PILOTO DATASANDBOX COLOMBIA: IDENTIFICACIÓN DE AREAS CONSTRUIDAS A PARTIR DE IMÁGENES SATELITALES

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Desarrollo Digital - Unidad de Científicos de Datos • Ministerio de Tecnologías de la Información y las Comunicaciones • BEXTechnology
Sector	Planeación
Tecnologías utilizadas	Microsoft Azure (Databricks-Python), QGIS
Fuentes de datos	Google Earth Engine, Instituto Geográfico Agustín Codazzi (IGAC)

Contenido

1. Presentación 2
2. Objetivos del proyecto 3
3. Metodología 3
4. Resultados 8
5. Conclusiones y recomendaciones 11
6. Socialización 12

Contacto 12

ANEXOS 13

Anexo 1 Modelos y configuración óptima de parámetros 13

1. Presentación

En Colombia alrededor del 40% de los desastres naturales están asociados a inundaciones causadas por el desbordamiento de cuerpos de agua. En el marco de la evaluación y prevención de estos desastres es de suma importancia identificar infraestructuras construidas en zonas de susceptibilidad. En 2019 la Unidad de Científicos de Datos (UCD) del Departamento Nacional de Planeación (DNP) desarrolló una herramienta para identificar ocupación o infraestructura construida dentro de las zonas demarcadas como ronda hídrica[1]. Estas zonas se caracterizan por tener una alta susceptibilidad de daños o perdidas por inundaciones debido al desbordamiento de los ríos. El objetivo de la herramienta es generar insumos a las direcciones técnicas para la planificación territorial y la prevención de desastres naturales.

En el presente piloto se aborda un proceso intermedio que permite el adecuado funcionamiento de la herramienta. Este consiste en permitir el acceso a información geoespacial del conjunto de construcciones. En Colombia solo el 5% del territorio nacional cuenta con un catastro actualizado[2], lo cual sugiere que la información disponible es escaza. Para contrarrestar esta falta de información, la herramienta mencionada emplea técnicas de predicción que permiten identificar áreas con infraestructura construida a partir de imágenes satelitales. El objetivo de este proyecto es mejorar dichas técnicas de predicción a través de la búsqueda de un modelo de aprendizaje de máquinas utilizando las tecnologías suministradas por el DataSandbox Colombia[3].

In Colombia, around 40% of natural disasters are the consequence of water floods. In the context of evaluating and preventing natural disasters is extremely important to identify infrastructures at risk of damage or loss. During 2019 the Data Scientists Unit of the National Planning Department developed a tool that identifies constructions in water-round zones[4], which are at risk in case of rivers overflooding due to their nearness to the river's basin. Its main goal is to provide valuable information for territorial planning and prevent natural disasters.

This project tackles an intermediate process of the mentioned purpose by identifying areas of built infrastructure. In Colombia, only 5% of the territory has an update cadaster[5]. Thus, the mentioned tool uses prediction techniques to face the lack of geospatial information. This work aims to improve the actual prediction model by searching a machine learning algorithm using the technologies available in the DataSandbox Colombia[6].

---

[1] El decreto 2811 de 1974, artículo 83, consagra que debe existir una franja paralela como mínimo de 30 metros al cauce de los ríos, la cual se denomina **ronda hídrica**.

[2] Dentro del Plan Nacional de Desarrollo 2018-2022, se prioriza la actualización catastral como eje central para el desarrollo, la transformación y consolidación territorial. Se espera que para el 2022 se logre una actualización catastral del 60% y del 100% del territorio en 2025.

[3] El DataSandbox Colombia es un espacio experimental que busca promover el uso de *Big Data* en el sector público a nivel nacional.

[4] *The Decree 2811 of 1974, Article 83, expose there must be a parallel strip of at least 30 meters on both sides of the riverbed. This area is called **Water-Round** and should remain unpopulated.*

[5] *In the National Development Plan 2018-2022, cadastral updating prioritization is crucial for regional development and consolidation. The protections for 2022 are that 60% of the cadaster is updated and 100% by 2025.*

[6] *DataSandbox Colombia is an experimental space that aims to promote the use of Big Data in public entities.*

2. Objetivos del proyecto

2.1. General

Contribuir a la identificación de infraestructuras susceptibles de inundación al generar un modelo automático de detección de construcciones a partir de imágenes satelitales. De esta manera se busca mejorar los insumos de las direcciones técnicas para la prevención de riesgos por inundación y toma de decisiones en Planes de Ordenamiento Territorial.

2.2. Específicos

1. Mejorar el modelo de clasificación de imágenes actual al hacer uso de nuevas metodologías compatibles con el DataSandbox Colombia.
2. Integrar el uso de información georreferenciada del IGAC referente a la capa de construcciones.
3. Promover el uso de Big Data en el sector público a través del DataSandbox Colombia.

3. Metodología

El desarrollo de este proyecto se divide en tres etapas principales; ingesta de datos, procesamiento y modelado. De manera general, la ingesta de datos consistió en obtener información georreferenciada de los municipios seleccionados para entrenar el modelo de predicción. Dicha información hace referencia a la capa de construcciones y su respectiva imagen satelital. En segunda instancia, los datos obtenidos fueron procesados para ser utilizados en la búsqueda y entrenamiento del modelo de predicción. Finalmente, los resultados obtenidos consisten en un proceso automático que toma coordenadas de la región a analizar y produce una predicción de áreas de construcción sobre la región de análisis.

Cada una de las etapas mencionadas se desarrollaron en Microsoft Azure, herramienta suministradas por el DataSandbox Colombia. A continuación, se presenta detalladamente cada una de estas etapas.

3.1. Ingesta de Datos

- ***IGAC: Información geoespacial***

Para identificar si una construcción puede estar en peligro de sufrir daños o perdidas por inundaciones es necesario tener información acerca de los mapas hídricos del territorio nacional y las construcciones. En Colombia, la entidad encargada de producir la cartografía básica de Colombia y elaborar el catastro nacional es el IGAC. En septiembre de 2020, esta entidad anunció la actualización catastral de municipios de diez departamentos del país. En el presente proyecto se utilizaron datos catastrales de 7 municipios ubicados en dos de los diez departamentos en mención. Estos municipios corresponden a La Plata, Pitalito y Garzón, que pertenecen al departamento del Huila, y los municipios Piendamó, Puerto Tejada, Santander de Quilichao y Popayán, que hacen parte del departamento del Cauca.

Los datos utilizados corresponden a la capa de construcciones rural y urbana. La cual consiste en una serie de polígonos georreferenciados. Estos datos fueron obtenidos de la Base de Datos Geográfica Catastral[7]. En la *Figura 1* se muestra un ejemplo de esta información para el municipio de La Plata – Huila.

[7] Esta base de datos se puede obtener en el siguiente enlace: https://geoportal.igac.gov.co/contenido/datos-abiertos-catastro. Allí se puede acceder a diferentes conjuntos de datos en múltiples formatos, lo cual permite el análisis de estos en lenguajes de programación tales como Python.

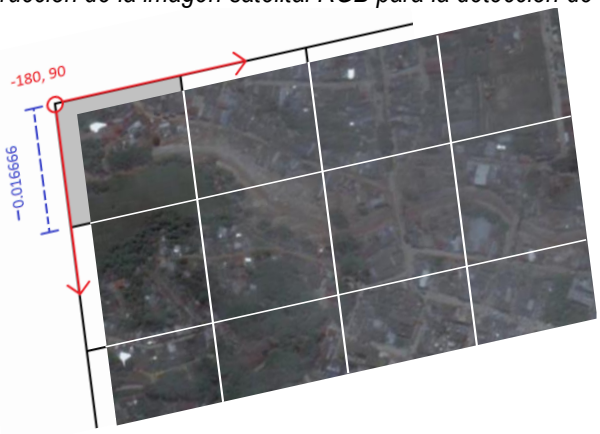
Figura 1: Capa de construcciones Rural y Urbana del municipio de La Plata

Fuente: Elaboración propia con datos de la Base de Datos Geográfica Catastral del IGAC (2020)

- **Imágenes Satelitales**

Como se mencionó anteriormente, la información de construcciones del IGAC se encuentra en proceso de expansión y no está disponible para la mayoría del territorio nacional. Por esta razón es necesario estimar una aproximación de las capas de construcciones mediante el análisis de imágenes satelitales. Estas imágenes fueron obtenidas de los mapas base de *GoogleMaps* y *ESRI* a través de servicios web que se encuentran disponibles en estos portales. Las imágenes recopiladas están conformadas por una serie de baldosas georreferenciadas dentro del casco urbano de los municipios mencionados. Dichas imágenes tienen tres canales: rojo, verde y azul (RGB por sus siglas en inglés). La *Figura 2* muestra cómo se reconstruye la imagen de la región de interés a partir de las baldosas georreferenciadas.

Figura 2: Construcción de la imagen satelital RGB para la detección de construcciones



Fuente: *GoogleMaps y ESRI*

Cada baldosa descargada tiene un tamaño de 256x256 pixeles, lo cual implica que las dimensiones de las imágenes de análisis sean múltiplos de 256 en ambas dimensiones y el tamaño mínimo sea de 256x256 pixeles. Además, cada baldosa es capturada a un nivel de zum 17, para el cual cada píxel cubre un área de 1.089 metros cuadrados y cada baldosa cubre un área de 71320 metros cuadrados aproximadamente.

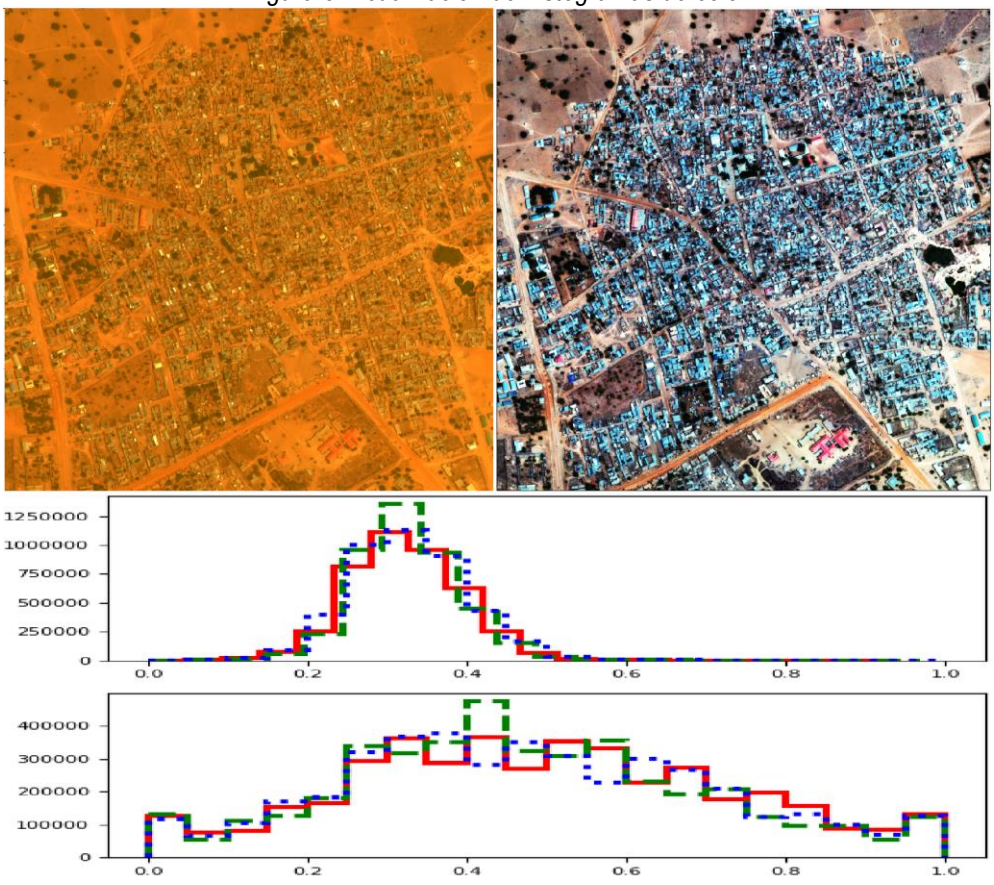
3.2. Procesamiento

Luego de obtener los dos elementes esenciales para el análisis (imagen satelital y capa de construcciones), es necesario procesar dicha información para que esta pueda ser utilizada en la búsqueda y entrenamiento del modelo de predicción. Para ello se requiere de dos etapas; la extracción de características de las imágenes y la etiqueta de áreas de construcción. A continuación, se presenta con más detalle estas tareas.

- ***Procesamiento de imágenes satelitales***

La etapa de procesamiento inicia con la normalización de las intensidades de la imagen satelital debido a que un canal de color puede ser dominante con respecto a los demás, lo que entorpecería el proceso de clasificación. Para ello se hace una ecualización de los histogramas de color por cada canal. Como se evidencia en la *Figura 3*, este preprocesamiento mejora la calidad de la imagen. Además, en el panel inferior de la figura se muestra la distribución antes y después de aplicar la ecualización. Allí se evidencia que los histogramas de color se distribuyen de manera más homogénea a lo largo de todas las intensidades de color.

Figura 3: Ecualización de histogramas de color



Fuente: Elaboración propia.

Luego de que la imagen pasa por el proceso de ecualización de histogramas se procede a extraer las características que se utilizarán en el proceso de modelado. Para ello se divide la imagen en cuadrículas de análisis de tamaño kxk pixeles. Posteriormente se cambia el espacio de color de la imagen, de RGB a matiz-saturación-valor (HSV, por sus siglas en ingles). Una vez la imagen pasa al espacio de color HSV se extraen N características por cada canal H-S-V. En suma, cada cuadrícula de kxk pixeles tendrá $Nx3$ características asociadas que serán utilizadas en el proceso de entrenamiento. La siguiente figura ilustra la división de la imagen satelital en cuadrillas de $135x135$ pixeles y la creación de la tabla de atributos que se utilizará en la búsqueda del modelo de clasificación.

Figura 4: Construcción de tabla de Atributos

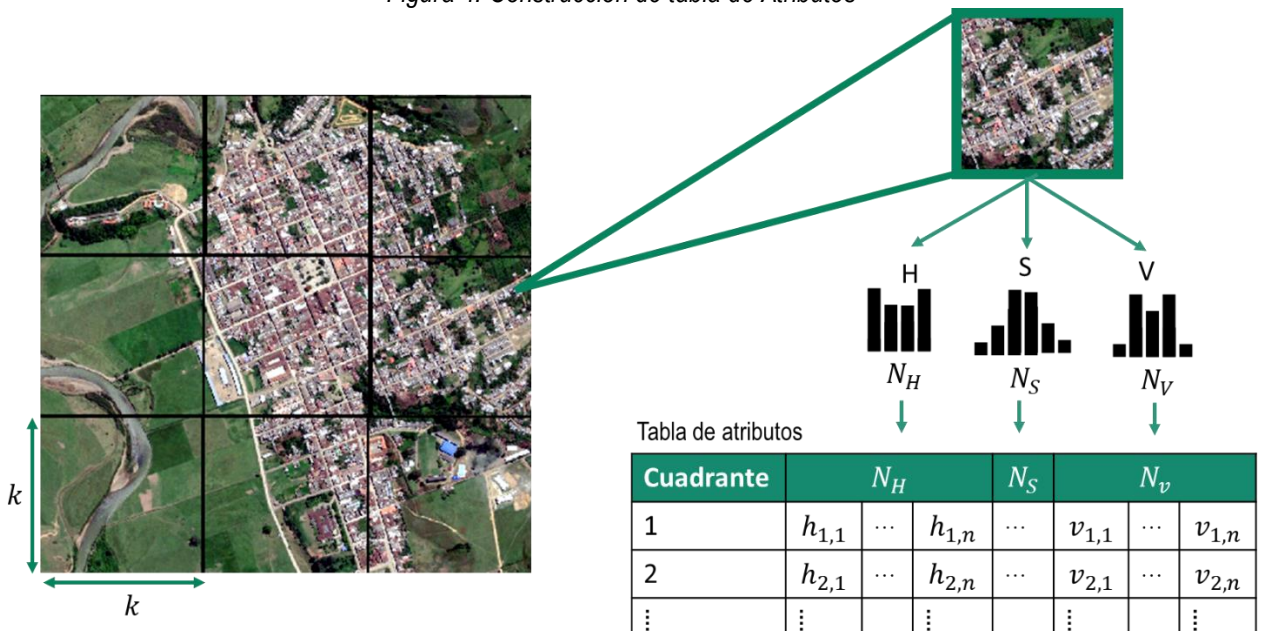


Tabla de atributos

Cuadrante	N_H			N_S	N_v		
1	$h_{1,1}$	⋯	$h_{1,n}$	⋯	$v_{1,1}$	⋯	$v_{1,n}$
2	$h_{2,1}$	⋯	$h_{2,n}$	⋯	$v_{2,1}$	⋯	$v_{2,n}$
⋮	⋮		⋮		⋮		⋮
C	$h_{c,1}$	⋯	$h_{c,n}$	⋯	$v_{c,1}$	⋯	$v_{c,n}$

Fuente: Elaboración propia con imagen tomada de GoogleMaps.

- ***Etiqueta de imágenes por cuadrantes***

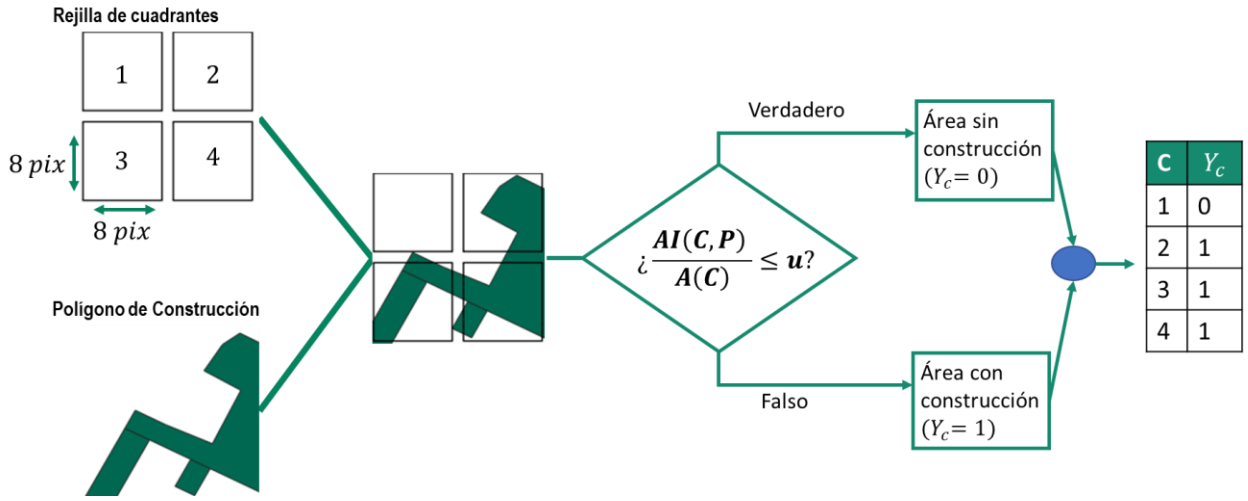
Una vez la imagen es procesada y se ha obtenido la tabla de atributos, el paso a seguir es elaborar el vector de etiquetas asociado a dicha tabla. Esto con el fin de abordar el problema de clasificación de manera supervisada. El vector de etiquetas tendrá elementos binarios que indican si dentro de cada cuadrante existe o no una infraestructura construida. Esto implica que se debe usar la misma máscara de cuadrantes que se utilizó en el proceso de extracción de atributos de la imagen. Para etiquetar cada cuadrante se realizó una operación de intersección entre la máscara de rejillas y la capa de construcciones obtenida en la ingesta de datos. De esta manera, si una construcción ocupa una porción determinada del cuadrante, dicho cuadrante es etiquetado como área construida. De lo contrario es etiquetado como área no construida. La siguiente ecuación muestra la manera en que se realiza la asignación de etiquetas y la figura 5 ilustra el proceso.

$$y_c = \begin{cases} 0 \leftrightarrow \dfrac{\boldsymbol{AI(C,P)}}{\boldsymbol{A(C)}} \leq \boldsymbol{u} \\ 1 \leftrightarrow \dfrac{\boldsymbol{AI(C,P)}}{\boldsymbol{A(C)}} > \boldsymbol{u} \end{cases}$$

Donde:

- $\boldsymbol{AI(C,P)}$ hace referencia al área de intersección entre el cuadrante $(\boldsymbol{C})$ y el polígono de construcción $(\boldsymbol{P})$.
- $\boldsymbol{A(C)}$ hace referencia al área del cuadrante $(\boldsymbol{C})$, que es de aproximadamente 69.649 metros cuadrados.
- $\boldsymbol{u}$ hace referencia a la proporción determinada como límite de decisión.

Figura 5: Construcción de Vector de Etiquetas



Fuente: Elaboración propia.

Luego de obtener la etiqueta de construcciones se puede construir la tabla de atributos utilizada en la etapa de búsqueda y entrenamiento del modelo de clasificación. En la *tabla 1* se muestra el resultado de la etapa de procesamiento descrita.

Tabla 1: Conjunto de Datos de entrenamiento

Cuadrante	N_H			N_S	N_v			Y
1	$h_{1,1}$	…	$h_{1,n}$	…	$v_{1,1}$	…	$v_{1,n}$	Y_1
2	$h_{2,1}$	…	$h_{2,n}$	…	$v_{2,1}$	…	$v_{2,n}$	Y_2
⋮	⋮		⋮		⋮		⋮	⋮
C	$h_{c,1}$	…	$h_{c,n}$	…	$v_{c,1}$	…	$v_{c,n}$	Y_C

Fuente: Elaboración propia.

3.3. Modelado

En la etapa de modelado se abordó un problema de clasificación supervisado que consistió en estimar la probabilidad de que cada cuadrilla de kxk pixeles fuese un área con infraestructura construida. Para ello se debe tomar una muestra representativa que contenga la información descrita en la etapa de ingesta de datos. La muestra debe estar balanceada de tal que el número de cuadrillas con construcción sea igual al número de cuadrillas sin construcción. Al final se debe contar con una tabla de datos de C registros y $((Nx3) + 1)$ columnas correspondientes a los cuadrantes de kxk pixeles y a las N características por canal (HSV), más el vector de etiquetas o valores verdaderos, respectivamente. Posteriormente se deben dividir los datos en un conjunto de entrenamiento y un conjunto de prueba cuya proporción permita una adecuada evaluación de desempeño de los modelos utilizados.

Luego de contar con los insumos necesarios para la búsqueda y entrenamiento del modelo de predicción se procede a utilizar el conjunto de datos de entrenamiento para evaluar múltiples modelos de clasificación de manera simultánea. Esto se hizo por medio de una red de búsqueda empleando las tecnologías suministradas por el DataSandbox Colombia. La búsqueda consistió en evaluar múltiples modelos con diferentes combinaciones de sus respectivos parámetros. Para obtener las métricas de desempeño de cada uno de los modelos se empleó un método de validación cruzada de 10 separaciones y un tamaño de prueba correspondiente al 19% de los datos de entrenamiento. Para seleccionar el conjunto de parámetros ideal para cada modelo se utilizó la media y desviación estándar del Valor F1 (*f1-score)* a lo largo de las 10 separaciones de validación cruzada. Esta métrica es la media armónica entre la precisión y la sensibilidad, por lo que permite comparación entre estas dos características. La fórmula de dicha métrica se presenta a continuación:

$$Valor\ F_1 = 2 \times \frac{precisión \times Sensibilidad}{precisión + sensibilidad}$$

Donde:

- $\boldsymbol{precisión}$ es la proporción de áreas correctamente predichas por el modelo como áreas con construcción.
- $\boldsymbol{sensibilidad}$, o exhaustividad, es la proporción de áreas predichas por el modelo que en efecto son áreas con infraestructura construida.

Luego de escoger los modelos con mejor desempeño se debe realizar la configuración óptima del umbral de decisión a utilizar. Dicho umbral de decisión determina qué probabilidad estimada se requiere para considerar una cuadrilla cono área de infraestructura construida. A continuación, se presentan los resultados obtenidos después de aplicar la metodología anteriormente descrita.

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte de la comunidad es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

En primer lugar, luego de realizar una búsqueda heurística sobre el tamaño del área de análisis y el límite de decisión deseado ($\boldsymbol{k}\ y\ \boldsymbol{u}$, respectivamente), se utilizaron cuadrículas de 8x8 pixeles, lo que equivale a 69.649 metros cuadrados aproximadamente y un límite de decisión de 0.1. En este sentido, si la intersección entre un cuadrante de 8x8 pixeles y un polígono de construcción supera el 10% del área total del cuadrante, el cuadrante se etiquetó como área construida. Esta asignación se llevó a acabo siguiendo lo expuesto en la *Figura 5.*

Por otra parte, la base de datos utilizada para la búsqueda y entrenamiento del modelo de clasificación se construyó a partir de la imagen satelital e información de construcciones de los municipios de La Plata, Pitalito, Garzón y Piendamó. Luego de unir los datos de estos 4 municipios se balanceó la muestra de tal manera que el número de

cuadrillas con construcción fuese igual al número de cuadrillas sin construcción. Al final se obtuvo una tabla de datos de 65072 registros y 106 columnas correspondientes a cuadrantes de 8x8 pixeles y 35 características ($\boldsymbol{N = 35}$) por canal (HSV) más el vector de etiquetas o valores verdaderos. El número de características por canal también se escogió luego de una búsqueda heurística. Luego de obtener los datos necesarios siguiendo la etapa descrita de procesamiento se dividió la muestra en un conjunto de entrenamiento y un conjunto de prueba cuya proporción fue de 0.85 y 0.15 del total de los datos, respectivamente.

A continuación, se presenta el resultado de la búsqueda y entrenamiento de los modelos considerados. La siguiente tabla muestra el desempeño de estos sobre el conjunto de datos de prueba. En el *Anexo 1* se muestra la configuración de parámetros óptima para cada uno de los modelos presentados.

Tabla 2: Resultado de Evaluación de Modelos

Modelo	Precisión	Sensibilidad	Exactitud	Valor F1
XGboost[8]	0.841	0.886	0.861	**0.863**
Support Vector Classifier (SVC)	0.834	0.886	0.856	**0.859**
Bagging SVC	0.834	0.886	0.856	0.859
K vecinos más cercanos	0.804	0.899	0.841	0.849
Adaboost Classifier	0.825	0.855	0.838	0.840
Random Forest	0.793	0.867	0.822	0.829
Logistic Clasiffier	0.808	0.823	0.816	0.816
Stochastic gradient descent	0.809	0.823	0.816	0.816
Gaussian Naive Bayes	0.738	0.869	0.782	0.798
Arboles de Decisión	0.775	0.768	0.775	0.772

Nota: Métricas obtenidas sobre 9761 observaciones[9].

Como se evidencia en la *Tabla 2*, los modelos que obtuvieron el mejor desempeño fueron el *XGboost* y el *SVC*. Sin embargo, los modelos presentados en la *Tabla 2* utilizan un umbral de decisión de 0.5 sobre la probabilidad predicha por los modelos. El umbral de decisión determina la probabilidad mínima necesaria para considerar una cuadrilla como área de construcción. Este parámetro puede ser modificado con el fin de mejorar el desempeño de los modelos. Como último paso del proceso de modelado se realizó la búsqueda del umbral de decisión que arrojase los mejores resultados para los modelos *XGboost* y *SVC*. A continuación, se presenta el desempeño de dichos modelos luego de optimizar el umbral de decisión.

[8] *XGboost* hace referencia al modelo *Extreme Gradient boosting.*
[9] Estas observaciones fueron escogidas de manera aleatoria y corresponden al 15% del conjunto de datos procesados. Además, para asignar etiquetas de clasificación se utilizó un umbral de decisión de 0.5 sobre la probabilidad predicha por los modelos.

Tabla 3: Resultado de Evaluación de Modelos XGboost y SVC optimizados

Modelo	Precisión	Sensibilidad	Exactitud	Valor F1	Umbral de Decisión
XGboost	0.830	0.907	0.862	0.867	0.441
Support Vector Classifier (SVC)	0.816	0.913	0.855	0.862	0.368

Nota: Métricas obtenidas sobre 9761 observaciones.

En suma, los modelos presentados en la tabla anterior son los modelos con los cuales se implementa la predicción de áreas de construcción a partir de imágenes satelitales. La principal diferencia entre ambos es el tiempo computacional que requieren para realizar predicciones. Mientras que el modelo *XGboost* se toma alrededor de 2 segundos para realizar predicciones sobre una imagen, el modelo *SVC* se toma alrededor de 60 segundos. Esto, sumado a un mejor desempeño, hace que sea preferible realizar las predicciones a través del modelo *XGboost.*

Ahora bien, el proceso de generación de resultados consiste en generar datos georreferenciados de áreas de construcción predichas por el modelo. La implementación de la metodología descrita consiste en asignar una probabilidad a cada cuadrilla de 8x8 pixeles perteneciente a la región de análisis. Esta región está determinada por las coordenadas suministradas al proceso. Por último, se generan polígonos correspondientes a las regiones clasificadas como construcción, teniendo en cuenta el umbral de decisión establecido. A continuación, se presenta un ejemplo del proceso para el municipio de Puerto Tejada utilizando el modelo de predicción *XGboost*.

Figura 6: Predicción de áreas de construcción sobre Puerto Tejada

Fuente: Elaboración propia con imagen tomada de *GoogleMaps.*

En total, se obtuvieron resultados correspondientes a 11 municipios de Colombia. De los cuales 4 corresponden a municipios de entrenamiento, 3 corresponden a municipios de prueba y 4 municipios que no cuentan con la información oficial respecto a la capa de construcciones. En la siguiente tabla se presentan las métricas de desempeño del modelo respecto a la predicción sobre los 3 municipios de prueba, incluyendo la predicción presentada en la *Figura 6*.

Tabla 4. Desempeño de modelo de predicción XGboost sobre municipios de prueba

Municipio	Precisión	Sensibilidad	Exactitud	Valor F1	Observaciones
Puerto Tejada	0.750	0.921	0.807	0.827	22888
Santander de Quilichao	0.718	0.868	0.771	0.786	25599
Popayán	0.746	0.857	0.783	0.798	29132

Fuente: Elaboración propria con imágenes tomadas de *GoogleMaps*

5. Conclusiones y recomendaciones

Como se mencionó en la presentación, el objetivo de este proyecto es mejorar el proceso de predicción de construcciones empleado en la herramienta de identificación de construcciones susceptibles a sufrir daños por inundaciones. En este sentido, la metodología presentada anteriormente cumple con dicho objetivo y permite identificar áreas con infraestructura construida a partir de imágenes satelitales. El proceso desarrollado en este piloto toma como insumo las coordenadas de la región que se desea analizar y genera los cuadrantes predichos como áreas de construcción con su respectiva probabilidad asociada. Dichos cuadrantes son georreferenciados por lo que pueden ser utilizados en programas de análisis GIS como ArcMap y QGIS.

En conclusión, si bien se mejoró el método de predicción de construcciones sobre imágenes satelitales, este todavía tiene oportunidades de mejora. En especial, la cantidad de cuadrillas asignadas a la clase de áreas con construcción sin en realidad serlo sigue siendo elevada. Esto se evidencia en la precisión que arrojó el modelo para los municipios de prueba. Ahora bien, para mejorar el desempeño del modelo se podrían aumentar el número de imágenes satelitales utilizadas en el periodo de búsqueda y entrenamiento de los modelos de predicción, así como implementar nuevas tecnologías que pueden llegar a tener un mejor desempeño si la cantidad de datos es lo suficientemente amplia. Sin embargo, como se mencionó en la presentación del proyecto, existe una baja disponibilidad de datos oficiales respecto a la capa de construcciones en la mayoría del territorio nacional. Esto hace que el número de imágenes satelitales adicionales que se podrían emplear para el análisis sea limitado. Sin embargo, se pueden consultar bases de datos públicas que contengan imágenes satelitales y su respectiva información de construcciones. Además, como se mencionó anteriormente, se espera que para el 2025 la actualización catastral sea del 100% por lo que es importante emplear alternativas de predicción mientras esto ocurre. En suma, es necesario realizar una revisión respecto a la disponibilidad de datos que podrían llegar a servir para mejorar aún más el proceso de identificación de construcciones.

6. Socialización

El proyecto fue publicado en el repositorio de *Github* de Inundaciones. Además, se realizó la publicación de resultados en el portal de Datos Abiertos de Colombia. Estos datos fueron generados el 28 de diciembre del 2020 y los datos están proyectados en formato MagnaSirgas (EPSG:3115). La siguiente tabla resume los datos publicados en el portal.

Tabla 4: Publicación en Datos Abiertos

Nombre del Archivo	Municipio	Número de Registros	Enlace
DNP-XGboostPitalito	Pitalito	8732	DNP-XGboostPitalito | Datos Abiertos Colombia
DNP-XGboostPiendamo	Piendamó	4058	DNP-XGboostPiendamo | Datos Abiertos Colombia
DNP-XGboostLaPlata	La Plata	16642	DNP-XGboostLaPlata | Datos Abiertos Colombia
DNP-XGboostGarzon	Garzón	10348	DNP-XGboostGarzon | Datos Abiertos Colombia
DNP-XGboostCerete	Cereté	11233	DNP-XGboostCerete | Datos Abiertos Colombia
DNP-XGboostCartago	Cartago	13882	DNP-XGboostCartago | Datos Abiertos Colombia

Fuente: Elaboración propia

Estos datos pueden ser descargados en formatos que permiten su utilización en programas de análisis GIS como ArcMap y QGIS.

Contacto
Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.

ANEXOS

Anexo 1 Modelos y configuración óptima de parámetros

Tabla A.1: Lista de modelos y parámetros óptimos

Modelo	Parámetro	Valor
XGboost	*early_stopping_rounds*	50
	learning_rate	0.069559608
	max_depth	19
	maximize	False
	min_child_weight	11.9950663029316
	num_boost_round	1000
	reg_alpha	0.341448739
	reg_lambda	0.155133834
	verbose_eval	True
Support Vector Classifier (SVC)	c	3.095578547
	kernel	rbf
Bagging SVC	*base_estimator*	SVC (c = 3.095578547)
	boostrap	*False*
K Vecinos más cercanos	*n_neighbors*	*29*
Adaboost Classifier	*algorithm*	*'SAMME.R'*
	n_estimators	*770*
Random Forest	*criterion*	*gini*

	max_depth	*5*
Logistic Classifier	*c*	*0.28088685*
	solver	*lbfs*
Stochastic gradient descent	*alpha*	*0.001086987*
	l1_raio	*0.14251322*
	loss	*modified_huber*
Gaussian Naive Bayes	*var_smoothing*	*1.00E-08*
Árboles de decisión	*ccp_alpha*	*0*
	criterion	*'gini'*
	splitter	*'best'*