Identificación de patrones submarinos en especies de macroinvertebrados de monitoreo ecológico

Eduardo Grande

Carlos Peñarubia

Ahmed Begga

Minería de Datos
Máster en Ciencia de Datos
Curso 2022-2023

Máster en Ciencia de Datos
Universidad de Alicante

Índice

1. Introducción

La fundación Charles Darwin realiza incursiones en todas las islas que conforman las Galápagos. El fondo submarino se encuentra dividido en diferentes partes, diferenciando la isla, la región, y el aspecto más concreto sería el transecto. Un transecto es una pequeña porción de fondo marino, identificada con un código único.

En cada incursión que se realiza, se anotan todos los datos de la ubicación dónde se realiza el avistamiento, así como los datos de la especie marina observada.

Estas observaciones se realizan a lo largo de varios años, para así poder tener una visión global del fondo marino. Fenómenos meteorológicos, cambios en el ambiente y otros fenómenos pueden cambiar la distribución de especies, por tanto, tener todos los datos bien clasificados y estructurados puede permitir a largo plazo ver los cambios que se han producido y predecir que puede pasar en el futuro.

En el presente trabajo se analizarán los datos proporcionados por la fundación Charles Darwin, los cuales contienen información del rango temporal entre 2010 y 2020. Se explicará a lo largo de la memoria el trabajo realizado y los resultados obtenidos.

2. Recolección y pre-procesamiento de datos

El punto de partida del trabajo será explorar los datos recibidos, para familiarizarse con los mismos. Una vez conocidos los datos, se procederá a realizar una limpieza.
Observando los datos, estos tienen 27 columnas, entre las que se encuentran la fecha en la que se ha realizado el buceo, el transecto, la especie avistada…
Se cuenta con 7025 filas, es decir, con 7025 observaciones a lo largo de 10 años.
Una vez conocido el dataset, se procede a realizar una limpieza del mismo. Esta limpieza se ha realizado generalmente columna a columna, si bien en alguna ocasión se ven varias columnas a la vez. Se detalla a continuación el proceso por columna:

- **Columna *id***: De esta columna se comprueban los valores únicos, para ver si hay algún valor anómalo. Se ve que todo está de forma correcta.
- **Columnas *dive_data, dive_month* y *year***: Se han estudiado estas tres columnas de forma conjunta al estar todas referidas a la fecha del buceo. Primero se comprueban los valores únicos de *year*, viendo que no hay valores anómalos, al haber valores entre 2010 y 2020. A continuación, se explora si hay filas con valores nulos en *dive_date* y en *dive_month* a la vez. Se comprueba que no hay ningún caso de estos. Dónde sí que hay filas con valores nulos es si se buscan solo en *dive_date*, por tanto, esas filas se eliminan al no tener información relativa respecto a la fecha de la incursión, ni tampoco poder ser deducida de otras columnas. Además, son pocas filas, 93 concretamente, por lo que la pérdida de información no es relevante (apenas un 1,3%). Además, se ha convertido la columna *dive_date* al tipo *datetime*, para que pueda ser más fácilmente manejable. Una vez hecho, se puede comprobar si el mes extraído de esa columna (*dive_date.dt.month_name()*) se corresponde con el mes *dive_month*. Se comprueba que en todos los casos coincide, por tanto, todo está de forma correcta. La columna *dive_month* ya no aporta información adicional, al poder como se ha visto sacar el mes de *dive_date*, por tanto, se elimina esa columna. Se realiza el mismo proceso pero en este caso con el año, comprobando si el valor de *year* coincide con el de *dive_date.dt.year*. En este caso, hay 23 registros en los que el año no coincide. A pesar de eso, se ha podido regularizar el problema, ya que estos datos están todos rodeados de datos de 2011, por tanto, se entiende que corresponden a ese año, ya que los datos de 2020 (el otro año que aparecía, 2011 ≠ 2020), aparecen al final del dataset (todos los registros con posiciones mayores a 6300), mientras estos están en posiciones bajas (menores a 1500). Solucionado el problema, se eliminará la columna *year*.
- **Columna *island***: En esta columna se comienza comprobando si hay algún valor nulo. Se comprueba que hay 106 nulos. ¿Es posible obtener la isla de otro modo?

Las dos primeras letras del transecto corresponden a las islas de las Galápagos, por tanto, dadas esas dos letras, es posible deducir que islas son. En todos los casos de valores nulos, las islas eran las de Pinta (PI) y San Cristóbal (SB). Por tanto, se cambian los nulos por los valores correspondientes.

- **Columna *Transect_code***: Como se ha comentado en la columna anterior, las dos primeras letras de este código corresponden a las islas. Por tanto, se comprueba que para todos los valores de esta columna, el dato de la columna isla correspondiente. Se comprueba así que todos los valores coinciden. En este caso no se elimina la columna *island*, ya que pese a poder deducir la columna dado el *Transect_code*, en esa columna tenemos el nombre de la isla completa.
- **Columna *Bioregion***: Como se hace siempre, se comprueba si hay valores nulos, cosa que sí que pasa en esta columna. Todos los nulos corresponden a valores de la isla Fernandina. Se comprueba que el *Transect_code* también contiene dos dígitos (el tercero y cuarto) que corresponden a la bioregión, pero en este caso, no se han podido obtener los valores correspondientes ya que los códigos de las bioregiones que faltan no aparecen en ninguna fila con información completa. Se dejarán los valores nulos.
- **Columna *MPA_Status***: Se comprueba que esta columna no tiene valores nulos. Ahora, se comprueba que para cada isla y para cada transecto (4 primeros dígitos de *Transect_code*), siempre el status de área protegida sea el mismo (de lo contrario, para un mismo transecto e isla, no puede ser que el estado sea diferente). Se comprueba que todos los valores son consistentes.
- **Columna *Sum_ind***: Para esta columna, se comprueban los valores mínimos y máximos, así como valores nulos. No hay nulos, y los valores mínimos y máximos están dentro de unos límites razonables, por lo que no se aprecian valores anómalos.
- **Columnas *TaxonID*, *Domain*, *Kingdom*, *PhylumOrDivision*, *Class*, *Order*, *Family*, *ScientificName*, *CommonNameEnglish*, *CommonNameSpanish***: Se han estudiado todas estas columnas juntas ya que todas ellas son elementos de la taxonomia de los animales. Para cada identificador, se cuenta con un valor para cada columna (excepto para *CommonNameSpanish*, que puede tener más de un valor). En muchas ocasiones, solo se cuenta con el *ScientificName*, por lo que se ha buscado la forma de completar el resto de aspectos. Para ello, **se hace uso del *Datazone* de la web de la Charles Darwin Fundation**. Se ha analizado la web para saber cómo realizar peticiones y obtener así todos los campos que faltan, dado el *ScientificName*. Se ha creado una función que permite eso, realizando peticiones *HTTPS* a la web. Por ello, se gira sobre las más de 340 filas a las que les falta algún

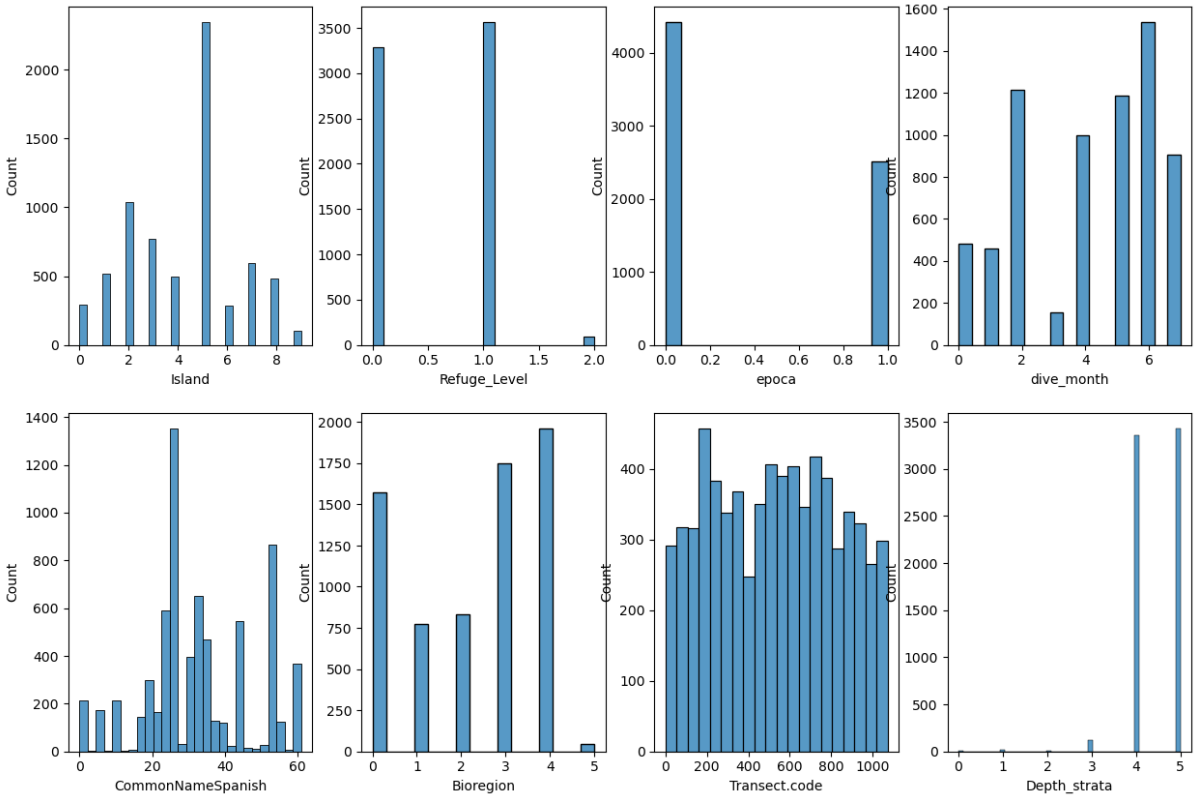
valor de las diferentes columnas de la taxonomia, llamando a la función definida, obteniendo los datos de la web, y completando así los datos. De todas esas filas, se consiguen rellenar más de dos tercios, siendo el tercio restante errores, de los cuales algunos se han podido solucionar manipulando manualmente el *ScientificName* y poniendo uno existente (en algún caso estaba puesto de forma abreviada y por eso no se encontraba). Tras todo el procesamiento, se tienen apenas 100 filas con errores, las cuales se eliminan al faltar de ellas bastante información.

- **Columna *Site***: Se comprueba si hay valores nulos (no hay) y se exploran los valores únicos de esta columna para comprobar que no haya valores anómalos (no se han encontrado).
- **Columna *Latitude* y *Longitude***: Se establece un rango de latitud y longitud en la que se sitúan las islas, y después se comprueba que todos los valores estén dentro de ese rango. Se comprueba que no hay ningún valor erróneo.
- **Columna *Subzone.name***: Se comprueban los valores únicos, viendo que no hay ningún anómalo, si bien hay algún valor nulo. Se comprueba que todos las filas que tienen esta columna nula también tienen nula la columna *Bioregion*, además de tener nula la longitud y latitud. Debido a tanta fila de información en esas filas, se eliminan.
- **Columna *Refuge_Level***: Se comprueban los valores únicos, viendo que no hay ningún valor anómalo, pero si nulos. Se exploran las filas donde hay nulos, intentando ver si se puede deducir de alguna forma la información faltante, pero como no se consigue, se eliminan las aproximadamente 40 filas.
- **Columna *depth_strata***: Se comprueba los valores únicos, viendo en este caso que hay un valor anómalo, el '-', valor que se puede interpretar como que falta especificar la profundidad. Por tanto, se exploran esos datos viendo si se puede deducir la profundidad de otro modo, pero como no se consigue, se eliminan las 8 filas.
- **Columna *epoca***: Se comprueban los valores únicos, viendo que no hay ninguno anómalo o nulo.

Una vez realizado todo este proceso, se genera un nuevo archivo con los datos limpios, para que puedan ser usados en fases posteriores.

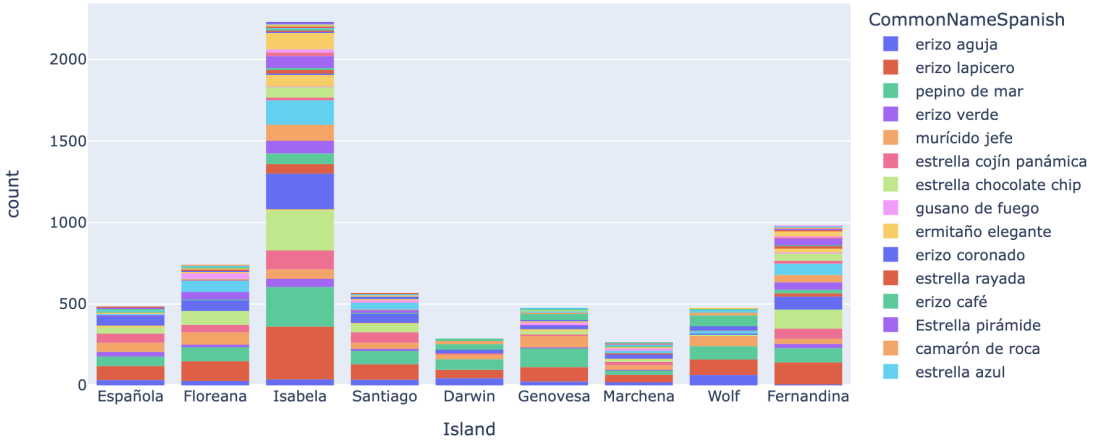
3. Visualización y Distribución de los datos

Para el apartado de visualización, hemos realizado el típico análisis exploratorio que se puede hacer en los datos en machine learning. Concretamente hemos realizado el plot de la matriz de correlación de las variables, obteniendo la siguiente figura:

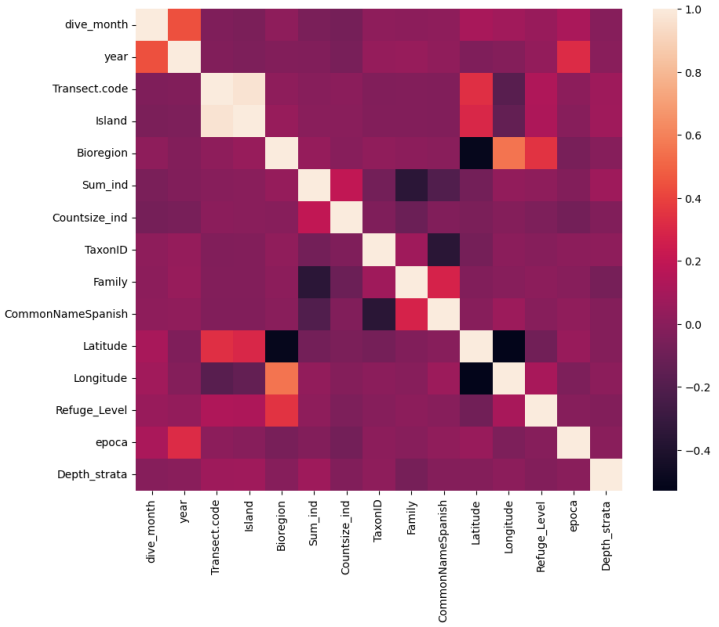


Donde podemos ver que en la quinta isla (Isabela) es la que más animales tiene, en lo que los niveles de refugio, podemos ver que el tipo de refugio (santuario/Extractive use) son similares, es decir, es una zona rica en santuarios. También podemos ver que los datos eran más propensos en obtenerse en épocas calientes que en frías y sobre todo en meses de verano [5,7). El animal del que se tiene más registro (CommonNameSpanish) es el erizo lapicero con 1063 muestras.

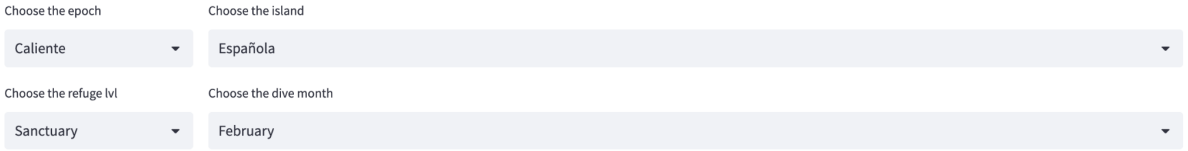
En lo que respecta a distribuciones que dependen de dos variables, como puede ser el animal y la isla, hemos graficado la siguiente figura, donde se puede ver en cada una de las islas la densidad de animales que hay en ella. Por ejemplo, en la isla Darwin, abundan los pepinos de mar y los erizos lapiceros.



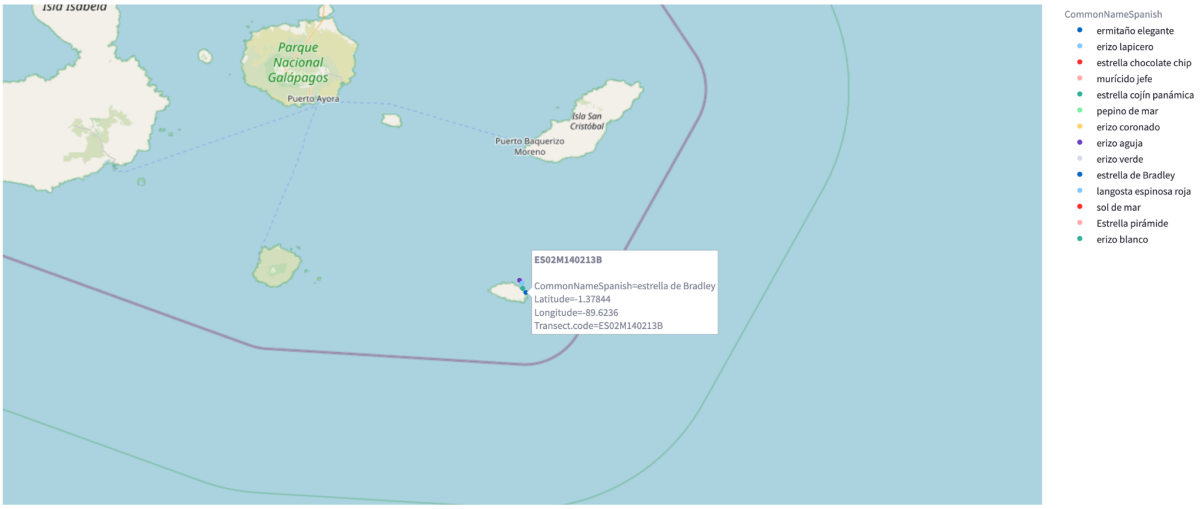
Una vez hecha la visualización de las correlaciones entre las variables seleccionadas, podemos ver como estamos en un dataset débilmente correlacionado, debido a que el mapa de calor es muy uniforme fuera de la diagonal principal. Si que hay casos en la que hay correlación negativa, como puede ser el caso de latitud y Bioregion, pero realmente no aporta ninguna información. Creemos que estas correlaciones tan débiles pueden suponer un problema a la hora de resolver cualquier tarea de clasificación. Por ejemplo, si vamos a predecir el Transect.Code, tendríamos problemas debido a que no hay apenas correlaciones, salvo con la longitud y la latitud.



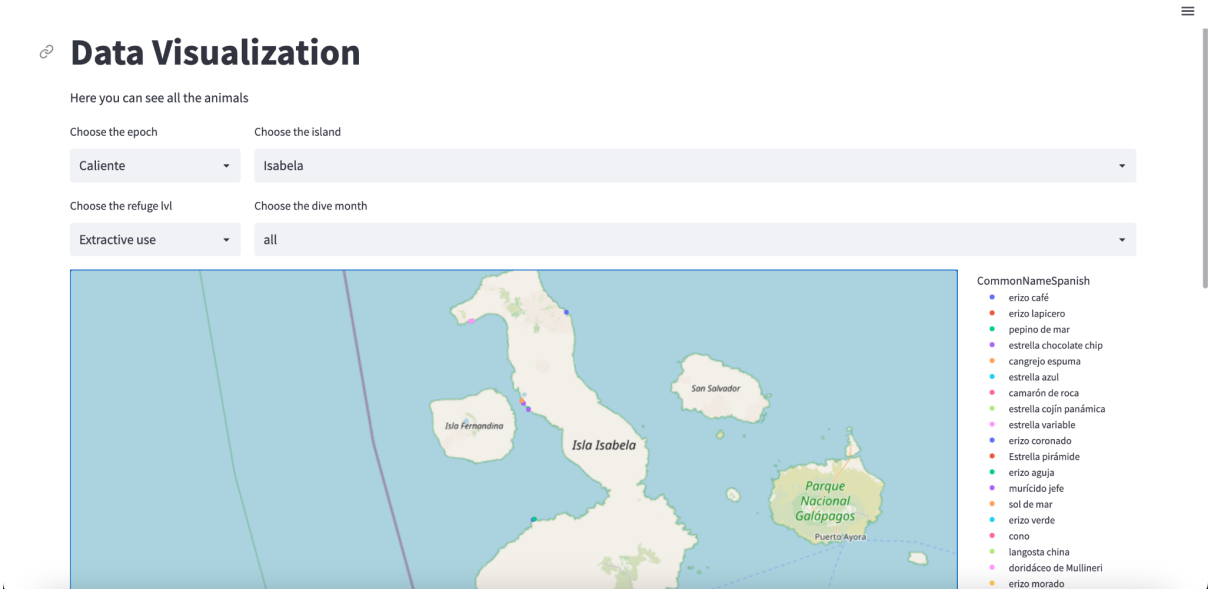
Finalmente, como se trataban de datos tabulares sobre ubicaciones, hemos decidido innovar y hacer una página web desde la cual podamos visualizar todos los datos de golpe sobre un mapa. Sobre dicho mapa, podemos visualizar que especies estamos viendo en qué época [‘Caliente’,’Fria], en que mes o isla. Además de poder filtrar sobre qué nivel de refugio se encuentra la especie:



Una vez seleccionados los parametros de busqueda, se nos mostrara el mapa con todas las muestras que cumplen dichos filtros:



Como podemos observar, si clicamos sobre cualquiera de los puntos resultantes, podemos ver que se nos despliega un cuadro de texto con información de la longitud y latitud, además del Transec.code.



Para más detalle, visita nuestra página: https://bloque-3-mineria-de-datos.streamlit.app/

4. Predicciones

Cuando se extraen patrones de comportamiento de manera eficiente se es capaz de predecir dada una especie en qué zona o región se encuentra una especie en concreto. Por ello, en esta sección nos vamos a centrar en en un modelo predictivo, en el que dada la especie y el mes del año, nos prediga en qué zona de las islas galápagos se encuentra. Existen distintas variables que hacen referencia a zonas de las islas galápagos, como el transecto, la isla, o la bioregión. Mediante distintos modelos predictivos, comprobaremos si mediante la especie y el mes de año se es capaz de predecir en qué zona se encuentra dicha especie, viendo así si hay una mayor preponderancia a que aparezca en una zona dada, y ver con qué granularidad se puede predecir la zona específica en la que habita dicha especie.

Se van a usar dos tipos de modelos para predecir, uno basado en árboles de decisión y otro basado en redes neuronales. Los modelos basados en árboles de decisión presentan una clara ventaja, y es que son altamente explicables, por lo que dado el modelo, veremos qué decisiones toma para predecir una zona u otra. Por otra parte, los modelos basados en redes neuronales carecen de explicabilidad, por lo que dada una entrada es muy difícil de determinar qué parte de la red hace que se produzca su correspondiente salida. A cambio suelen presentar rendimientos más altos en tareas complejas.

Por tanto, para las variables explicativas especie (TaxonID) y mes (dive_month) probaremos a predecir las variables transecto (Transect.code), isla (Island) y bioregión (bioregion), de mayor a menor granularidad respectivamente.

Cabe destacar que ambas variables explicativas las hemos codificado, correspondiendo para la variable mes las siguientes etiquetas

ID	Mes
0	abril
1	agosto
2	febrero
3	julio
4	junio
5	marzo
6	mayo
7	noviembre

● Árboles de decisión

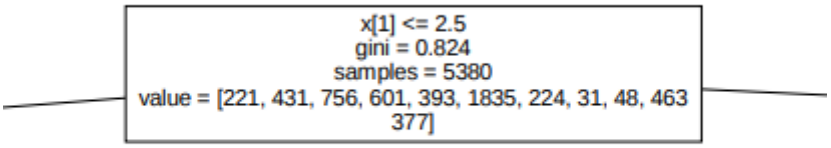
En los árboles de decisión se ha usado una partición de datos de 80% train y 20% test. Además, en el árbol de decisión, para que una rama se divida la impureza que elimine dicha división debe superar 0.0001. De esta manera evitaremos el excesivo overfitting.

Transecto
Mediante árboles de decisiones el accuracy que se obtiene a la hora de predecir el transecto es del 0%. Esto se debe a que existen 1058 transectos, y resulta muy complicado, que la mayoría de una especie dado un mes habite mayoritariamente un transecto específico, es decir, habite en un cuadrado de 1x5 metros.

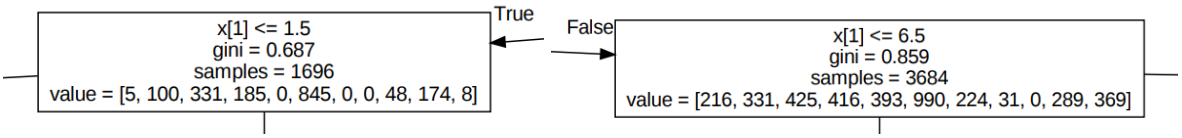
Isla
El accuracy que se obtiene a la hora de predecir la isla es del 39%. Este resultado se podría considerar relativamente bueno, ya que una especie en un mes concreto puede estar en varias islas a la vez, por lo que tener un 100% de accuracy resultaría imposible debido a que todos los clasificadores asumen que las clases son excluyentes, es decir, dada una muestra solo se le atañe una clase, no varias. De hecho, atendiendo a nuestro dataset, suponiendo que dada una especie y un mes se clasifica a la isla en la que más veces aparece dicha especie en dicho mes, el máximo de accuracy que se podría obtener es de 44.33%, por lo que el margen de mejora es pequeño y el árbol de decisiones generaliza suficientemente bien.

	precision	recall	f1-score	support
0	0.11	0.04	0.06	47
1	0.23	0.20	0.21	107
2	0.36	0.06	0.10	196
3	0.26	0.26	0.26	148
4	0.22	0.30	0.26	79
5	0.46	0.80	0.59	477
6	0.25	0.04	0.07	51
7	0.00	0.00	0.00	9
8	0.33	0.07	0.12	14
9	0.30	0.08	0.12	119
10	0.35	0.28	0.31	98
accuracy			0.39	1345

Atendiendo al árbol de decisión generado, se destacan las siguientes decisiones:
La primera decisión que toma el árbol, es decir, aquella que elimina mayor impureza, es comprobar si la variable mes es menor que menor que 2.5, es decir, si estamos en los meses de abril, agosto o febrero o no.



En los dos nodos hijos se vuelve a comprobar a qué meses corresponde la muestra que se quiere predecir. Esto nos deja ver que claramente hay una vinculación entre el mes del año y las especies, ya que dependiendo de la temporada del año una misma especie podrá migrar entre islas y ser más preponderante en una zona en concreto.

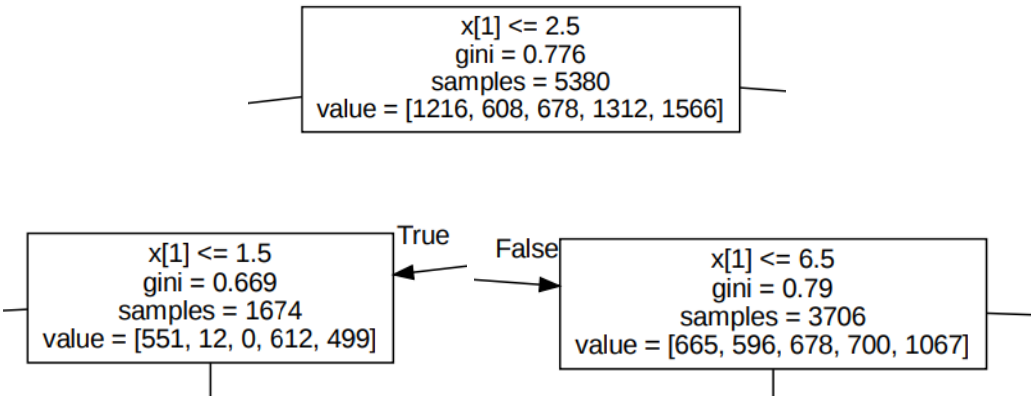


Bioregión

El accuracy que se obtiene a la hora de predecir la bioregion es de 43%. Este resultado es ligeramente mejor que el que se obtiene con respecto a las islas. Sin embargo, con las islas teníamos 11 posibles clases (una por isla), y con las bioregiones 5 (Bahía Elizabeth, Lejano Norte, Norte, Oeste, Sureste), es decir, que aunque se haya reducido en más de la mitad las posibles clases, no ha aumentado el accuracy en gran medida. Esto es debido potencialmente a que igual que antes, una especie en un mes concreto puede estar en distintas bioregiones. En el caso de las bioregiones, el accuracy máximo que se puede obtener usando como variables explicativas la especie y el mes es del 52%, por lo que sigue siendo bastante limitado. Esto a su vez nos puede vislumbrar que muchas especies se muevan atendiendo a las islas en las que se encuentran (ya que cada isla presentará unas características favorables para dicha especie en una temporada concreta) y no con respecto a la zona de la bioregión.

	precision	recall	f1-score	support
0	0.43	0.30	0.36	324
1	0.47	0.53	0.50	159
2	0.29	0.20	0.23	147
3	0.47	0.36	0.41	355
4	0.38	0.60	0.46	360
accuracy			0.41	1345

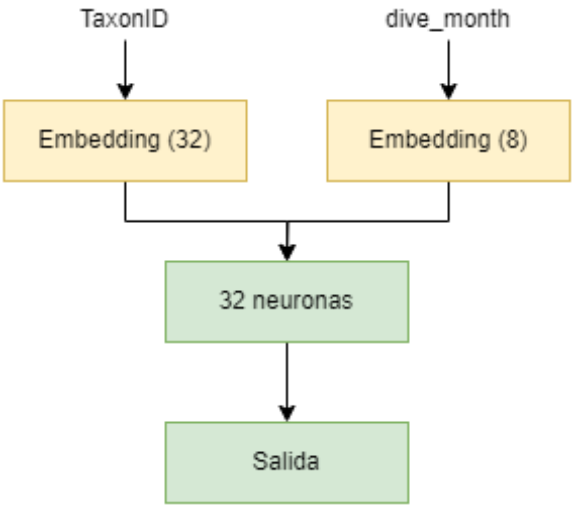
Con respecto al árbol de decisiones generado, cabe destacar que las decisiones del comienzo del árbol son idénticas. Esto nos hace ver que claramente hay una estrecha relación entre la temporada del año y la zona que habitan distintas especies.



- **Deep Learning**

Lo que se busca con este modelo es dilucidar si con un técnicas de aprendizaje supervisado más potentes se consigue mayor rendimiento, o por el contrario, debido a las características del dataset, con un árbol de decisiones es suficiente y es capaz de sacar el máximo partido de los datos.

Puesto que las variables TaxonID y dive_moth son categóricas, se van a usar dos embeddings, de tamaño 32 y 8 respectivamente. En la naturaleza existen especies que tienen comportamientos y patrones parecidos, al igual que existen meses que tienen efectos en las especies semejantes. Por tanto, el objetivo es que los embeddings extraigan características y agrupen especies y meses parecidos. Estos embeddings se unen y pasan por una capa oculta de 32 neuronas, para finalmente producir la salida con tantas neuronas como clases tenga la variable a predecir. La función de activación usada es ReLU, menos la capa de salida que es softmax. La función de pérdida es categorical cross entropy y el optimizador Adam con un learning rate de 0.0003. Se ha usado además un early stopping de 10 de paciencia.



Transecto
De la misma manera que con árboles de decisiones y siguiendo las mismas razones, el accuracy que se obtiene es de un 0%.

Island

El accuracy obtenido a la hora de predecir la isla es del 40%. Como se puede observar, el resultado es muy parecido. Esto nos indica que para este problema, usar técnicas más avanzadas de machine learning no proporciona apenas mejora. Además, al usar estos métodos estamos perdiendo por completo la explicabilidad del modelo, por lo que usarlos para este caso no es recomendable, sería más adecuado quedarnos con los árboles de decisión, que mediante su explicabilidad nos revela que factores son los más importantes y los que mayor peso tienen a la hora de producir el resultado.

	precision	recall	f1-score	support
0	0.10	0.02	0.04	45
1	0.28	0.18	0.22	110
2	0.33	0.03	0.05	195
3	0.33	0.24	0.28	162
4	0.27	0.38	0.31	93
5	0.45	0.86	0.59	447
6	0.12	0.01	0.03	71
7	0.00	0.00	0.00	7
8	0.00	0.00	0.00	18
9	0.39	0.09	0.14	104
10	0.35	0.46	0.40	93
accuracy			0.40	1345

Bioregión

El accuracy obtenido a la hora de predecir la bioregión es de 46%. De la misma manera que usando árboles de decisiones, los resultados son muy parecidos, además de no obtener una mejora significativa con respecto a cuando se usan las islas para predecir, aunque este valor de accuracy sí que está más cercano al 52% de accuracy máximo que se puede lograr a obtener.

	precision	recall	f1-score	support
0	0.46	0.38	0.42	313
1	0.49	0.60	0.54	138
2	0.36	0.27	0.31	171
3	0.49	0.39	0.43	326
4	0.46	0.62	0.53	397
accuracy			0.46	1345

5. Discusión y conclusión

A lo largo de este trabajo hemos alcanzado con éxito los objetivos planteados en la práctica, habiendo realizado una limpieza de los datos atendiendo a que los valores sean consistentes y no tengan nulos, un análisis exploratorio de los datos visualizando en un mapa de las islas galápagos la distribución de las especies atendiendo a distintos parámetros pudiendo de ésta manera extraer de manera interactiva y visual patrones de comportamiento, y finalmente realizando modelos que puedan dada una especie y una época del año predecir en qué zona estará dicha especie con mayor probabilidad.

Como se ha visto, extraer patrones de comportamiento es una tarea compleja que requiere de un conjunto amplio y consistente de datos, por lo que la correcta información y limpieza de los datos es crucial para capturar de manera efectiva cambios en dichos patrones a lo largo del tiempo, y sacar las correctas conclusiones con respecto a los problemas actuales, como la pérdida de biodiversidad o la injerencia humana en los hábitats de los animales, afectando al ecosistema en su conjunto. Además, para que un modelo aprenda y sea aplicable y útil en casos reales, se requiere que aprenda sobre datos correctos. Por otra parte, mediante el análisis de los modelos también se pueden extraer conclusiones muy valiosas sobre los datos. Por ello, en nuestro caso hemos utilizado árboles de decisiones, ya que presentan una alta explicabilidad. Posteriormente, mediante técnicas de redes neuronales nos hemos centrado más concretamente en mejorar la calidad de las predicciones, dejando la explicabilidad en un margen. Por último, el análisis exploratorio de los datos de manera visual es clave. Con un mapa interactivo se consigue extraer información y valor de los datos por sí mismos, sin la necesidad de complejas transformaciones o métodos estadísticos.

En líneas futuras se podría plantear el uso de métodos de clasificación multietiquetas, de manera que se pueda predecir dada una especie y época todas las zonas en las que ésta pueda aparecer. También se podría ampliar las capacidades de visualización del mapa, añadiendo datos sobre animales terrestres, y pudiendo comprobar así si hay interacciones entre los animales terrestres y marinos.