Proyecto 1

Campaña para fomentar el turismo de avistamiento

Rogelio Garcia - 201326488
Stephannie Jimenez - 201423727

0. Misión

Somos *Mapache Inc*, una empresa que trabaja con inteligencia de negocios con un énfasis especial para la conservación del medio ambiente, turismo ecológico, responsabilidad ambiental y temas afines. Fuimos contratados por el Instituto Humboldt en el marco de una campaña que busca incentivar el ecoturismo de una manera responsable. El objetivo de negocio es poder crear un esquema en donde visitantes colombianos o extranjeros, adultos o niños, puedan ir a parques naturales de Colombia, realizar actividades de avistamiento de especies y luego registrar dichos avistamientos en línea, en lenguaje natural y con poco detalle. El trabajo de análisis de datos que realizaremos es un crear un modelo predictivo para clasificar las entradas de turistas geográficamente y temporalmente.

1. **Selección de Fuentes**

Se encontraron dos bases de datos que contienen información sobre el tema de biodiversidad en Colombia. La primera base de datos pertenece a la Universidad Nacional y contiene un gran número de datos sobre la flora y la fauna que se puede encontrar en diferentes parques nacionales del país[1]. Tiene alrededor de los cien mil registros con 44 columnas de información del hábitat y de la especie siguiendo el formato de *DarwinCore*. Esta base de datos contiene las plantas y animales que se han visto en el parque natural desde 1995 hasta la actualidad. Por su parte, la segunda base de datos se obtuvo de la página de datos abiertos del gobierno colombiano[2]. Esta contiene un gran número de información sobre los visitantes que han visitado los parques desde 1995 hasta 2014. Esta se encuentra definida por cada uno de los parques con la cantidad total de visitantes, y su diferenciación por edad.

2. **Descripción y Preparación de los Datos**

Descripción

Biodiversidad en Colombia - Universidad Nacional	
Número de columnas	44
Número de filas	109270
Frecuencia de generación	Anual*
Frecuencia de actualización	Mensual*

*Se publicaron los datos en diciembre de 2016 y ha visto 41 revisiones menores, pocas de ellas con actualizaciones de datos.

Sobre la fuente de datos 2 del ingreso de visitantes a Parques Naturales Nacionales, se tienen los siguientes metadatos:

Ingreso de Visitantes Parques Naturales	
Fuente	Parques Nacionales Naturales de Colombia
Número de columnas	55
Número de filas	972
Frecuencia de generación	Anual

La fuente 2 de datos tiene 55 columnas. 1 que describe la dirección territorial del parque, otra para el nombre del parque (área protegida), otra para el mes que abarca la toma de los datos. Las demás columnas corresponden a los datos. Las siguientes 32 columnas muestran el número de visitantes y están desglosadas por tipo de población (adultos nacionales, niños y estudiantes, extranjeros, extranjeros residentes, niños residentes,

[1] https://www.gbif.org/dataset/79684ec1-01e8-46cc-83cb-cd5bdfb469fe#dataDescription

[2] https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Ingreso-de-visitantes-a-Parques-Nacionales-Nat/vq7a-34kf

adultos residentes, niños excentos no residentes[sic], adultos excentos no residentes[sic]) por cada año entre 2011 y 2014. Las últimas 20 columnas corresponden al total de visitantes año por año desde 1995 hasta 2014.
La columna dirección territorial es una variable categórica en texto simple con seis posibles valores: DIRECCION TERRITORIAL AMAZONIA, DIRECCION TERRITORIAL ANDES OCCIDENTALES, DIRECCION TERRITORIAL ANDES NORORIENTALES, DIRECCION TERRITORIAL CARIBE, DIRECCION TERRITORIAL ORINOQUIA, DIRECCION TERRITORIAL PACIFICO. La columna de área protegida tiene 27 posibles valores: ANU Los Estoranques, PNN El Cocuy, PNN Pisba, PNN Tamá, PNN Chingaza, PNN Sumapaz, PNN Tuparro, PNN Amacayacu, PNN Cueva de los Guacharos, PNN Nevados, PNN Puracá, PNN Corales del Rosario, PNN Macuira, PNN Old Providence, PNN Sierra Nevada, PNN Tayrona, PNN Gorgona, PNN Utría, SFF Galeras, SFF Colorados, SFF Flamencos, SFF Isla de la Corota, SFF Otún Quimbaya, SFF Guanenta - Alto Rio Fonce, SFF Iguaque, SFF Malpelo, VP Isla de Salamanca. La columna de mes tiene 12 posibles valores, uno por cada mes. Cada fila de datos se repite tres veces (más al respecto en *calidad de datos*), por lo que hay 36 filas para cada área protegida, tres por cada mes y por lo tanto, 36 apariciones de cada dirección territorial por área protegida relacionada. Es decir, el número de registros es:

Columna	Valor categórico	Registros
dirección territorial	AMAZONAS (1 área protegida)	36
	CARIBE (8 áreas protegidas)	288
	ANDES OCCIDENTAL (6 áreas protegidas)	216
	ANDES NORORIENTAL (6 áreas protegidas)	216
	ORINOQUIA (3 áreas protegidas)	108
	PACIFICO (3 áreas protegidas)	108
área protegida	todos	36
mes	todos	81

Los valores numéricos se caracterizaron por aparte, uniéndolos en dos categorías: los valores desglosados por tipo de visitante y los valores totales por año.

Tipo	Caracterización
Valores desglosados por tipo de visitante	Mínimo: 0. Máximo: 68886. Este fue el número de visitantes adultos nacionales para el parque nacional natural Corales del Rosario (Caribe) en el 2013.
Totales anuales	Mínimo: 0. Máximo: 297948, visitantes totales para el parque nacional natural Corales del Rosario (Caribe) en 1996.

Es importante notar que el mínimo para los totales anuales es 0. Esto es sin lugar a dudas ruido y falencia en los datos. No hay datos nulos, sin embargo hay ceros y se deben de alguna manera filtrar los ceros reales y los ceros que afectan la calidad y confiabilidad de los datos.

Preparación de los datos

La transformación de los datos está documentada en su totalidad en el archivo 'preprocesamiento.py' y 'Merge_Reports.py'. Aquí un resumen de la preparación de datos para cada fuente:

Fuente 1:

- Se borraron 32 columnas innecesarias según el análisis de calidad
- Se borraron avistamientos no provenientes de Colombia
- Se borraron registros anteriores a 1995.
- Se borraron registros que no tuvieran datos de año, clase, orden o familia.
- Las columnas con caracteres se cambiaron a mayúsculas únicamente (reino, especies, nombre científico, parque, región natural)
- Se quitaron los caracteres especiales y se reemplazaron por caracteres normales.
- Los datos de lugar del avistamiento, que están escritos en lenguaje natural, fueron contrastados con los parques del piloto de la fuente 2 de datos. Si contenía el nombre del parque, se guardaba la fila y el lugar del avistamiento se corregía, y si no, se borraba la fila.

Fuente 2:

- Se capitalizaron las columnas con caracteres (no numéricas): direccion_territorial, area_protegida, mes.
- Se cambiaron caracteres especiales.
- Se cambiaron los nombres de las áreas protegidas (parques) por nombres sencillos fáciles de buscar en el lugar del avistamiento de la fuente 1.
- El mes se cambió a número
- Se categorizaron los datos según dos escalas diferentes: una para las columnas segmentadas por tipo de visitante y otra para las columnas con un total anual.

Segmentados	Totalizados
0: 0	0: 0
1-100: 1	1-50: 1
101-500: 2	51-500: 2
501-1000: 3	501-2000: 3
1001-5000: 4	2001-15000: 4
5001+: 5	15000+: 5

- Así se imprimió un archivo con los datos segmentados, que van del 2011 al 2014, con sus respectivos totales anuales, y otro con los totales anuales incluyendo años donde no se tienen datos mensuales, sino solo anuales. Para ello, se sumaron todos

los datos mensuales y se eliminó la columna de mes. Hubo entonces **dos archivos generados** para la fuente 2 de datos.

Por último, se unieron los archivos generados utilizando el parque como llave. El archivo generado de fuente 1 con cada archivo generado de la fuente 2. Se generó un tercer archivo únicamente con los datos del avistamiento (la especie y taxonomía), la región y el parque. Estos archivos se resumen aquí en el orden que fueron mencionados:

Archivo	merge_fuente1_segmentos	merge_fuente1_totales	merge_fuente1_totales_animales
Filas	10800	900	900
columnas	50	31	11

3. Análisis de Calidad de los Datos

Para determinar la calidad de los datos de las bases de datos con las cuales se planea trabajar, se utilizó la herramienta pandas profiling en Python.

Fuente 1: Biodiversidad en Colombia - Universidad Nacional

El resumen de esta base de datos se ve como indica la siguiente figura:

Dataset info

Number of variables	44
Number of observations	109270
Total Missing (%)	0.0%
Total size in memory	36.7 MiB
Average record size in memory	352.0 B

Variables types

Numeric	13
Categorical	19
Boolean	0
Date	0
Text (Unique)	0
Rejected	12
Unsupported	0

Utilizando la información del reporte, se encontró que hay 12 columnas que tienen un valor constante. Es decir, no se deberían tomar en cuenta para el análisis final porque no aportan ningún valor de importancia significativa en los datos. No tiene ninguna fila repetida, por lo cual se puede determinar que solo hay una especie por cada una de las filas. Adicionalmente, se encontró que hay tipos de variables numéricas y categóricas.

Entre el país de residencia de la especie, se encontró que hay algunos registros que no pertenecen a Colombia. Por esta razón, se va a tener en cuenta que tienen que ser eliminados para que no agreguen ruido al objetivo de negocio.

countrycode
Categorical

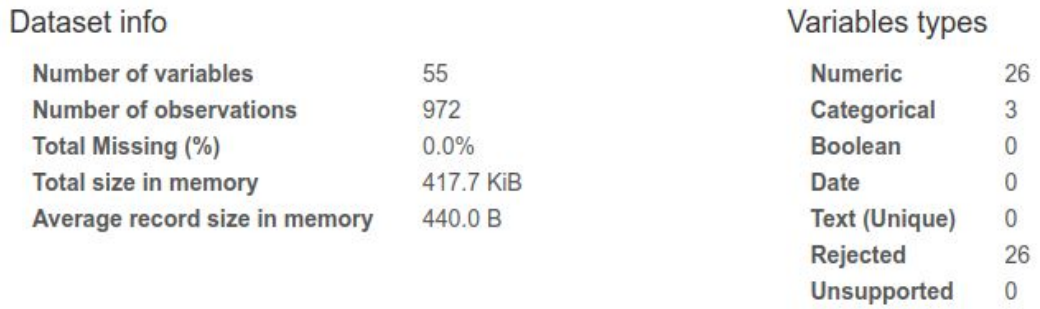
Distinct count	33
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

CO	108516
EC	176
BO	143
Other values (30)	435

De igual forma, se encontraron varias columnas que tienen una gran cantidad de datos nulos. Debido a que la mayoría de filas contienen esta información en nulo, para el análisis final de los datos se deberían eliminar. Por ejemplo, las columnas de “identified by” o “infra specific epithet”, que para efectos de la fuente, son datos opcionales.

Fuente 2: Ingreso de Visitantes en Parques Naturales

El resumen de esta base de datos se encuentra en la siguiente figura:



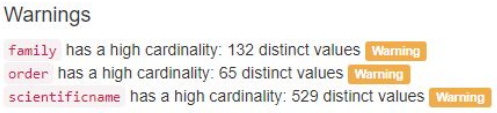
Adicionalmente, se encontró que el dataset tiene un total de 648 filas duplicadas. Dado que el dataset está desglosado por parque natural y por mes en términos de filas, para las columnas de totales anuales no tiene sentido hacer una suma mensual. Como una estrategia para salvaguardar la integridad de los datos del documento, estas columnas únicamente tienen datos en el mes de enero. Esto facilita hacer operaciones como, por ejemplo, sumar datos mensuales, pero dificulta la lectura de la tabla (sobre todo teniendo en cuenta los valores repetidos) y es un problema de calidad de datos ya que se tienen dos dimensiones inherentemente incompatibles. En el ejemplo anterior de El Cocuy, las columnas de los totales anuales muestran lo siguiente (se tomaron solo algunas columnas):

direccion_territorial	area_protegida	mes	total_2007	total_2008	total_2009	total_2010
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Junio	0	0	302	200
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Junio	0	0	302	200
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Junio	0	0	302	200

Los años a partir del 2009 muestran totales por mes. Los demás años totalizan el número de visitantes en enero:

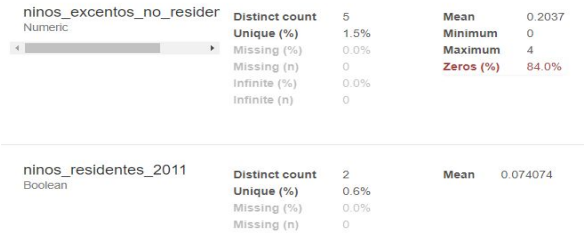
direccion_territorial	area_protegida	mes	total_2007	total_2008	total_2009	total_2010
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Enero	3073	3639	3282	4174
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Enero	3073	3639	3282	4174
DIRECCION TERRITORIAL ANDES NOR	PNN El Cocuy	Enero	3073	3639	3282	4174

Como se mencionó anteriormente, hay datos nulos que deben ser filtrados, ya que hay ceros reales (por ejemplo, que no haya habido niños extranjeros no residentes exentos de tarifa de entrada durante un mes específico), pero un cero en el total de visitantes de un parque natural nacional en un año entero no es significativo.



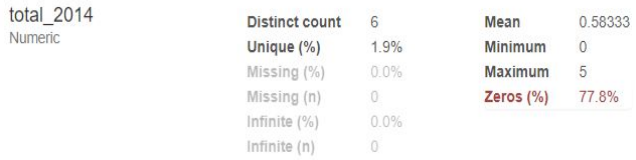
El análisis luego del preprocesamiento de datos es más favorable. Para la fuente 1 no hubo mayor advertencia, solo la alta cardinalidad de los datos:

Para la fuente 2 hay varias advertencias por un alto número de ceros en los datos. Sin embargo, estos ceros son tanto reales como falsos. Un ejemplo de un candidato de cero real es el siguiente:



El número de niños exentos no residentes que asiste a un parque natural es muy específico, y pueden pasar meses sin que haya datos para esta columna. La columna de niños residentes para el mismo año no tiene ceros, por otro lado.

Un candidato para un cero falso es el siguiente:



Hay ceros en los datos del total de visitantes anuales en el 2014. Esto corresponde a parques poco conocidos cuyos datos puede que no sean de la calidad más alta. Puede que sea, sin embargo, debido a que algún parque cierra sus puertas por temporadas.

En el segundo archivo resultante de la fuente 2 (fuente_2_procesada_totales) tiene menos incongruencias, y la mayor cantidad de ceros se borran porque se suma anualmente, así se borran los ceros por falta de datos o porque el parque realmente no abra en ciertas temporadas.
Las advertencias son únicamente por correlación:

Sin embargo, los datos para 1995 tienen ceros de todas maneras. Estos son ceros reales probablemente.

4. Resolución del Problema de Negocio

Oportunidad/problema de negocio	El ecoturismo en Colombia ha estado en gran auge en los últimos años. Los viajeros se están interesando en poder ir a diferentes parques naturales y una de las motivaciones es el avistamiento de varias especies de la flora y la fauna. Tanto compañías de viajes como el ministerio de turismo quiere encontrar la fuente de los datos recolectados de avistamientos de flora y fauna en el país. Por esta razón, decidieron contratar a un equipo especializado que permita desarrollar un modelo predictivo en un piloto de 27 parques naturales.	
Descripción del requerimiento desde el punto de vista de minería de datos	Segmentar los datos para encontrar las variables que determinan el avistamiento de las especies encontradas en las bases de datos.	Clasificar los parques según el avistamiento de las especies e información adicional que disponga el usuario.

Detalles de la actividad de minería de datos		
Tarea	Técnica	Algoritmo y parámetros utilizados
Agrupación por afinidad	Clusterización no supervisada	K means, con K = 5 para generar cinco centroides sobre los datos
Clasificación	Árbol de decisión	C4.5 donde sus parámetros se definen completamente en las siguientes secciones del documento

5. Resultados de Modelos Analíticos

Clasificación No Supervisada

Se quería encontrar entre todas las variables del dataset los que generan una mayor variación a la hora de poder avistar una ave. Como no es claro que es lo que se quiere, se tomó la decisión de llevar a cabo una clasificación no supervisada utilizando el algoritmo de K-means. El parámetro fundamental de esta técnica es el número de clusters, o número de grupos que se quieren armar. Se decidió tener un número no tan grande de segmentos debido a que se querían encontrar las variables que tienen una mayor influencia.

Para ello, se utilizó la implementación en Python de K-means y se utilizaron las variables numéricas. Aquellas que eran categóricas y eran fundamentales como por ejemplo el parque se categorizaron en un espacio numérico. El archivo *"kmean_results.csv"* contiene los resultados realizados al ingresar todos los datos al modelo y correrlo con un K=5. Cabe resaltar, que los resultados involucraron que las variables más influyentes en los datos son el mes y la llave única de cada especie.

Clasificación Supervisada

Para poder predecir el parque según las especies y la región se determinó que la mejor estrategia es realizar un árbol de decisión. Para ello, se utilizó la herramienta Knime que permite cargar la base de datos como tal y generar un flujo que realiza tanto el entrenamiento como la prueba del clasificador. Los parámetros significativos de este algoritmo solamente incluyen los datos, lo demás se determinó como el default de la aplicación.

Se desarrollaron dos clasificadores que varían en los datos de entrada. En el primer clasificador se tuvieron en cuenta todas las variables presentes en el archivo resultante del preprocesamiento merge_fuente1_segmentos. En cambio, el segundo clasificador se generó utilizando la información de las especies y la región, presente en merge_fuente1_totales_animales, donde se decidió dejar de lado la información sobre los visitantes de cada uno de los parques naturales.

En un análisis superficial se encontró que los factores más determinantes para poder realizar una buena clasificación son la región natural, el tipo y el número de visitantes. De igual forma, el segundo clasificador se comporta de tal manera que realiza una predicción muy similar en precisión que el primer clasificador, utilizando la región natural y la taxonomía. Por último, los valores obtenidos de estos, permiten analizar que el error de la clasificación es muy bajo. Por lo que el método es capaz de generar una clasificación correcta con alta precisión.

6. Análisis de los Modelos

Análisis del modelo de clasificación no supervisada

Al realizar la segmentación de los datos, se encontró que se pudieron generar cinco clusters de espacio homogéneo en la nueva representación vectorial. El algoritmo arrojó un número entre 0 y 4 para referenciar el segmento al que pertenece cada tupla de la base de datos. Este dividió de forma uniforme las tuplas, por lo cual hay aproximadamente 180 tuplas por cada una de las categorías que encontró.

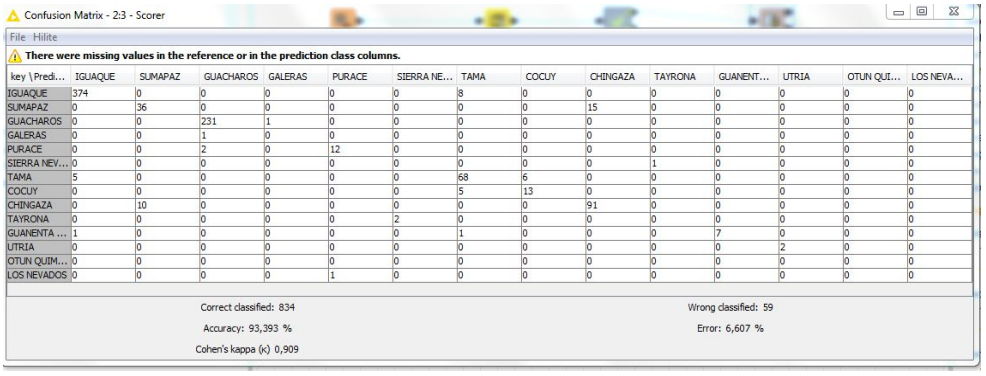
	A	C	D	E	F	AA
1		key	month	taxonkey	year	kmeans
2	0	1	6.0	5281044	2001.0	0
3	1	1	12.0	5288981	2002.0	0
4	2	1	5.0	3032536	2003.0	1
5	3	1	12.0	5287810	2002.0	0
6	4	1	12.0	5291985	2002.0	0
7	5	1	12.0	2717987	2002.0	1
8	6	1	12.0	6127	2002.0	3
9	7	1	6.0	2669988	2001.0	1
10	8	1	6.0	2689384	2001.0	1
11	9	1	12.0	2694600	2002.0	1
12	10	1	12.0	2695281	2002.0	1
13	11	1	12.0	2695281	2002.0	1
14	12	1	12.0	2694820	2002.0	1
15	13	1	12.0	2694820	2002.0	1
16	14	1	6.0	2679471	2001.0	1
17	15	1	6.0	2679471	2001.0	1
18	16	1	12.0	4659	2002.0	3
19	17	1	6.0	9043960	2001.0	2
20	18	1	6.0	9043960	2001.0	2
21	19	1	7.0	8311593	2001.0	2
22	20	1	6.0	2670289	2001.0	1
23	21	1	6.0	2670289	2001.0	1
24	22	1	7.0	2689032	2001.0	1
25	23	1	12.0	8079242	2002.0	2
26	24	1	12.0	4673	2002.0	3

Como ya se mencionó con anterioridad, los datos más significativos sobre los clusters y la similitud son la especie y el mes que se avistó. A continuación se muestran algunos de los resultados encontrados para este modelo.

En este caso, se puede observar que el cluster con referencia 1 está relacionado con la llave taxonómica que está alrededor del 26,000,000. Mientras que el cluster 3 se encuentra más identificado con las tuplas que se encuentran alrededor del 4,000. De esta forma, se puede determinar que el avistamiento es una parte fundamental de la segmentación de los datos. Para la organización esto tiene un valor fundamental por lo que determina que hay grupos de especies que se encuentran relacionados entre sí. Lo que involucra que un visitante puede estar interesado en el avistamiento de uno y por ende se encuentre dispuesto a ver el otro que está “cerca” de este primero. Adicionalmente, con los datos de la especie se pueden realizar otro tipo de análisis como el siguiente método propuesto por el equipo.

Análisis del modelo de clasificación supervisada

El primer árbol de clasificación se realizó con los datos segmentados y no segmentados. La diferencia fue muy pequeña, un error de menos del 5%, por lo que se optó por utilizar el *dataset* con la información completa, datos segmentados por tipo de visitante, total de visitantes y la información del avistamiento.



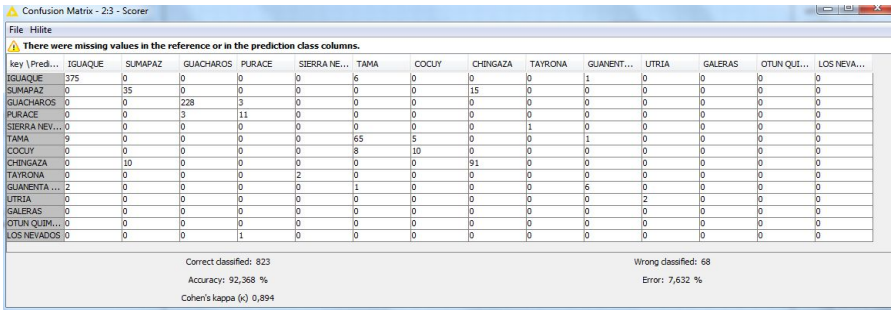
Este modelo logró clasificar el parque en un 93,4% de los casos, clasificando correctamente 834 parques de los 893. La estrategia que utilizó el árbol fue decidir basándose principalmente en los siguientes campos (en ese orden):

- Dirección territorial
- Tipo de visitante

- Temporada de visitas
- Total de visitantes
- Taxonomía

El segundo árbol de clasificación se realizó, como se mencionó anteriormente, con un archivo que solo contiene la información del avistamiento (información taxonómica), y la región. Se dejaron de lado los demás datos porque se buscaba una predicción con esas restricciones. Clasifica de manera correcta un menor número de avistamientos, pero se comporta muy bien de todas maneras:

Clasificó de manera correcta un 92,4% de los casos. Los valores usados para la predicción fueron la región y la taxonomía, en ese orden de importancia.



key \ Predi...	IGUAQUE	SUMAPAZ	GUACHAROS	PURACE	SIERRA NE...	TAMA	COCUY	CHINGAZA	TAYRONA	GUANENT...	UTRIA	GALERAS	OTUN QUI...	LOS NEVA...
IGUAQUE	375	0	0	0	0	6	0	0	0	1	0	0	0	0
SUMAPAZ	0	35	0	0	0	0	0	15	0	0	0	0	0	0
GUACHAROS	0	0	228	3	0	0	0	0	0	0	0	0	0	0
PURACE	0	0	3	11	0	0	0	0	0	0	0	0	0	0
SIERRA NEV...	0	0	0	0	0	0	0	0	1	0	0	0	0	0
TAMA	9	0	0	0	0	65	5	0	0	1	0	0	0	0
COCUY	0	0	0	0	0	8	10	0	0	0	0	0	0	0
CHINGAZA	0	10	0	0	0	0	0	91	0	0	0	0	0	0
TAYRONA	0	0	0	0	2	0	0	0	0	0	0	0	0	0
GUANENTA ...	2	0	0	0	0	1	0	0	0	6	0	0	0	0
UTRIA	0	0	0	0	0	0	0	0	0	0	2	0	0	0
GALERAS	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OTUN QUIM...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LOS NEVADOS	0	0	0	1	0	0	0	0	0	0	0	0	0	0

7. Estrategias para la Organización

Con esta información, la organización puede poner en marcha los planes iniciales, planteados en los puntos 0 y 4. La prueba piloto demostró que sí se puede clasificar los avistamientos según su ubicación y el momento en que fueron tomados para decidir sobre el parque donde se realizó el avistamiento. Por ello, las estrategias que debería tomar la organización son poner en marcha mecanismos para fomentar el turismo de avistamiento, incentivos para que las personas registren esos avistamientos en la plataforma, y campañas de preservación de estas especies. Según el análisis, especies relacionadas se pueden encontrar en el mismo parque, por lo que se puede promocionar el avistamiento no solo de una especie sino de varias que estén relacionadas y se sepa que están en el mismo parque. El Instituto Humboldt podría utilizar estos datos para generar una base de datos colaborativa entre turistas e investigadores.

Bibliografía

Raz L, Agudelo H (2016). ICN - Universidad Nacional de Colombia. Version 2.2. Universidad Nacional de Colombia. Occurrence Dataset https://doi.org/10.15472/v2lnzj accessed via GBIF.org on 2018-03-04.