3.7), la función sigmoide es continua, monótona, y asintótica. Esto quiere decir que el rango de esta función tiende a 0 y 1 cuando el predictor lineal tiende a $-\infty$ y ∞, respectivamente. Debido a esta condición, el valor estimado nunca alcanzará los valores extremos del rango ni será entero. En este sentido, en un problema de clasificación, el resultado de esta función puede interpretarse como la probabilidad que tiene una observación de pertenecer a determinada clase, donde un valor cercano a 1 es un indicador de alta probabilidad. Cuando se desea evaluar el desempeño de la clasificación, se considera que una observación es de una clase solo si supera un determinado umbral. En este caso, es usual tomar 0.5 como umbral de clasificación, aunque este valor podría ajustarse dependiendo del desempeño del modelo.

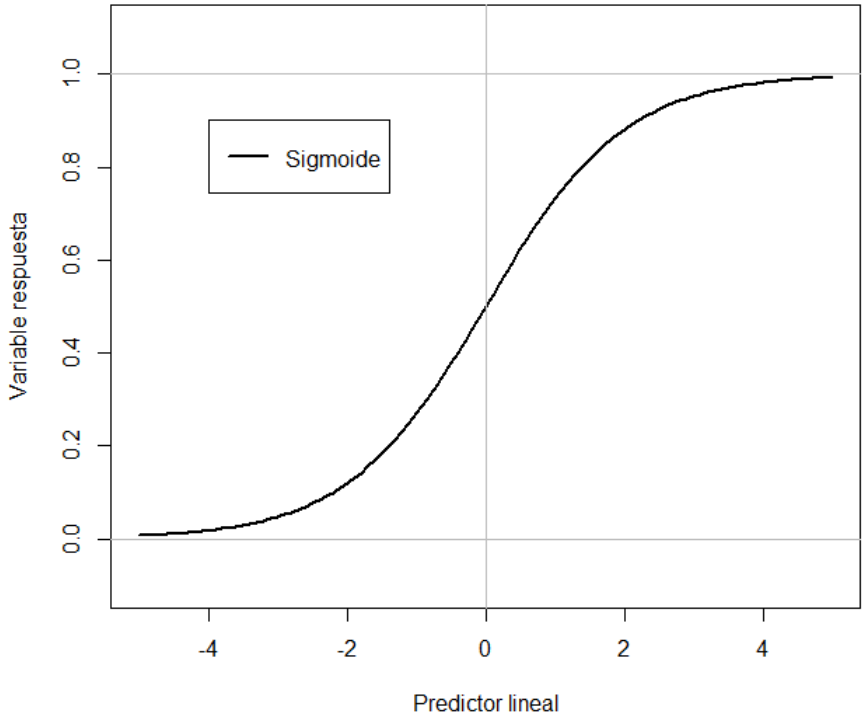

Figura 3.7: Representación gráfica de una función logística en el plano cartesiano.

La estimación de los parámetros más adecuados del modelo se suele realizar mediante el método de máxima verosimilitud. Con este método se buscan los parámetros que maximicen una función de verosimilitud construida con los datos observados y se suele requerir del uso de herramientas computacionales durante su resolución. Un descripción clara y didáctica de las propiedades de la regresión logística y los algoritmos de estimación de parámetros pueden encontrarse en Hosmer et al. y McCullagh and Nelder (1989).

3.2.2. Random forest

Random forest es un método de clasificación y regresión que resulta del ensamblaje de múltiples árboles de decisión. En este sentido, es importante primero comprender las propiedades de estos árboles. Los árboles de decisión son herramientas muy utilizadas en el aprendizaje automatizado y poseen una estructura jerárquica, similar a la de un árbol. El objetivo de este árbol es permitir la predicción de un resultado solo mediante el recorrido de los distintos niveles de jerarquía del árbol. Dicho de otra forma, la predicción se logra siguiendo el camino de las ramas, o divisiones, desde la raiz hasta los nodos u hojas finales. En la figura 3.8 se puede apreciar una representación gráfica de un árbol de decisión. En esta se puede observar que cada nodo se bifurca de acuerdo a determinado criterio de división. El primer nodo superior suele ser denominado *raiz*; mientras que los últimos nodos inferiores suelen llamarse *hojas*. Si se recorre el árbol desde la raiz a hasta las hojas se podrá determinar la clase que se desee predecir. Nótese que cada nivel del árbol construye un clasificador lineal. La combinación de estos múltiples clasificadores lineales permite obtener predicciones incluso en problemas de clasificación no separables linealmente. Adicionalmente, estos árboles son populares debido a que, por la naturaleza jerárquica de las divisiones, son robustos frente a los valores atípicos e insensibles a transformaciones monotónicas de las variables (Cutler et al., 2012)

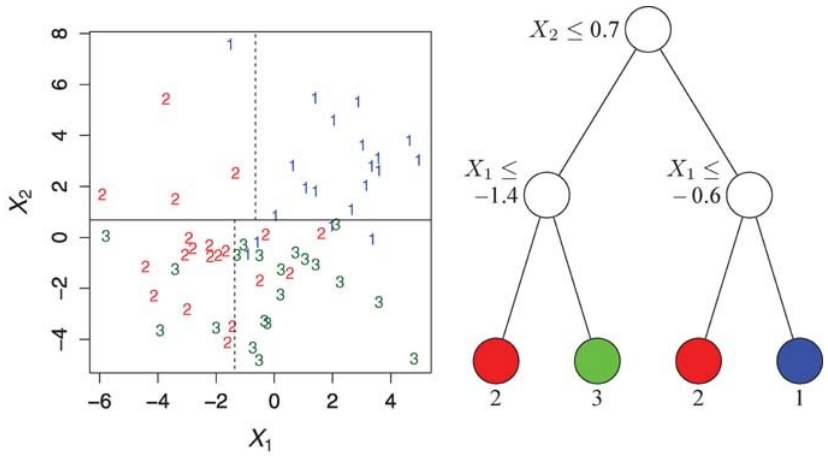

Figura 3.8: Ejemplo gráfico de un árbol de decisión. El árbol construido genera multiples separaciones binarias para determinar la clase a la cual pertenece el dato a predecir. Extraído de Loh (2011)

Como se puede notar, la clave en la construcción de un árbol de decisión se encuentra en selecciónar un adecuado criterio de divisón. Respecto a este último, existen dos métodos muy populares utilizados en la literatura: CART (Breiman, 1984) y C4.5(Salzberg, 1994). En el primer caso, se utiliza la impureza de Gini como criterio de división. Como se ve en la ecuación 3.3, para una variable E con N clases, la impureza se acercará a 1 mientras exista heterogeneidad en los datos; mientras que en el caso contrario, se acercará a 0 mientras los datos sean más homogéneos. El árbol comenzará la división con la variable que tenga la mayor impureza. En el segundo caso, la entropía es un indicador de la ganancia de información que se tiene al utilizar la variable E. Al igual que en el caso de la impureza de Gini, el árbol será construido siguiendo el orden de las variables con mayor ganancia de información. La ecuación 3.4 sirve para calcular la ganancia de información de una variable E con N clases.

$$Gini(E) = 1 - \sum_{n=1}^{N} p_n^2 \tag{3.3}$$

$$Entropia(E) = -\sum_{n}^{N} p_n \log p_n \tag{3.4}$$

Random forest es, en esencia, una extensión del método *baggings* propuesto por Breiman (1996). *baggings* es la abreviación de *bootstrap aggregation* y consiste en la construcción de multiples árboles de decisión entrenados con sub-muestras del conjunto de entrenamiento obtenidos mediante *bootstraping* (i.e., muestreo aleatorio con repeticiones). La predicción resulta de la combinación (e.g., votación o promedio) de las predicciones realizadas por los distintos árboles. Random forest, propuesto por Breiman (2001), explota el concepto de *baggings* al construir los árboles con muestras *bootstrap* y seleccionando un sub-espacio aleatorio de variables para cada árbol. Los árboles entrenados tienen poco sesgo pero alta varianza. Sin embargo, el poder de este método radica en que la combinación de estos árboles genera una clasificación con bajo sesgo y poca varianza. El ensamble de árboles de decisión no tiene, necesariamente, un mejor desempeño que el mejor árbol dentro del bosque de clasificadores; sin embargo, al promediar los resultados de muchos árboles se reduce la probabilidad de tener un solo árbol con poco desempeño (Zhang and Ma, 2012).

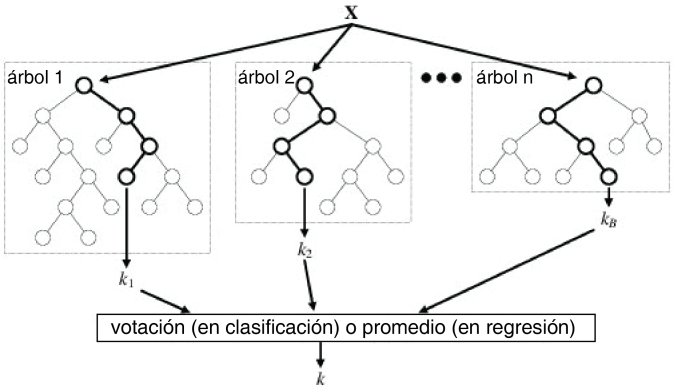

Figura 3.9: Representación gráfica de un modelo de random forest. Adaptado de Verikas et al. (2016)

3.2.3. Redes Neuronales Artificales

Las redes neuronales artificiales (ANN, por sus siglas en inglés) son configuraciones de operaciones secuenciales que permiten computar información utilizando una arquitectura inspirada en el funcionamiento cerebral del ser humano. El objetivo del uso de estas arquitecturas es modelar fenómenos mediante el entrenamiento o calibración de la misma a partir de observaciones de la realidad. Este tipo de modelo goza en la actualidad de mucha popularidad debido a su capacidad de modelar fenómenos con comportamientos no lineales. Sin embargo, pese a este rampante aumento de interés, las ANN ya eran estudiadas desde inicios de los años 50. Es en McCulloch and Pitts (1943) donde se propone por primera vez una estructura de resolución lógica similar al comportamiento sináptico del cerebro humano. Aunque existió un entendible revuelo al respecto en aquellos años, estos prometedores modelos fueron siendo dejados de lado debido a su restringida capacidad para entrenarse o resolver determinados problemas. No fue sino hasta la primera década del 2000 en el que el uso de las ANN se popularizó debido a los sorprendentes resultados obtenidos en los campos de la visión computacional, el reconocimiento de voz, la generación de textos, entre otros (Alom et al., 2018). Desde entonces, las ANN han ido evolucionando en su estructura y en su capacidad de abstraer fenómenos de la realidad. De acuerdo a Bebis and Georgiopoulos (1994), las ANN se pueden definir, formalmente, como un sistema que mapea una función no lineal $\hat{y} = G(x)$. Esta función se construye durante la etapa de entrenamiento y vincula los datos de entrada x con los datos de salida y mediante distintos parámetros. La hipótesis fundamental de este tipo de modelos es que existen, en teoría, infinitas configuraciones de red que podrían mapear los valores de x a y. Encontrar la configuración adecuada que determine este mapeo es la tarea que otorga sentido al arte de la construcción de modelos ANN (Bebis and Georgiopoulos, 1994).

Una ANN tiene tres elementos fundamentales en su estructura: una capa de entrada, una o más capas ocultas, y una capa de salida. Cada una de estas capas, a su vez, cuenta con elementos fundamentales denominados *neuronas*. Utilizando el símil propuesto por McCulloch and Pitts (1943), se puede decir que cada *neurona* recibe información de la combinación de la información de otras *neuronas* multipicadas por un determinado *peso*, referido también como la fuerza sinapsis. Para que esta neurona receptora envíe información a la siguiente neurona se requiere superar cierta valla. A esta valla requerida para continuar con la sinapsis se le denomina *activación* y determina si la neurona permitirá que el flujo de información continue (McCulloch and Pitts, 1943; Krogh, 2008). En la Figura 3.10 se representa la arquitectura más básica de una ANN, denominada perceptrón simple. En la primera capa del perceptrón las neuronas reciben los datos de entrada x_i y son multiplicados por los pesos w_i uno a uno, donde i es el número de variables. Como se ve en la figura, el grosor

de las líneas representa el valor de w y refleja la importancia que la neurona receptora otorga a esa conexión. En la neurona receptora se realiza la sumatoria $\sum x_i w_i$ para finalmente ser introducida dentro de una función $g(f(x))$denominada *función de activación* (función sigmoidal para el ejemplo) que define cual será la salida final de la neurona. En el caso del perceptrón, la salida de la neurona final representa el valor estimado ŷ. A partir del entendimiento de esta arquitectura sencilla es posible construir arquitecturas mucho más complejas debido que, en escencia, toda neurona tendrá entradas y salidas de la forma expuesta anteriormente. Aunque esta comparación es didáctica, lo cierto es que esta representación no está ni cerca de la complejidad del funcionamiento del cerebro humano; por lo que su uso es una atribución tomada por el autor de esta tesis.

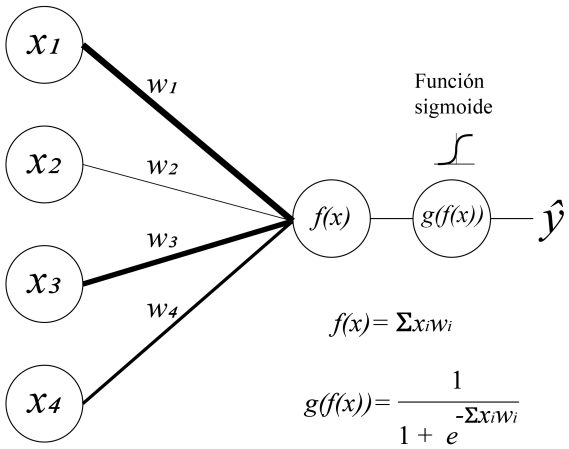

Figura 3.10: Representación gráfica del perceptron simple

La arquitectura de ANN profunda suele tener mucho más de una capa profunda y miles de parámetros. Una arquitectura de ANN profunda de un solo sentido (*feedforward*) puede verse en la Figura 3.11. La configuración que se designe a la red depende del tipo de problema planteado, de las características de los datos y de la respuesta que se esté buscando. En Van Veen (2016) puede encontrarse una clara descripción del desarrollo de las diversas arquitecturas de redes encontradas en la literatura.

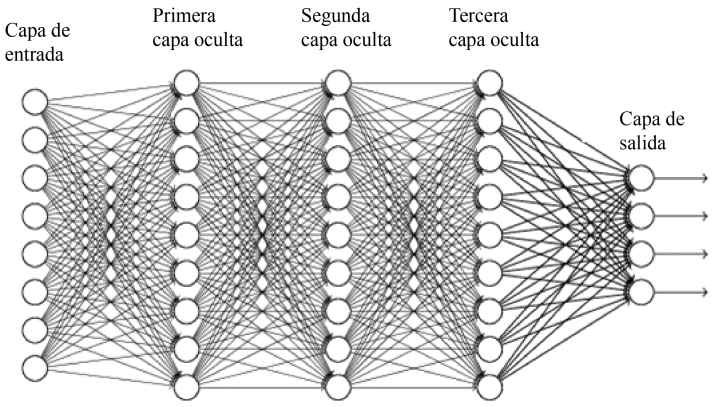

Figura 3.11: Arquitectura de una red ANN profunda en un solo sentido

Cual fuese la configuración elegida en la construcción de una red, la estimación de los mejores parámetros se realiza mediante la optimización de una función de costo. Esta función depende de los

parámetros y se utiliza para penalizar los errores de predicción del conjunto de datos. La mayoría de problemas de clasificación, hoy en día, entrenan los modelos con el método de máxima verosimilitud, por lo que la función de costo es el negativo de la log-verosimilitud (Goodfellow and Bengio, 2016) Esto es equivalente a utilizar la entropía cruzada de los datos de entrenamiento y los datos estimados. En la Ecuación 3.5, N indica el conjunto de datos, n es una de las observaciones; y k, la clase correspondiente a esa observación.

$$E_{\text{entropy}} = -\sum_{n}^{N} [t_k^n \ln y_k^n + (1 - t_k^n) \ln(1 - y_k^n)] \tag{3.5}$$

Finalmente, esta función de costo es optimizada, en practicamente la mayoría de casos, utilizando un método llamado descenso de gradiente. Con este método se calculan las gradientes de la función de costo con respecto a cada parámetro del modelo. Posteriormente, cada parámetro es corregido sumandole su gradiente correspondiente escalada por un valor denominado *tasa de aprendizaje*. Este proceso iterativo busca, eventualmente, modificar los parámetros en dirección opuesta a la gradiente a fin de minimizar la función de costo. El cálculo de las gradientes se realiza utilizando un algoritmo denominado *retropropagación* y puede ser consultado en Hecht-Nielsen (1989).

3.3. Estimación de emisiones de GEI

La estimación de las emisiones de gases de efecto invernadero se realizó utilizando los resultados obtenidos en los distintos modelos de regresión. Se estimó un nivel de corte óptimo para el mejor modelo. A partir de este valor se consideró si un pixel sería deforestado o no. Finalmente, se sumaron todos los valores de densidad de carbono correspondientes a cada píxel deforestado. La transformación del carbono total a CO_2eq se realizó multiplicando el valor total por 3.66, factor de transformación obtenido a partir del peso molecular. La ecuación 3.6 muestra la operación de cálculo de emisiones de GEI, donde $\boldsymbol{N}$ es el número de píxeles, $\boldsymbol{c}$ el contenido de carbono del píxel $\boldsymbol{n}$, y $\boldsymbol{p}$ la probabilidad de deforestación del píxel $\boldsymbol{n}$.

$$CO_2 = 3.66 \sum_{n}^{N} [c_n p_n] \tag{3.6}$$

3.4. Implementación del sistema de trabajo en la nube

La tangibilización del sistema de trabajo propuesto se realizó utilizando un proceso de adquisión y análisis de datos basado completamente en la nube. Este sistema utilizó la plataforma GEE para realizar los cálculos y los preprocesamientos requeridos, así como la importación y exportación de los mapas ráster en cada etapa del proyecto. Esta plataforma permite programar en el lenguaje javascript las distintas operaciones de forma declarativa, estas son posteriormente enviadas a los servidores de GEE para su cálculo y este último devuelve los resultados para ser visualizados (ver figura 3.12). El muestreo fue también realizado en esta plataforma y exportado a un *bucket* alojado en los servidores de Google Cloud Plataform. Desde este lugar, la información es llamada desde cualquier entorno de python siquiendo su dirección en la nube (i.e., 'gs//: ....'). En este sentido, los datos pudieron ser utilizados al mismo tiempo desde diferentes plataformas. Se utilizó Google Colab, un entorno *ijupyter* en la nube, como cuaderno de python3 para experimentar con los datos muestreados y se utilizaron tarjetas gráficas (GPUs) para acelerar los cálculos. La búsqueda de los hiperparámetros se realizó utilizando herramientas bastante exploradas en la literatura (Schratz et al., 2018). En este caso, se utilizó un método denominado búsqueda de grilla e iteraciones implementadas en el script de cómputo. Se utilizó la librería Scikit-learn (Pedregosa et al., 2011) para la experimentación con los modelos RF, NB y LR. Los modelos de redes neuronales fueron implementados utilizando las librerías Tensorflow y Keras. Las tres librerías están implementadas en python y fue en este lenguaje en el cual se realizó el análisis.

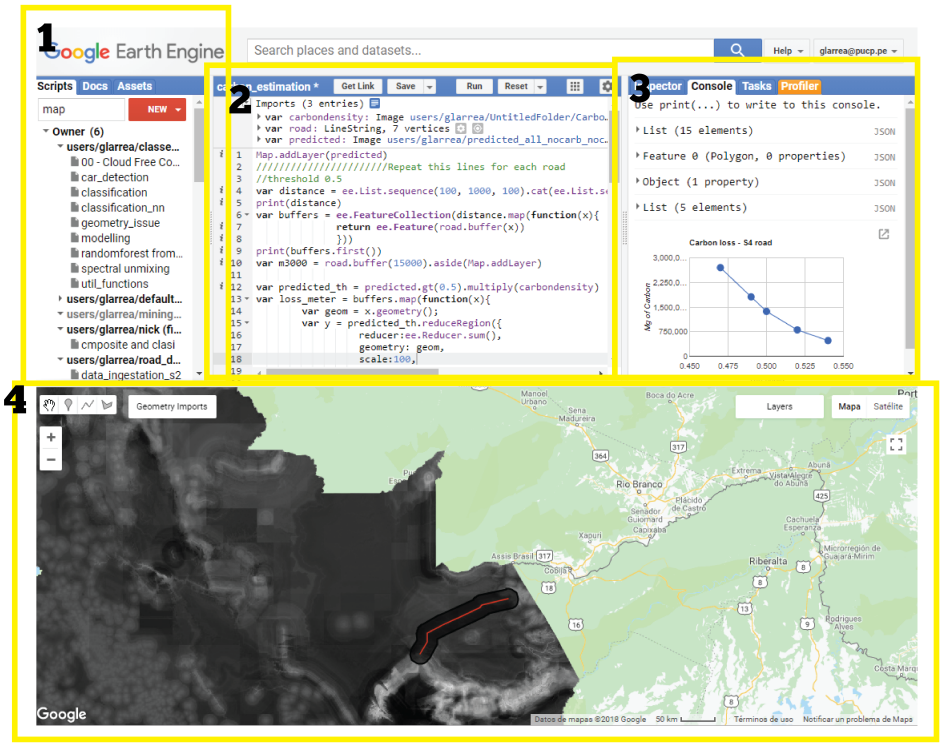

Figura 3.12: Captura de pantalla de el interfaz gráfico de Earth Engine. El recuadro 1 muestra el repositorio y la documentación. 2 muestra el cuaderno de trabajo. 3 muestra la consola donde se exhiben resultados numéricos y se realiza la depuración. 4 muestra la pantalla de visualización

Una vez estimados los hiperparámetros adecuados, se regreso a Earth Engine para realizar la clasificación de los modelos pero escalados a todo el territorio nacional. Las imágenes clasificadas se almacenaron en la nube en formato TIFF y pueden ser visualizadas con cualquier explorador web. Para facilitar la difusión de los resultados, se implementó una aplicación de javascript que utilizá el API de Earth Engine. Esta aplicación permite visualizar los resultados desde cualquier explorador o aparato móvil sin necesidad detener una cuenta de Earth Engine o pagar por un servicio de Cloud Plataform. Finalmente, todo el flujo descrito en esta sección (ver Figura 3.13) fue utilizado en cada iteración de los distintos modelos a fin de que estos sean afinados.

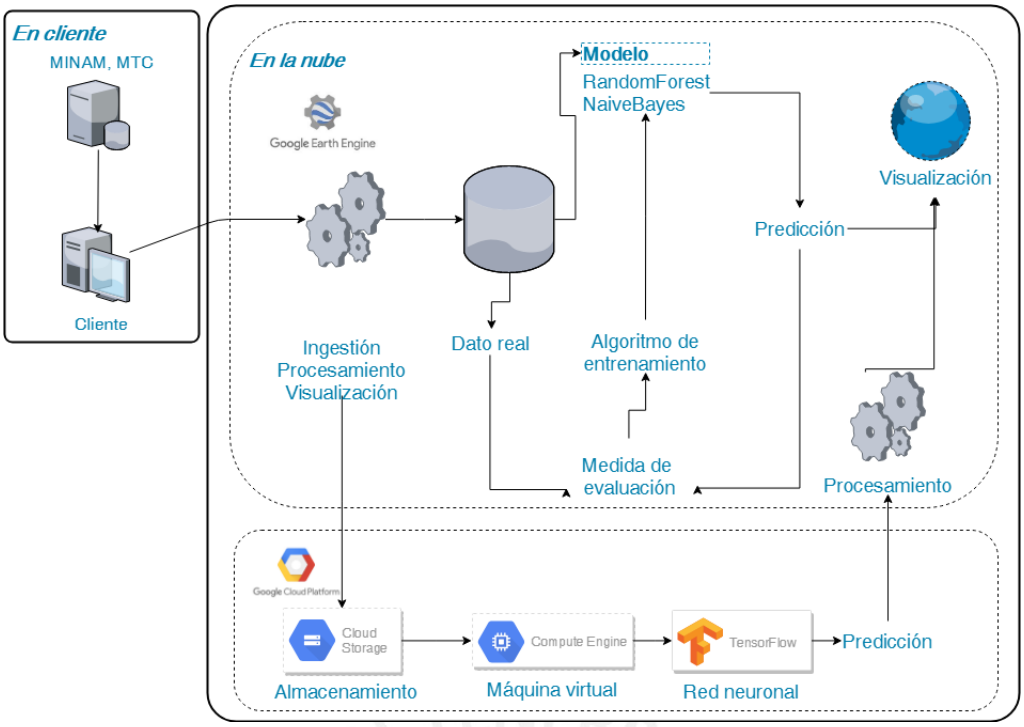

Figura 3.13: Flujo operacional en la nube seguido durante la investigación

Finalmente, en la Tabla 3.4 se muestran los modelos entrenados y la descripción de sus características.

Tabla 3.4: Descripción de los modelos analizados en este estudio

Modelo	**Código**	**Descripción**
Random forest	RF	Utiliza impureza de gini y 120 árboles.
Naive Bayes	NB	–
Regresión Logística	LR	Utiliza regularización l2, con parámetro $B = 0.01$
Red Neuronal	NN_{256}	Red con 6 capas ocultas de 256, 128, 64, 32, 16 y 8 nodos. Utiliza datos normalizados y una capa dropout.
Red Neuronal	NN_{60}	Red con 4 capas ocultas de 60, 60, 30, y 8 nodos. Utiliza datos normalizados.

Capítulo 4

Resultados y discusión

4.0.1. Análisis de datos

Siguiendo el paradigma del análisis de datos mencionado en el Capítulo 1, se realizó un análisis explorativo de los datos utilizados en el modelo. Todas variables son númericas y continuas, a excepción de la variable a predecir, que es binaria. Para entender la relación que existe entre las variables se construyó una matriz de correlaciones (ver figura 4.1) con todas las variables predictoras. Esta matriz utiliza el coeficiente de correlación de Pearson para determinar el grado de correlación que existe entre cada par de variables. La correlación (i.e., relación lineal) es positiva cuando el valor del coeficiente se aproxima a 1; y negativa, cuando el valor se aproxima -1. Un coeficiente de 0 se interpreta como correlación nula. En la Figura 4.1 se aprecia que no existe una correlación importante entre ninguna de las variables, a excepción de las variables *Distance a ZA* y *Distancia a ANP*. Esta correlación puede deberse a que la existencia de una ZA está condicionada a la existencia de una ANP. Dicho de otro modo, una ZA es creada con la finalidad de proteger la integridad de una ANP, por lo que todo píxel muestreado que se encuentre cerca a una ANP se encuentra, estrictamente, también cerca a una ZA. Otro caso particular es el que se observa con las variables *latitud* y *longitud*. En este caso, la correlación es negativa (i.e., -0.53) y se debe a que los pares de coordenadas se encuentran entre un rango limitado de valores debido a que corresponden a muestras de zonas específicas. Esta bajas correlaciones, sin embargo, no representan problemas para la construcción de los modelos. Por el contrario, suelen ser de utilidad al momento de implementar modelos estadísticos paramétricos (i.e., regresión logística).

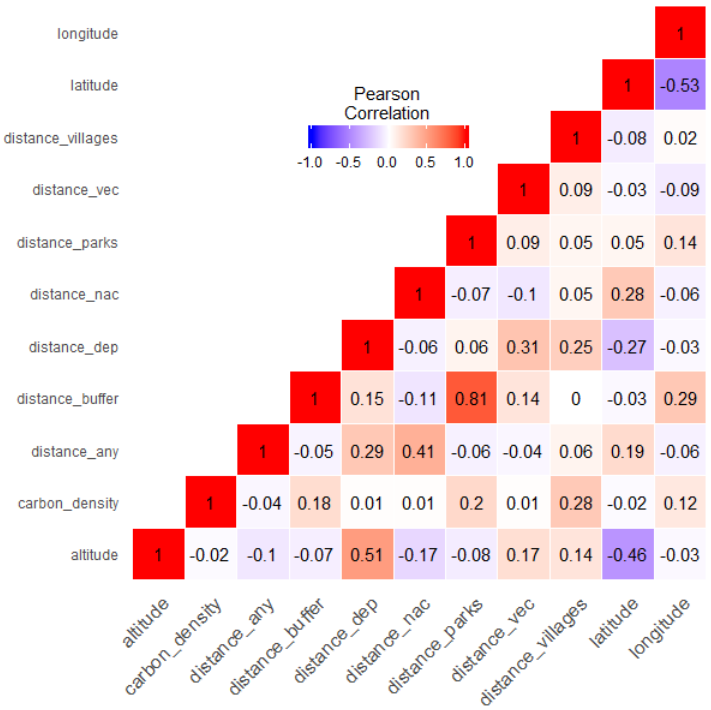

Figura 4.1: Matriz de correlaciones entre variables. Se utiliza el índice de correlación de Pearson para determinar el grado de correlación existente entre las variables.

Para entender el comportamiento individual de cada variable con respecto al fenómeno de interés, se construyeron histogramas (ver Figura 4.2) de ocurrencia de deforestación a partir de las variables predictoras más relevantes para el clúster C1. Debido a que los datos eran numéricos y continuos, los histogramas fueron elaborados a partir de intervalos de distancias. Es decir, se cuantificó la cantidad de píxeles deforestados en cada intervalo de distancia, desde 0 hasta 25000 metros, obteniendose un valor que representa la tasa de deforestación para cada intervalo. Respecto a las tendencias de los histogramas, se puede apreciar que la deforestación disminuye considerablemente a medida que la distancia al punto de referencia aumenta. Esta disminución puede ser lineal, en el caso de *Distancia a carretera nacional* y *Distancia a ZA*, o exponencial, en el caso de *Distancia a carretera vecinal*. En el caso de las variables restantes, se puede apreciar que la disminución tiene una tendencia errática. Por ejemplo, en *Distancia a ANP* la tasa de deforestación en los primeros 25000 metros parece ser constante; mientras que en *Distancia a centro poblado*, esta tasa aumenta en los primeros 5000 metros para luego disminuir considerablemente. Una posible interpretación del comportamiento no monotónico de *Distancia a centro poblado* es que las zonas muy cercanas a los centros poblados ya atravesaron por un proceso de urbanización y consecuente cambio de uso de suelos. Sin embargo, los alrededores de estos centros urbanos (e.g., 5000 metros) suelen ser zonas destinadas a la expansión urbana o la práctica de agricultura y ganadería. En lo que respecta a las magnitudes, se puede apreciar que la mayor ocurrencia de deforestación se da en *Distancia a carretera vecinal*, seguida de *Distancia a carretera nacional* y *Distancia a centro poblado*. En el caso de *Distancia a ANP* y *Distancia a ZA*, los valores de deforestación son, relativamente, mucho menores que en el resto de casos. Este hecho permite inferir que las zonas destinadas a la protección de biodiversidad sí cumplen la función para la cual fueron diseñadas. Esta inferencia es consecuente con el discurso expuesto en Miranda et al. (2014).

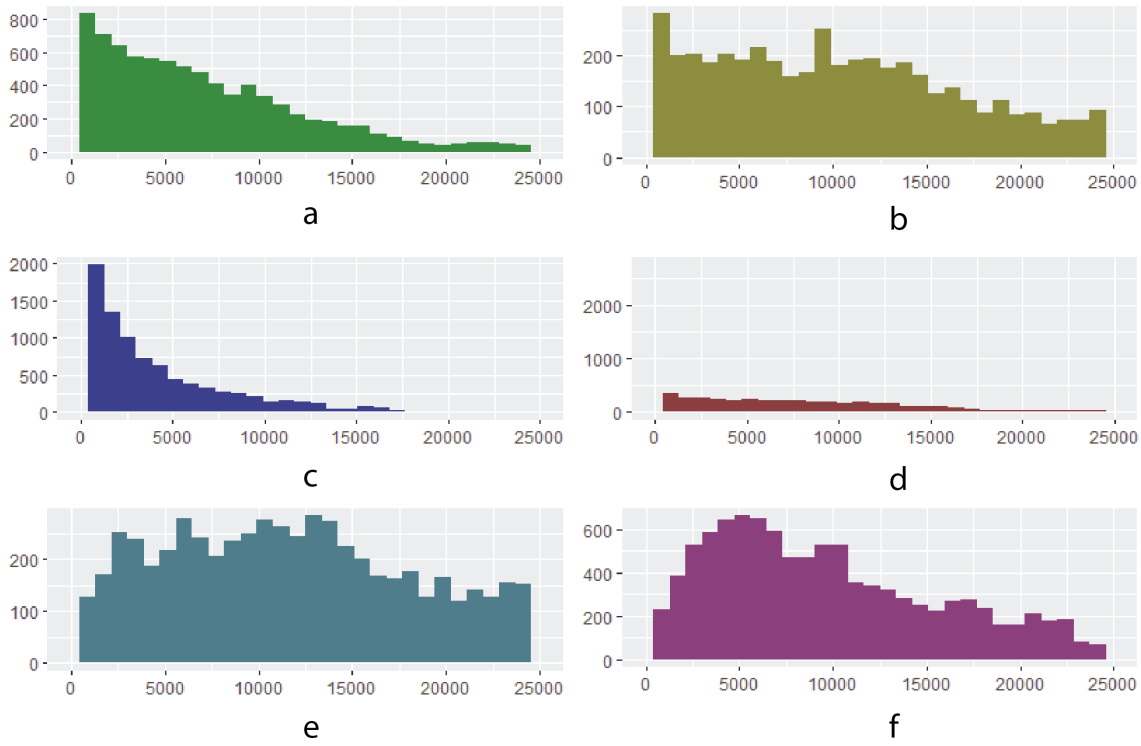

Figura 4.2: Histograma de ocurrencia de deforestación - Distancia (m) para distancias a carretera nacional (a), departamental (b), vecinal (c), zona de amortiguamiento (d), Área Natural Protegida (e) y centro poblado (f)

Adicionalmente, se analizó el comportamiento de la deforestación en los diferestes clústers propuestos. Al explorar la variable *Distancia a carretera nacional* (ver Figura 4.3) se observó que, dependiendo del cluster analizado, la tendencia de la deforestación puede ser completamente diferente. Por ejemplo, en los clústers C1, C6 y CT es posible observar una tendencia decreciente lineal. Sin embargo, en los clusters C3, C4 y C8 el comportamiento es errático y la deforestación no sigue el patrón esperado. Un análisis visual de los clusters (ver Figura 4.4) permitió determinar que, en el caso del clúster C8, el patrón de deforestación es bástante difuso y disperso. Esto se debe, principalmente, a que está zona experimentó un importante crecimiento agrícola en los últimos años, tanto de palma aceitera como de agricultura a pequeña escala (Vijay et al., 2018).

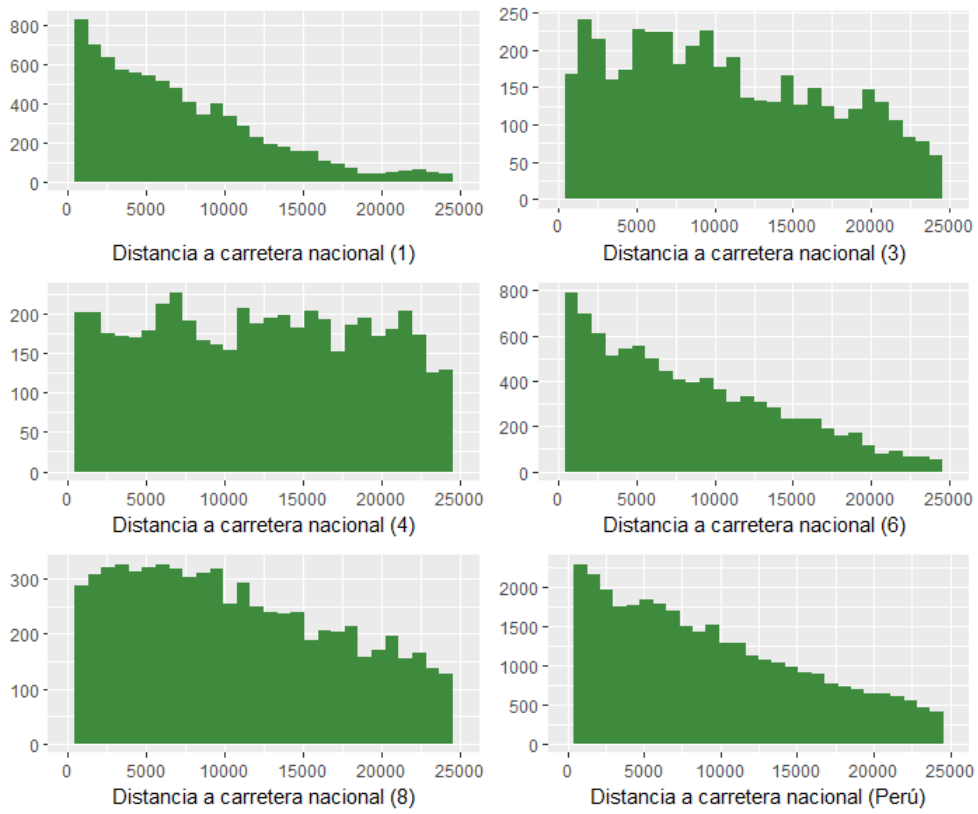

Figura 4.3: Histograma de ocurrencia de deforestación - Distancia a carretera nacional (m) para cada cluster analizado

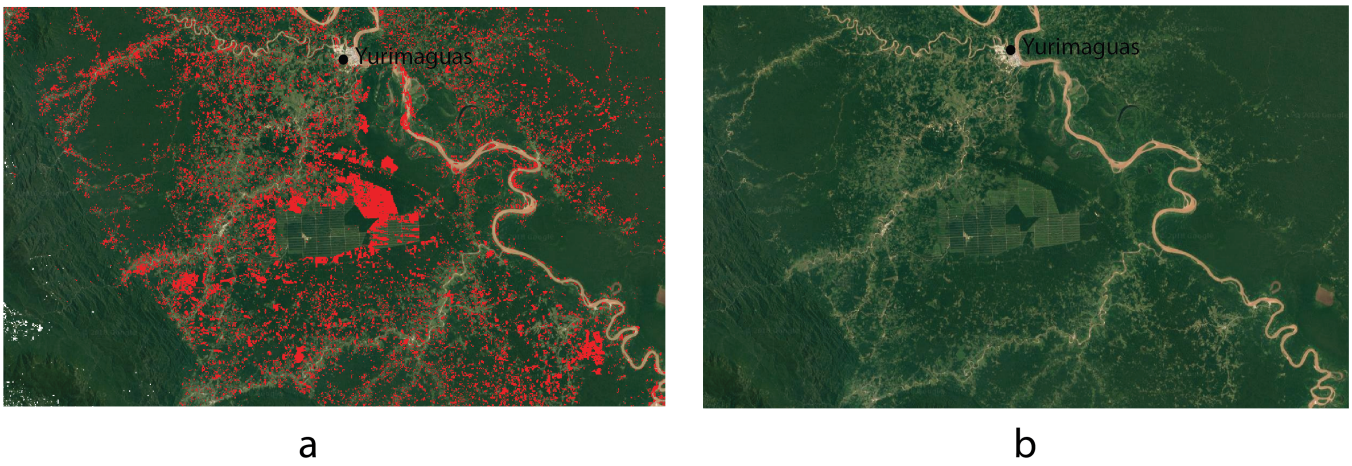

Figura 4.4: Imágen satelital de la deforestación ocurrida en el cluster 8, en los alrededores de Yurimaguas. Comparación entre los píxeles deforestados utilizados de datos (a) y las imágenes satelitáles

4.0.2. Búsqueda de hiperparámetros

En cada iteración mencionada en el capítulo anterior se ejecutó una búsqueda de grilla para determinar la mejor combinación de hiperparámetros. En el caso de RF, se determinó la cantidad de árboles óptima, la profundidad de los árboles y el criterio de división de cada nodo. La figura 4.5 muestra la variación en la precisión de acuerdo al número de árboles utilizado en cada iteración de la búsqueda de grilla, y, a su vez, en cada cluster analizado. Como se puede observar, en todos los casos, la precisión crece rápidamente a medida que se aumenta el número de árboles. Este crecimiento se

detiene cuando el modelo utiliza 50 árboles y se observa que, a partir de este número, la precisión se estabiliza. Respecto a los valores finales de precisión una vez alcanzada la estabilidad, se puede decir que en 5 de los 6 casos mostrados, la precisión se encuentra en el rango de 0.775 a 0.825. De hecho, cuando se toman 120 árboles, los clústers C4,C6 y C1 tienen precisiones muy parecidas. Otro aspecto que se puede notar es que todos los clusters, excepto el clúster C8 mantienen su ranking de precisión independientemente del número de árboles utilizados. En contraste, el cluster C8 presenta precisiones inferiores en todo el experimento y en ningún caso llega a superar el valor de 0.700.

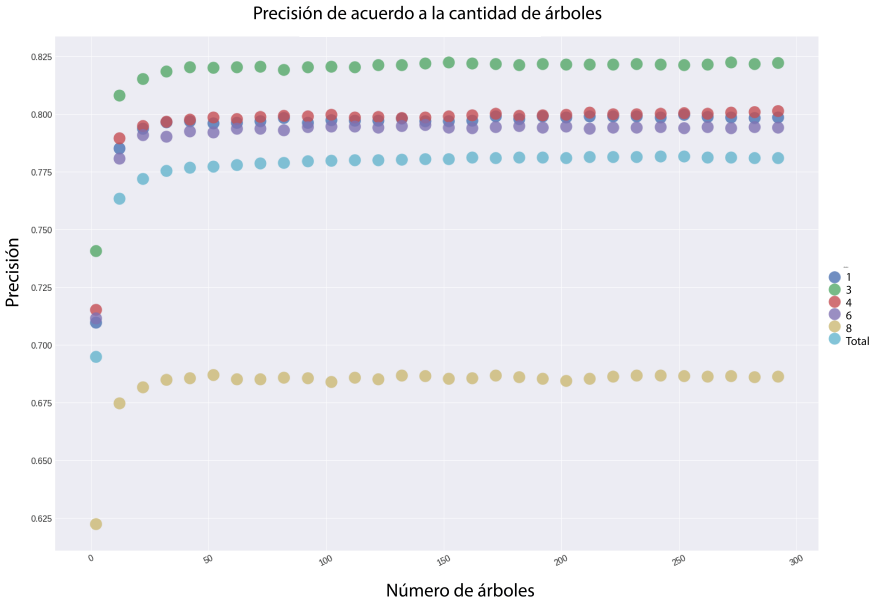

Figura 4.5: Variación en la precisión de acuerdo al número de árboles para cada cluster

Para las ANN, se experimentó con las distintas arquitecturas detalladas en el capítulo 3 y se analizó la evolución de la pérdida en los datos de entrenamiento y validación a lo largo de las distintas épocas de entrenamiento. En cada caso, se fue incrementando la complejidad de la arquitectura y se fueron añadiendo los criterios de regularización, normalización en bloque y *droput*. La figura 4.6 muestra un típico caso de sobre entrenamiento observado en ANN256-5, en esta iteración se estudió la capacidad predictiva del modelo cuando se utilizó una arquitectura relativamente sencilla pero sin considerar regularizadores y *dropout*. Como se observa, en todos los casos, a exepción del cluster T, pérdida del conjunto de validación cambia de sentido y comienza a aumentar. Este aumento se refleja en un precisión de validación que se estanca en un rango, mientras que la precisión de entrenamiento sigue en aumento. Esta característica es propia de un modelo con sobre entrenamiento. Sin embargo, si se observa el caso del cluster T, se puede ver que la pérdida sigue descendiendo aún después de la época 500. Aunque este descenso se puede considerar casi como una estabilización, el modelo es estable aún hasta después de las 500 épocas. Una explicación a este fenómeno puede darse al recordar que este conjunto de datos cuenta con 100000 observaciones, resultado de la combinación de los datos del resto de clusters. Este último hecho demuestra que el aumento de la cantidad de datos es una efectiva estrategia para reducir el sobre entrenamiento.

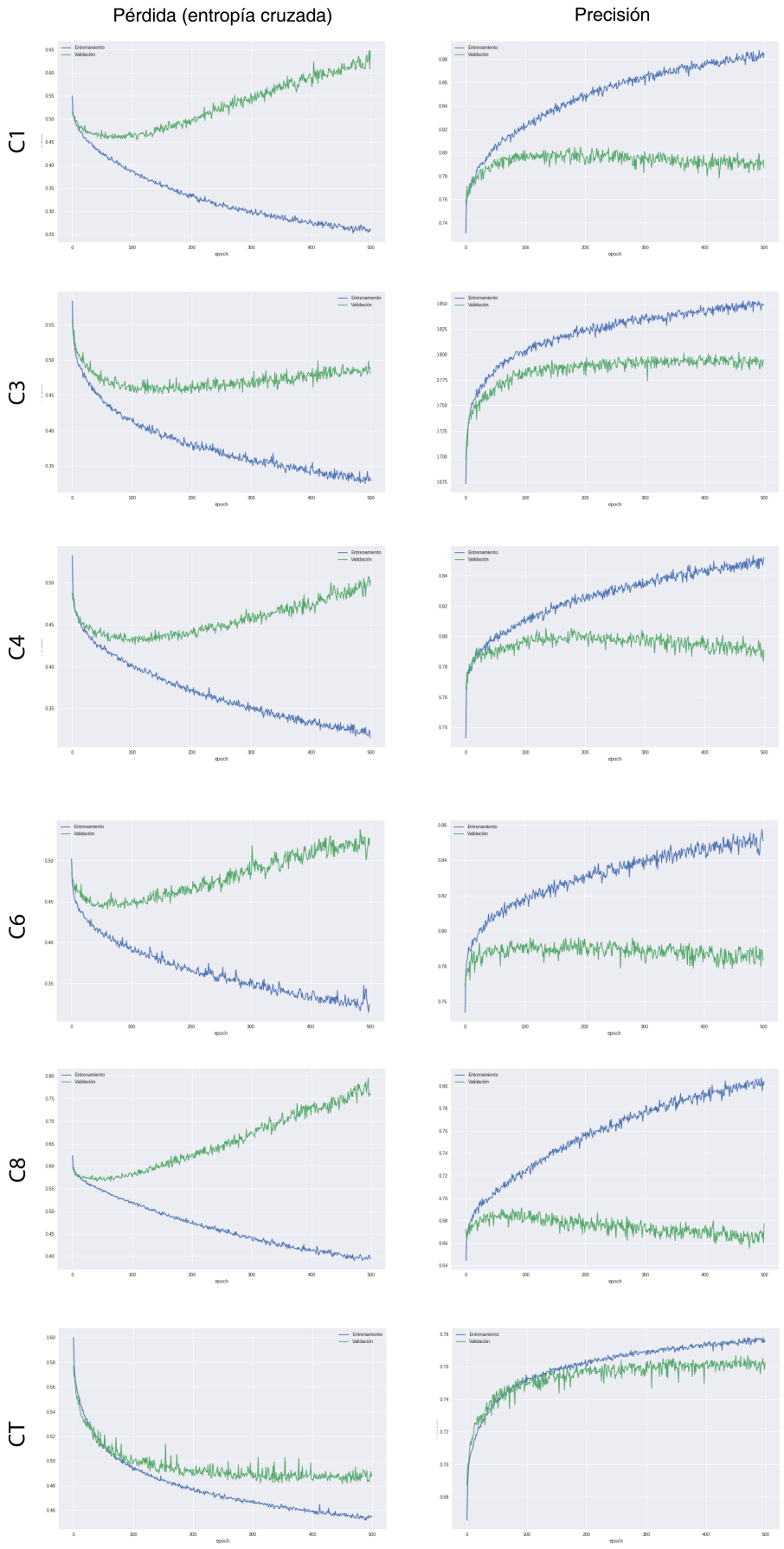

Figura 4.6: Variación de la pérdida y la precisión del modelo de red neuronal a lo largo de las distintas épocas de entrenamiento

4.0.3. Importancia de las variables

Se determinó la importancia que pueden tener las variables en los modelos utilizados. Para esto, se utilizó el atributo *feature importance* del objeto random forest en Scikit-learn. Este atributo permite observar la importancia cada variable de acuerdo a su relevancia en la división de los nodos en cada nivel del árbol de decisión. La figura 4.7 muestra un ranking de las variables más importantes extraída de una de las iteraciones del modelo. En este caso se puede observar que, practicamente, todas variables son determinantes a la hora de dividir los nodos. Dicho de otra manera, es necesario contar con todas las variables estudiadas debido a que cada una de estas contribuye a la construcción de los árboles de decisión.

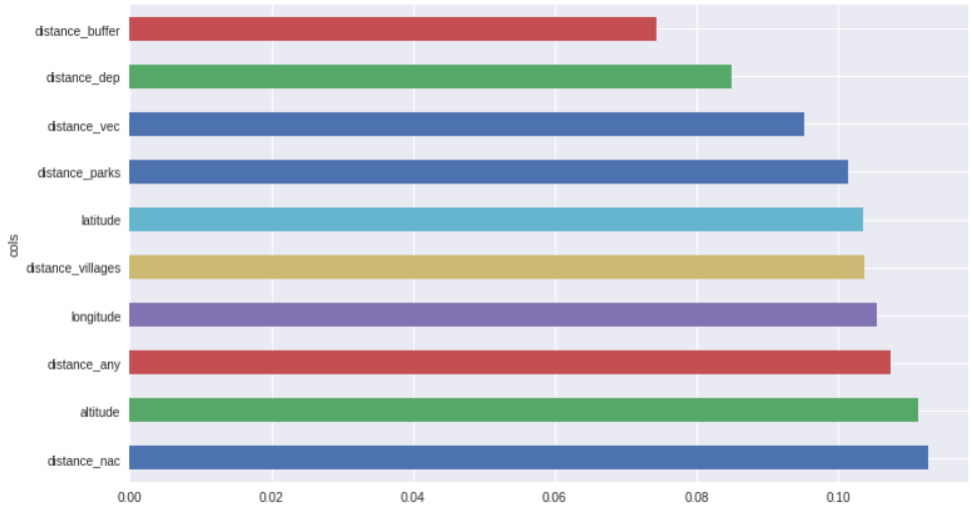

Figura 4.7: Importancia de las variables utilizadas en el modelo de random forest expresadas en porcentaje

4.0.4. Comparación entre modelos

Los diferentes modelos fueron comparados de acuerdo a la precisión obtenida por cada uno de estos utilizando los distintos datos de entrenamiento. Como muestra la figura 4.8, la precisión estuvo en el rango de 0.55 y 0.85. La figura muestra que no es posible determinar que modelo es complemente superior debido a dos razones fundamentales. La primera razón es que la dispersión de las precisiones es muy alta entre los diferentes clusters. La segunda razón es que en cada uno de los modelos, el ranking de precisiones no se mantiene constante. Por ejemplo, si se observa la ANN60-5, se puede ver que el modelo entrenado con el cluster 3 tiene menor que precisión que aquellos entrenados con los cluster 4,6 y 1; sin embargo, en RF, se observa que este comportamiento es practicamente inverso. La figura muestra también que en todos los casos, a excepción de NB, el cluster 8 es el que presenta las menores precisiones. Este comportamiento característico de los modelos entrenados en este cluster se debe a características propias de los datos de esta zona. En efecto, si se observa la figura 4.9 se puede notar que gran parte de los píxeles amarillos, correspondientes a deforestación, se encuentran agrupados en dos zonas en particular. Al realizar revisiones a las fotografías satelitales se pudo constatar que estas zonas corresponden a plantaciones de palma aceitera. Este hecho permite deducir que el modelo entrenado con los datos del cluster 8 recibe muchos datos con variables que no están relacionadas, en lo absoluto, con las características propias de las carreteras de la zona. En efecto, las plantaciones de palma aceitera son costosas, extensas y de muy largo plazo, por lo que su aparición responde a motivaciones económicas ajenas a la expansión vial.

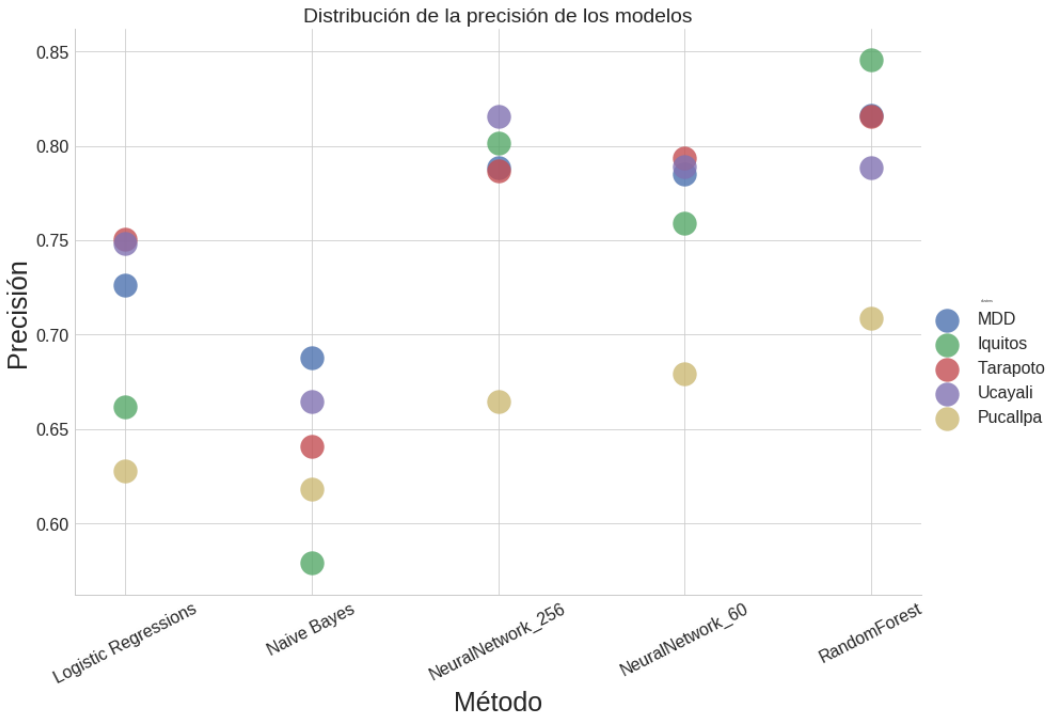

Figura 4.8: Distribución de la precisión de los distintos modelos entrenados con datos de los distintos clusters

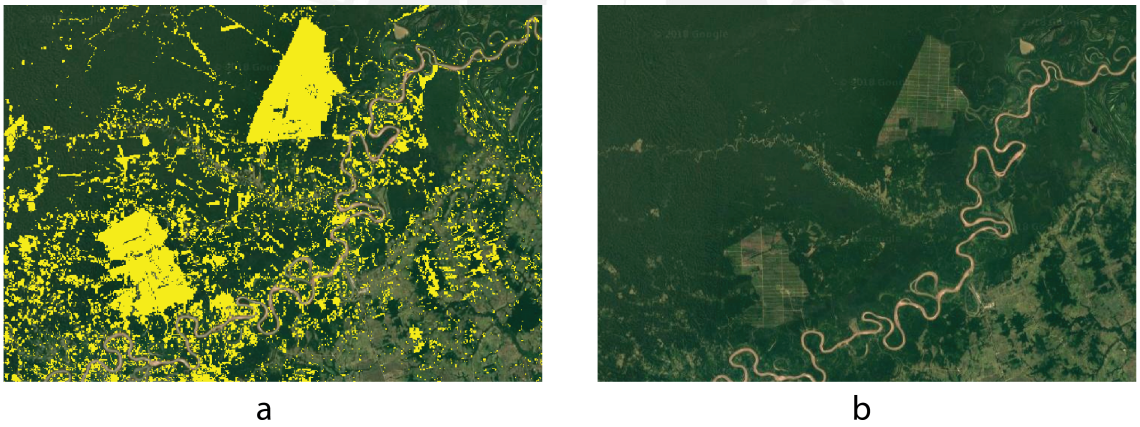

Figura 4.9: Imágen satelital que muestra que la zona deforestada (a) corresponde a una plantación de aceite de palma aceitera(b)

4.0.5. Visualización de resultados

Se seleccionó el mejor modelo RF como modelo principal para la construcción del mapa de riesgo de deforestación. Para ello, se realizó la predicción de cada uno de los píxeles del ROI de una nueva base de datos que incluía las carreteras proyectadas en el Plan Vial Nacional. La Figura 4.10 muestra el mapa de probabilidad de deforestación construido tomando los valores de probabilidad de cada predicción del modelo RF. Cada píxel tiene un valor en el rango de $[0, 1]$, donde 0 indica 0 % probabilidad; y 1, 100 % de probabilidad. Es importante mencionar que el modelo se entrena utilizando un valor de corte óptimo calculado a partir de una curva ROC (*Region Under the Curve*, por sus siglas en inglés); sin embargo, se utilizó la representación probabilistica de los resultados porque se desea hacer el análisis utilizando una lógica difusa.

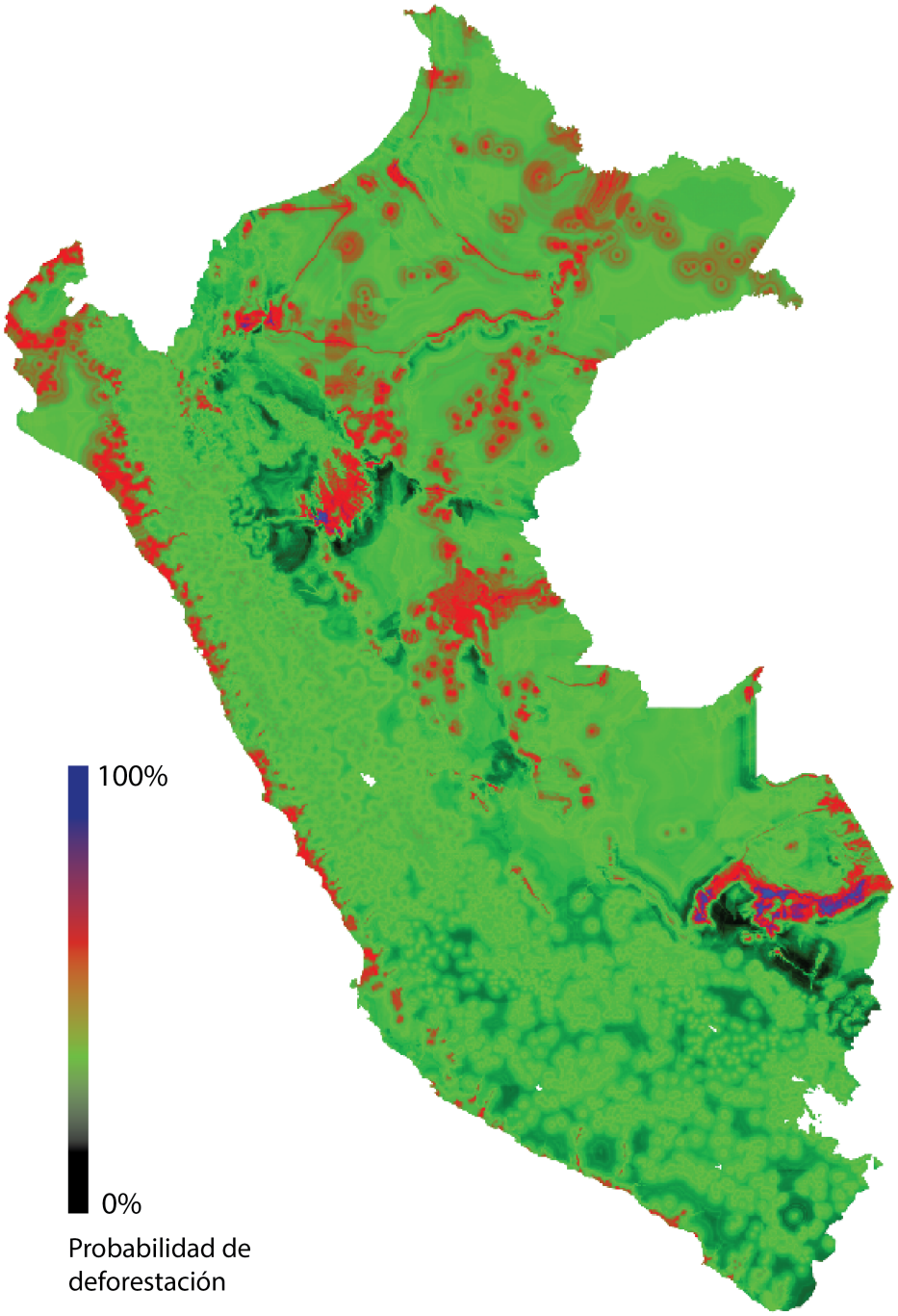

Figura 4.10: Mapa de probabilidad de deforestación construido con un modelo de random forest aplicado a carreteras proyectadas

Si se realiza un acercamiento al clúster C1, en particular a la carretera Boca Manu - Iberia (ver Figura 4.11), se puede observar que el riesgo de deforestación aumenta en la vía de la carretera. Se observa además que el color no es azul, por lo que el valor no está cerca a 1. Esto se puede interpretar como una baja probabilidad de deforestación si solo se evalua este proyecto vial. De igual forma, se debe notar que al insertar únicamente los datos de esta carretera nacional se asume que esta será la única carretera en los próximos 17 años. Esta suposición no es correcta debido a que se espera que esta vía principal permita el crecimiento de otras vías secundarias.

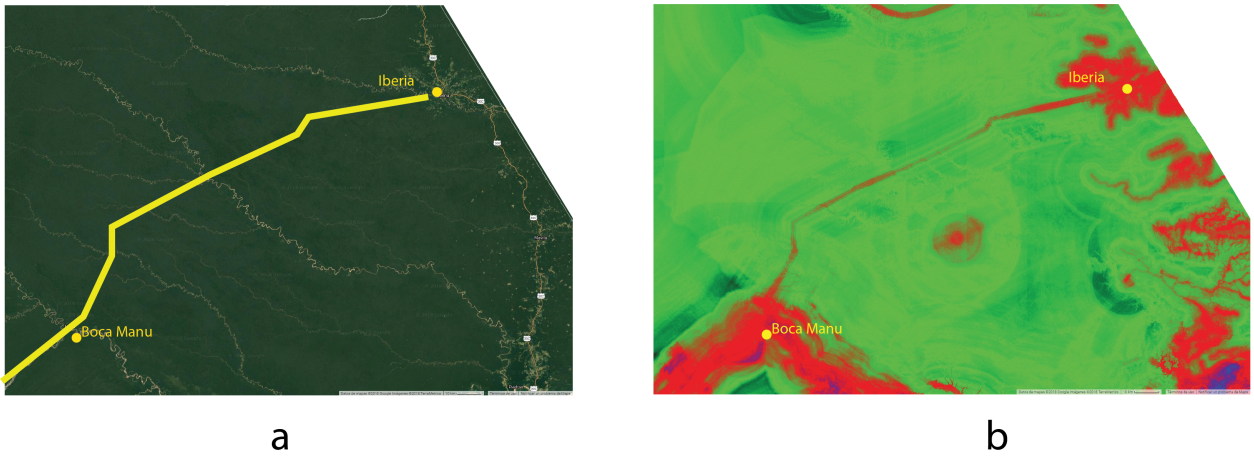

Figura 4.11: Mapa de probabilidad de deforestación (b) construido con los datos del proyecto de carretera Boca Manu - Iberia (a)

4.0.6. Cálculo de emisiones de carbono

Para obtener las emisiones de carbono se tomo como caso de estudio el proyecto MD-103, que conecta las comunidades de Salvación y Boca Manu. Para ello, primero se ingestó la información de la carretera en forma de imagen y se realizó la predicción. Luego, se selecciónó un criterio de corte a partir del cual un píxel se consideraría deforestado (i.e., 0.5). Finalmente, se multiplicó sumaron los valores de contenido de carbono correspondientes a cada píxel en los 30 kilómetros adyacentes a la carretera. La Figura 4.12 indica el procedimiento seguido para obtener la cantidad de carbono liberado.

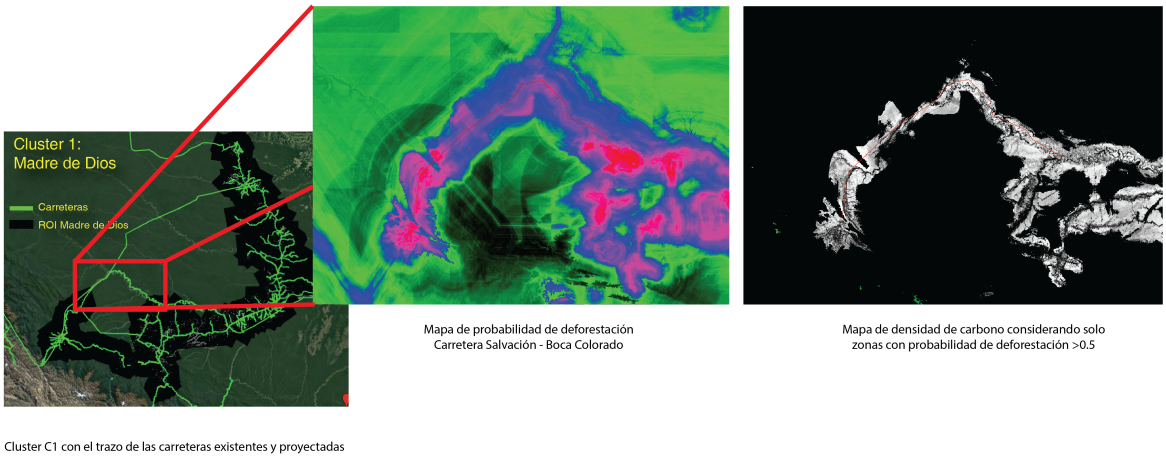

Figura 4.12: Proceso de cálculo de emisiones de CO_2 para el caso de estudio: Carretera MD-103

Adicionalmente, se construyeron dos gráficos (ver Figura 4.13) con información del carbono emitido a cierta distancia de la carretera. La Figura 4.13.a muestra el comportamiento de las emisiones en cada kilómetro de carretera. Como se puede observar, este valor disminuye considerablemente a medida que el píxel se encuentra más alejado de la carretera. Es importante mencionar que los cambios bruscos de pendiente se deben a que el contenido de carbono no es continuo a lo largo de la zona de estudio, por lo que la tasa de emisión dependerá de la zona en la que se realice el análisis. La Figura 4.13.b muestra las emisiones acumuladas en cada kilómetro de distancia a la carretera. Este valor se incrementa de

forma casi exponencial hasta estabilizarse en $2e^7 MgC$. Este valor puede ir aumentando si se toma una zona de afectación mayor a 30 kilómetros; sin embargo, se considera que esta distancia es prudente y que la emisión final no aumentará de manera significativa si aumenta los límites de análisis.

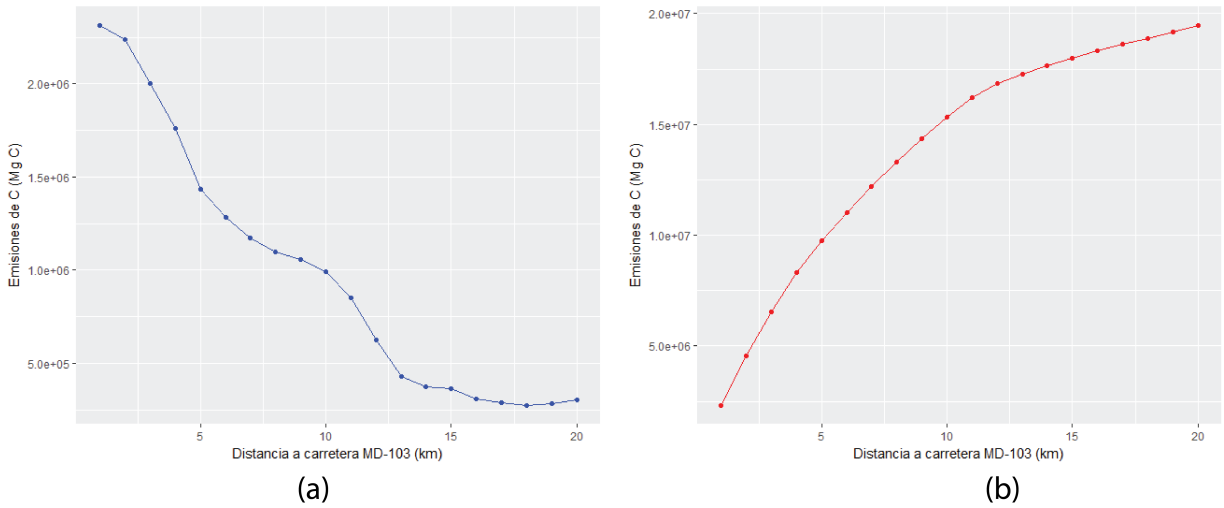

Figura 4.13: (a) Tasa de emisión de carbono por cada kilómetro de distancia a la carretera. (b) Emisión acumulada de carbono

Si se utiliza la Ecuación 3.6 para calcular las emisiones de CO_2eq se obtiene un valor de 73.2 Mt de CO_2eq. Este valor supera en demasía a las 60 Mt de CO_2eq estimadas por el Gobierno Peruano como meta de reducción. Esta estimación fue realizada a un solo proyecto vial, por lo que se puede suponer que al extender el análisis a todos los proyectos viales se obtendrás emisiones considerablemente mayores. Sin embargo, este valor estimado no representa un escenario de implementación de medidas. Los valores calculados en este estudio asumen que el comportamiento de los distintos actores será el mismo en los próximos 17 años. Bajo ese supuesto fuerte, se puede interpretar el resultado como *la cantidad de emisiones que se generarán a partir de la construcción de la carretera MD-103 si no se toman medidas*

Los códigos utilizados pueden ser visualizados accediendo a la plataforma de Earth Engine, siguiente el siguiente enlace: `https://code.earthengine.google.com/?accept_repo=users/glarrea/tesis_roads`

La aplicación de visualización puede accederse desde cualquier explorador web utilizando el siguiente enlace: https://glarrea.users.earthengine.app/view/roadstuff

Capítulo 5

Conclusiones

La presente investigación ha permitido explorar las potencialidades de diferentes herramientas computacionales del estado del arte para la resolución de distintas interrogantes. La interrogante H1 ha sido respondida mediante el cálculo de las emisiones de CO_2eq de un proyecto vial elegido como caso de estudio. Las emisiones de solo este proyecto superan a las estimaciones de reducción de todo el país propuestas por el Gobierno. En este sentido, es seguro afirmar que los compromisos ambientales planteandos por el Perú subestiman las emisiones que se podría generar en los próximos años. Sin embargo, es importante recalcar que las limitaciones del modelo utilizado no permiten que se distinga cual será el uso de suelo que tendrá la zona deforestada. Pese a esto, en cualquiera de los casos, el secuestro de carbono que ocurra no será mayor a la cantidad de carbono que se liberará a la atmósfera. Lo cierto es que el Estado Peruano ha propuesto una serie de medidas de acción para lograr la reducción deseada (MINAM (Ministerio del Ambiente), 2016). De hecho, estas medidas buscan mejorar la gestión del territorio en la Amazonía y destinar recursos para ejecutar las políticas de control, especialmente en el caso de actividades ilegales. Sin embargo, las medidas propuestas por El Gobierno no están vinculadas entre ellas. Un ejemplo de ello se da entre los sectores de transportes y USCUSS, en dónde no analizan las potenciales emisiones de actividades completamente legales, como el diseño vial o la construcción de carreteras (Vázquez-Rowe et al., 2019).

La hipótesis H2 fue validada mediante el análisis realizado a las precisiones de los diferentes modelos. En esta etapa se determinó que los modelos de aprendizaje de máquina tienen una mayor precisión que los modelos estadísticos tradicionales. No obstante, esta superioridad no es significativa si se comparán los tiempos de cómputo y la complejidad de la estructuración de los modelos de aprendizaje de máquina. Esta complejidad implica explorar múltiples combinaciones de parámetros y recurrir a herramientas computacionales que agilicen el entrenamiento.

La hipótesis H3 fue respondida al implementar un sistema entrenamiento y predicción de deforestación de toda la Amazonía peruana. Este sistema sí es facilmente replicable debido a que fue construído con lenguajes de código abierto y utilizando plataformas gratuitas. El escalamiento del modelo es posible mediante el uso de herramientas basadas en la nube, por lo que es plausible realizar predicciones de distintos escenarios y en poco tiempo. Por último, la utilización de esta herramienta en la toma de decisiones es factible con el uso de la aplicación creada para la visualización de los resultados. Se espera que en una futura mejora de la herramienta, personas con poca o nula experiencia en programación puedan acceder al modelo predictivo para evaluar diferentes escenarios.

Finalmente, este proyecto de investigación presentó un marco metodológico novedoso para construir e implementar modelos de predicción de deforestación para el cálculo de emisiones de GEI. Como ha sido presentado en este manuscrito, la simplicidad del proceso de construcción e implementación de los modelos permite generar cálculos rápidos. De igual manera, este estudio es pionero en realizar análisis de predicción de deforestación a escala nacional. De acuerdo a la flexibilidad de la metodología utilizada, espera que este enfoque de análisis pueda servir de complemento de otras metodologías que buscan integrar herramientas GIS en su estructura (Loiseau et al., 2018).

5.1. Agradecimientos

Agradezco a los profesores Ian Vázquez y Ramzy Kahhat por su participación durante la discusión y elaboración del presente manuscrito. Igualmente, deseo agradecer a CONCYTEC por el financiamiento y apoyo durante el desarrollo de esta investigación.

Bibliografía

Muzafar Ahmad Bhat, Razeef Mohd Shah, Bashir Ahmad, MARoad Srinagar, and Kashmir India. Cloud Computing: A solution to Geographical Information Systems (GIS) Cloud Computing and GIS. *International Journal on Computer Science and Engineering*, 2011. URL https://pdfs.semanticscholar.org/a93e/ed5c1980c56fdf9db61808e6aad2b8087dda.pdf.

Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, and Vijayan K. Asari. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. 3 2018. URL http://arxiv.org/abs/1803.01164.

A Angelsen and D Kaimowitz. Rethinking the causes of deforestation: lessons from economic models. *The World Bank research observer*, 14(1):73–98, 2 1999. ISSN 0257-3032. URL http://www.ncbi.nlm.nih.gov/pubmed/12322119.

David Alan Aschauer. Is public expenditure productive? *Journal of Monetary Economics*, 23 (2):177–200, 3 1989. ISSN 0304-3932. doi: 10.1016/0304-3932(89)90047-0. URL https://www.sciencedirect.com/science/article/pii/0304393289900470.

Gregory P Asner, William Llactayo, Raul Tupayachi, and Ernesto Ráez Luna. Elevated rates of gold mining in the Amazon revealed through high-resolution monitoring. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46):18454–9, 11 2013. ISSN 1091-6490. doi: 10.1073/pnas.1318271110. URL http://www.ncbi.nlm.nih.gov/pubmed/24167281http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3832012.

Gregory P. Asner, David E. Knapp, Roberta E. Martin, Raul Tupayachi, Christopher Anderson, Joseph Mascaro, F Sinca, K. Dana Chadwick, S Sousan, Mark Higgins, W Farfan, Miles Silman, William Llactayo, and A Neyra. The High-resolution Carbon Geography of Perú (Spanish) — Carnegie Airborne Observatory. Technical report, 2014.

Christopher Baraloto, Paula Alverga, Sufer Baéz Quispe, Grenville Barnes, Nino Bejar Chura, Izaias Brasil da Silva, Wendeson Castro, Harrison da Souza, Iracema Elisabeth de Souza Moll, Jim Del Alcazar Chilo, Hugo Dueñas Linares, Jorge Gárate Quispe, Dean Kenji, Matthew Marsik, Herison Medeiros, Skya Murphy, Cara Rockwell, Galia Selaya, Alexander Shenkin, Marcos Silveira, Jane Southworth, Guido H. Vasquez Colomo, and Stephen Perz. Effects of road infrastructure on forest value across a tri-national Amazonian frontier. *Biological Conservation*, 191:674–681, 11 2015. ISSN 0006-3207. doi: 10.1016/J.BIOCON.2015.08.024. URL https://www.sciencedirect.com/science/article/abs/pii/S0006320715300744.

Christopher P. Barber, Mark A. Cochrane, Carlos M. Souza, and William F. Laurance. Roads, deforestation, and the mitigating effect of protected areas in the Amazon. *Biological Conservation*, 177:203–209, 9 2014. ISSN 0006-3207. doi: 10.1016/J.BIOCON.2014.07.004. URL https://www.sciencedirect.com/science/article/abs/pii/S000632071400264X.

G. Bebis and M. Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13(4):27–31, 10 1994. ISSN 0278-6648. doi: 10.1109/45.329294. URL http://ieeexplore.ieee.org/document/329294/.

Elizabeth H. Boakes, Georgina M. Mace, Philip J. K. McGowan, and Richard A. Fuller. Extreme contagion in global habitat clearance. *Proceedings of the Royal Society B: Biological Sciences*, 277(1684):1081–1085, 4 2010. ISSN 0962-8452. doi: 10.1098/rspb.2009.1771. URL http://www.royalsocietypublishing.org/doi/10.1098/rspb.2009.1771.

Leo. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984. ISBN 9781351460491.

Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 08856125. doi: 10.1023/A:1018054314350. URL http://link.springer.com/10.1023/A:1018054314350.

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL http://link.springer.com/10.1023/A:1010933404324.

David Canning and Marianne Fay. The Effect of Transportation Networks on Economic Growth 12_-n. Technical report, 1993. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1030.1134&rep=rep1&type=pdf.

Nitesh V. Chawla. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*, pages 875–886. Springer US, Boston, MA, 2009. doi: 10.1007/978-0-387-09823-4{_}45. URL http://link.springer.com/10.1007/978-0-387-09823-4_45.

Russell G Congalton. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. Technical report, 1991. URL https://pdfs.semanticscholar.org/d7be/d062683df1fd6723fb9c0c1d26feddc8c133.pdf.

J Coronado. La brecha en infraestructura: Servicios públicos, productividad y crecimiento en el Perú (No. E10 C67). *Instituto Peruano de Economía*, 203.

Maureen Cropper, Jyotsna Puri, and Charles Griffiths. Predicting the Location of Deforestation: The Role of Roads and Protected Areas in North Thailand. *Land Economics*, 77(2):172, 5 2001. ISSN 00237639. doi: 10.2307/3147088. URL http://le.uwpress.org/cgi/doi/10.2307/3147088.

Adele Cutler, D. Richard Cutler, and John R. Stevens. Random Forests. In *Ensemble Machine Learning*, pages 157–175. Springer US, Boston, MA, 2012. doi: 10.1007/978-1-4419-9326-7{_}5. URL http://link.springer.com/10.1007/978-1-4419-9326-7_5.

Michele De Rosa. Land Use and Land-use Changes in Life Cycle Assessment: Green Modelling or Black Boxing? *Ecological Economics*, 144:73–81, 2 2018. ISSN 0921-8009. doi: 10.1016/J.ECOLECON.2017.07.017. URL https://www.sciencedirect.com/science/article/pii/S0921800916313647.

I César Delgado. Is the Interoceanic Highway exporting deforestation? *Masther Thesis*, (April):37, 2008.

Pedro Domingos and Pedro. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78, 10 2012. ISSN 00010782. doi: 10.1145/2347736.2347755. URL http://dl.acm.org/citation.cfm?doid=2347736.2347755.

P. M. Fearnside. Comment on "Determination of Deforestation Rates of the World's Humid Tropical Forests". *Science*, 299(5609):1015a–1015, 2 2003. ISSN 00368075. doi: 10.1126/science.1078714. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1078714.

Matt Finer, Sidney Novoa, Mikaela J Weisse, Rachael Petersen, Joseph Mascaro, Tamia Souto, Forest Stearns, and Raúl García Martinez. Combating deforestation: From satellite to intervention. *Science (New York, N.Y.)*, 360(6395):1303–1305, 6 2018. ISSN 1095-9203. doi: 10.1126/science.aat1203. URL http://www.ncbi.nlm.nih.gov/pubmed/29930127.

Geoffrey R. Gallice, Gustavo Larrea-Gallegos, and Ian Vázquez-Rowe. The threat of road expansion in the Peruvian Amazon. *Oryx*, pages 1–9, 6 2017. ISSN 0030-6053. doi: 10.1017/S0030605317000412. URL https://www.cambridge.org/core/product/identifier/S0030605317000412/type/journal_article.

Ian Goodfellow and Yoshua Bengio. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org.

Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 12 2017. ISSN 0034-4257. doi: 10.1016/J.RSE.2017.06.031. URL https://www.sciencedirect.com/science/article/pii/S0034425717302900.

Isabelle Guyon, André Elisseeff, and Andre@tuebingen Mpg De. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. URL http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

Haibo Haibo He and E.A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 9 2009. ISSN 1041-4347. doi: 10.1109/TKDE.2008.239. URL http://ieeexplore.ieee.org/document/5128907/.

A. Hall and D. Goodman. *The Future of Amazonia Destruction or Sustainable Development?*. Palgrave Macmillan Limited, 1991. ISBN 9781349210688. URL https://books.google.com.pe/books?hl=es&lr=&id=9K6vCwAAQBAJ&oi=fnd&pg=PR10&dq=Environmental+destruction+in+the+Amazon.+The+future+of+Amazonia:+destruction+or+sustainable+development%3F+&ots=sFCSHZkwhN&sig=LfF28zn9R5U_LllDfr3g0CGAve8#v=onepage&q=Environmen.

M C Hansen, P V Potapov, R Moore, M Hancher, S A Turubanova, A Tyukavina, D Thau, S V Stehman, S J Goetz, T R Loveland, A Kommareddy, A Egorov, L Chini, C O Justice, and J R G Townshend. High-resolution global maps of 21st-century forest cover change. *Science (New York, N.Y.)*, 342(6160):850–3, 11 2013. ISSN 1095-9203. doi: 10.1126/science.1244693. URL http://www.ncbi.nlm.nih.gov/pubmed/24233722.

Hecht-Nielsen. Theory of the backpropagation neural network. In *International Joint Conference on Neural Networks*, pages 593–605. IEEE, 1989. doi: 10.1109/IJCNN.1989.118638. URL http://ieeexplore.ieee.org/document/118638/.

David W. Hosmer, Stanley. Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. ISBN 9780470582473. URL https://www.wiley.com/en-pe/Applied+Logistic+Regression,+3rd+Edition-p-9780470582473.

R. A. Houghton, D. L. Skole, Carlos A. Nobre, J. L. Hackler, K. T. Lawrence, and W H. Chomentowski. Annual fluxes of carbon from deforestation and regrowth in the Brazilian Amazon. *Nature*, 403 (6767):301–304, 1 2000. ISSN 0028-0836. doi: 10.1038/35002062. URL http://www.nature.com/doifinder/10.1038/35002062.

INEI. Publicaciones Digitales, 2019. URL https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1483/index.html.

IPCC. 2006 IPCC Guidelines for National Greenhouse Gas, Volume 4, Agriculture, Forestry and Other Land Use. Technical report, 2006.

Ramzy Kahhat, Eduardo Parodi, Gustavo Larrea-Gallegos, Carlos Mesta, and Ian Vázquez-Rowe. Environmental impacts of the life cycle of alluvial gold mining in the Peruvian Amazon rainforest. *Science of The Total Environment*, 662:940–951, 4 2019. ISSN 0048-9697. doi: 10.1016/J.SCITOTENV.2019.01.246. URL https://www.sciencedirect.com/science/article/pii/S0048969719302736.

Anders Krogh. What are artificial neural networks? *Nature Biotechnology*, 26(2):195–197, 2 2008. ISSN 1087-0156. doi: 10.1038/nbt1386. URL http://www.nature.com/articles/nbt1386.

Gustavo Larrea-Gallegos, Ian Vázquez-Rowe, and Geoffrey Gallice. Life cycle assessment of the construction of an unpaved road in an undisturbed tropical rainforest area in the vicinity of Manu National Park, Peru. *The International Journal of Life Cycle Assessment*, 22(7):1109–1124, 7 2017. ISSN 0948-3349. doi: 10.1007/s11367-016-1221-7. URL http://link.springer.com/10.1007/s11367-016-1221-7.

William F. Laurance, Ana K. M. Albernaz, Gotz Schroth, Philip M. Fearnside, Scott Bergen, Eduardo M. Venticinque, and Carlos Da Costa. Predictors of deforestation in the Brazilian Amazon. *Journal of Biogeography*, 29(5-6):737–748, 5 2002. ISSN 0305-0270. doi: 10.1046/j.1365-2699.2002.00721.x. URL http://doi.wiley.com/10.1046/j.1365-2699.2002.00721.x.

William F. Laurance, Miriam Goosem, and Susan G.W. Laurance. Impacts of roads and linear clearings on tropical forests. *Trends in Ecology & Evolution*, 24(12):659–669, 12 2009. ISSN 0169-5347. doi: 10.1016/J.TREE.2009.06.009. URL https://www.sciencedirect.com/science/article/pii/S0169534709002067.

William F. Laurance, Anna Peletier-Jellema, Bart Geenen, Harko Koster, Pita Verweij, Pitou Van Dijck, Thomas E. Lovejoy, Judith Schleicher, and Marijke Van Kuijk. Reducing the global environmental impacts of rapid infrastructure expansion. *Current Biology*, 25(7):R259–R262, 3 2015. ISSN 0960-9822. doi: 10.1016/J.CUB.2015.02.050. URL https://www.sciencedirect.com/science/article/pii/S0960982215002195.

C. Le Quéré, R. J. Andres, T. Boden, T. Conway, R. A. Houghton, J. I. House, G. Marland, G. P. Peters, G. R. van der Werf, A. Ahlström, R. M. Andrew, L. Bopp, J. G. Canadell, P. Ciais, S. C. Doney, C. Enright, P. Friedlingstein, C. Huntingford, A. K. Jain, C. Jourdain, E. Kato, R. F. Keeling, K. Klein Goldewijk, S. Levis, P. Levy, M. Lomas, B. Poulter, M. R. Raupach, J. Schwinger, S. Sitch, B. D. Stocker, N. Viovy, S. Zaehle, and N. Zeng. The global carbon budget 1959–2011. *Earth System Science Data*, 5(1):165–185, 5 2013. ISSN 1866-3516. doi: 10.5194/essd-5-165-2013. URL https://www.earth-syst-sci-data.net/5/165/2013/.

Antoine Leblois, Olivier Damette, and Julien Wolfersberger. What has Driven Deforestation in Developing Countries Since the 2000s? Evidence from New Remote-Sensing Data. *World Development*, 92:82–102, 4 2017. ISSN 0305-750X. doi: 10.1016/J.WORLDDEV.2016.11.012. URL https://www.sciencedirect.com/science/article/pii/S0305750X16305411.

Xiaoping Liu, Guohua Hu, Yimin Chen, Xia Li, Xiaocong Xu, Shaoying Li, Fengsong Pei, and Shaojian Wang. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. *Remote Sensing of Environment*, 209:227–239, 5 2018. ISSN 0034-4257. doi: 10.1016/J.RSE.2018.02.055. URL https://www.sciencedirect.com/science/article/abs/pii/S003442571830066X.

Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23, 1 2011. ISSN 19424787. doi: 10.1002/widm.8. URL http://doi.wiley.com/10.1002/widm.8.

Eléonore Loiseau, Lynda Aissani, Samuel Le Féon, Faustine Laurent, Juliette Cerceau, Serenella Sala, and Philippe Roux. Territorial Life Cycle Assessment (LCA): What exactly is it about? A proposal towards using a common terminology and a research agenda. *Journal of Cleaner Production*, 176:474–485, 3 2018. ISSN 0959-6526. doi: 10.1016/J.JCLEPRO.2017.12.169. URL https://www.sciencedirect.com/science/article/pii/S0959652617331402.

J.F. Mas, H. Puig, J.L. Palacio, and A. Sosa-López. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling & Software*, 19(5):461–471, 5 2004. ISSN 1364-8152. doi: 10.1016/S1364-8152(03)00161-0. URL https://www.sciencedirect.com/science/article/pii/S1364815203001610.

Joseph Mascaro, Gregory P. Asner, David E. Knapp, Ty Kennedy-Bowdoin, Roberta E. Martin, Christopher Anderson, Mark Higgins, and K. Dana Chadwick. A Tale of Two "Forests": Random Forest Machine Learning Aids Tropical Forest Carbon Mapping. *PLoS ONE*, 9(1):e85993, 1 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0085993. URL https://dx.plos.org/10.1371/journal.pone.0085993.

Helen Mayfield, Carl Smith, Marcus Gallagher, and Marc Hockings. Use of freely available datasets and machine learning methods in predicting deforestation. *Environmental Modelling & Software*, 87:17–28, 1 2017. ISSN 1364-8152. doi: 10.1016/J.ENVSOFT.2016.10.006. URL https://www.sciencedirect.com/science/article/pii/S1364815216308428.

P. (Peter) McCullagh and John A. Nelder. *Generalized linear models*. Chapman and Hall, 1989. ISBN 9781351445849. URL https://books.google.com.pe/books?hl=es&lr=&id=UzmDDwAAQBAJ&oi=fnd&pg=PT14&dq=generalized+linear+models&ots=3W7WQfNY-g&sig=lXcRCwWdMgLtIPbYLZLfz7k_Urg&redir_esc=y#v=onepage&q=generalizedlinearmodels&f=false.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 12 1943. ISSN 0007-4985. doi: 10.1007/BF02478259. URL http://link.springer.com/10.1007/BF02478259.

MINAM (Ministerio del Ambiente). Sistema Nacional de Información Ambiental.

MINAM (Ministerio del Ambiente). Agenda para un desarrollo climáticamente responsable. Technical report, 2016. URL http://www.minam.gob.pe/cambioclimatico/wp-content/uploads/sites/11/2015/12/LA-CONTRIBUCI\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{O\global\mathchardef\accent@spacefactor\spacefactor}\accent19O\egroup\spacefactor\accent@spacefactor\futurelet\@let@token\penalty\@M\hskip\z@skipN-NACIONAL-DEL-PER\unhbox\voidb@x\bgroup\let\unhbox\voidb@x\setbox\@tempboxa\hbox{U\global\mathchardef\accent@spacefactor\spacefactor}\accent19U\egroup\spacefactor\accent@spacefactor\futurelet\@let@token\penalty\@M\hskip\z@skip1.pdf.

Juan Jose Miranda, Leonardo Corral, Allen Blackman, Gregory Asner, and Eirivelthon Lima. Effects of Protected Areas on Forest Cover Change and Local Communities: Evidence from the Peruvian Amazon. *SSRN Electronic Journal*, 12 2014. ISSN 1556-5068. doi: 10.2139/ssrn.2537829. URL http://www.ssrn.com/abstract=2537829.

G. C. Nelson and D. Hellerstein. Do Roads Cause Deforestation? Using Satellite Images in Econometric Analysis of Land Use. *American Journal of Agricultural Economics*, 79(1):80–88, 2 1997. ISSN 0002-9092. doi: 10.2307/1243944. URL https://academic.oup.com/ajae/article-lookup/doi/10.2307/1243944.

OECD. Land cover in countries and regions, 2015. URL https://stats.oecd.org/Index.aspx?DataSetCode=LAND_COVER.

Pontus Olofsson, Giles M. Foody, Martin Herold, Stephen V. Stehman, Curtis E. Woodcock, and Michael A. Wulder. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148:42–57, 5 2014. ISSN 0034-4257. doi: 10.1016/J.RSE.2014.02.015. URL https://www.sciencedirect.com/science/article/pii/S0034425714000704.

Stephen G. Perz, Youliang Qiu, Yibin Xia, Jane Southworth, Jing Sun, Matthew Marsik, Karla Rocha, Veronica Passos, Daniel Rojas, Gabriel Alarcón, Grenville Barnes, and Christopher Baraloto. Transboundary infrastructure and land cover change: Highway paving and community-level deforestation in a tri-national frontier in the Amazon. *Land Use Policy*, 34:27–41, 9 2013. ISSN 0264-8377. doi: 10.1016/J.LANDUSEPOL.2013.01.009. URL https://www.sciencedirect.com/science/article/pii/S026483771300029X.

Alexander S.P. Pfaff. What Drives Deforestation in the Brazilian Amazon?: Evidence from Satellite and Socioeconomic Data. *Journal of Environmental Economics and Management*, 37(1):26–43, 1 1999. ISSN 0095-0696. doi: 10.1006/JEEM.1998.1056. URL https://www.sciencedirect.com/science/article/abs/pii/S0095069698910567.

C Reid Turner, Alfonso Fuggetta, Luigi Lavazza, and Alexander L Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 12 1999. ISSN 0164-1212. doi: 10.1016/S0164-1212(99)00062-X. URL https://www.sciencedirect.com/science/article/pii/S016412129900062X.

Sassan S Saatchi, Nancy L Harris, Sandra Brown, Michael Lefsky, Edward T A Mitchard, William Salas, Brian R Zutta, Wolfgang Buermann, Simon L Lewis, Stephen Hagen, Silvia Petrova, Lee White, Miles Silman, and Alexandra Morel. Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24):9899–904, 6 2011. ISSN 1091-6490. doi: 10.1073/pnas.1019576108. URL http://www.ncbi.nlm.nih.gov/pubmed/21628575http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3116381.

Steven L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, 16(3):235–240, 9 1994. ISSN 0885-6125. doi: 10.1007/BF00993309. URL http://link.springer.com/10.1007/BF00993309.

Jannick H. Schmidt, Bo P. Weidema, and Miguel Brandão. A framework for modelling indirect land use changes in Life Cycle Assessment. *Journal of Cleaner Production*, 99:230–238, 7 2015. ISSN 0959-6526. doi: 10.1016/J.JCLEPRO.2015.03.013. URL https://www.sciencedirect.com/science/article/pii/S0959652615002309.

Ellen K. Silbergeld, Denis Nash, Circey Trevant, G. Thomas Strickland, Jose Maria de Souza, and Rui S.U. da Silva. Mercury exposure and malaria prevalence among gold miners in Pará, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 35(5):421–429, 10 2002. ISSN 0037-8682. doi: 10.1590/S0037-86822002000500001. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0037-86822002000500001&lng=en&tlng=en.

Stephen V. Stehman. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sensing Letters*, 3(2):111–120, 3 2012. ISSN 2150-704X. doi: 10.1080/01431161.2010.541950. URL http://www.tandfonline.com/doi/abs/10.1080/01431161.2010.541950.

Florence Van Stappen, Isabelle Brose, and Yves Schenkel. Direct and indirect land use changes issues in European sustainability initiatives: State-of-the-art, open issues and future developments. *Biomass and Bioenergy*, 35(12):4824–4834, 12 2011. ISSN 0961-9534. doi: 10.1016/J.BIOMBIOE.2011.07.015. URL https://www.sciencedirect.com/science/article/pii/S0961953411004119.

I. Vázquez-Rowe, R. Kahhat, G. Larrea-Gallegos, and K. Ziegler-Rodriguez. Peru's road to climate action: Are we on the right path? The role of life cycle methods to improve Peruvian national contributions. *Science of the Total Environment*, 659, 2019. ISSN 18791026. doi: 10.1016/j.scitotenv.2018.12.322.

Varsha Vijay, Chantal D Reid, Matt Finer, Clinton N Jenkins, and Stuart L Pimm. Deforestation risks posed by oil palm expansion in the Peruvian Amazon. *Environmental Research Letters*, 13 (11):114010, 11 2018. ISSN 1748-9326. doi: 10.1088/1748-9326/aae540. URL http://stacks.iop.org/1748-9326/13/i=11/a=114010?key=crossref.b79a451ac8c38ccebc29a57709336b95.

R. T. Watson, Daniel L. (Daniel Lee) Albritton, Intergovernmental Panel on Climate Change. Working Group I., Intergovernmental Panel on Climate Change. Working Group II., and Intergovernmental Panel on Climate Change. Working Group III. *Climate change 2001 : synthesis report*. Cambridge University Press, 2001. ISBN 0521807700.

Mikaela J. Weisse and Lisa C. Naughton-Treves. Conservation Beyond Park Boundaries: The Impact of Buffer Zones on Deforestation and Mining Concessions in the Peruvian Amazon. *Environmental Management*, 58(2):297–311, 8 2016. ISSN 0364-152X. doi: 10.1007/s00267-016-0709-z. URL http://link.springer.com/10.1007/s00267-016-0709-z.

I. H. (Ian H.) Witten, Eibe Frank, Mark A. (Mark Andrew) Hall, and Christopher J. Pal. *Data mining : practical machine learning tools and techniques*. 2017. ISBN 9780128043578. URL https://books.google.com.pe/books?hl=es&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=Data+Mining+Practical+Machine+Learning+Tools+and+Techniques+Witten&ots=8IEKveoEua&sig=mFej3m1MFvJymZAIUY-H6dx_lkk#v=onepage&q=DataMiningPracticalMachineLearningToolsandTec.

David H. Wolpert. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 10 1996. ISSN 0899-7667. doi: 10.1162/neco.1996.8.7.1341. URL http://www.mitpressjournals.org/doi/10.1162/neco.1996.8.7.1341.

Shobhna Yadav and Apoorvi Sood. Adaptation in Neural Networks : A Review. 2(11):3278–3281, 2013.

Cha Zhang and Yunqian Ma, editors. *Ensemble Machine Learning*. Springer US, Boston, MA, 2012. ISBN 978-1-4419-9325-0. doi: 10.1007/978-1-4419-9326-7. URL http://link.springer.com/10.1007/978-1-4419-9326-7.