IA para la predicción de la edad de los Abalones

I. Introducción

El abulón o abalón es un molusco gasterópodo de la familia Haliotidae. El calculo de la edad de uno estos moluscos se lleva a cabo a través de un proceso laborioso y tedioso el cual consiste en cortar a un ejemplar por la espira y se lija hasta exponer los anillos internos y posteriormente estos se cuentan con la ayuda de un microscopio. Se puede usar otras medidas más fáciles de obtener para predecir la edad aproximada. Esta información puede ser usada con el fin de apoyar la conservación de las especies al poder obtener estadísticas como el cambio del promedio de edad, adaptaciones a cambios en el ambiente, etc.[1]

Con el fin de realizar estas predicciones se realiza la implementación del algoritmo de Random Forest de Regresión

El algoritmo de Random Forest es un modelo conformado por la combinación de múltiples Arboles de decisión individuales, en estos se utiliza la técnica de Bagging la cual consiste en generar muestras aleatorias de los datos y usarlas para entrenar modelos independientes, de los cuales se usa el promedio del resultado de estos para producir una predicción mas precisa. Este algoritmo puede ser utilizado tanto para problemas de regresión como para problemas de clasificación.

Este algoritmo es popular debido a que reduce el riesgo de Overfitting esto debido a la toma de muestras de los datos; ofrece flexibilidad debido a que puede usarse para problemas de regresión y clasificación; y es fácil de determinar la importancia de las características que tienen mayor impacto en las predicciones.[2]

II. Dataset

El dataset proviene del repositorio UCI Machine Learning de la universidad de California, Irvine[3], el cual mantiene alrededor de 647 datasets para Machine Learning. El dataset usado para este proyecto contiene las medidas de ciertas características físicas de 4177 ejemplares de abulones, dentro de

estas se encuentra el sexo, longitud, diámetro, altura, pesos bajo diferentes condiciones y los anillos, estos últimos son usados para medir la edad.

Para el propósito de este proyecto se usaron las variables de:

- Sex_I

- length

- Diameter

- Height

- Whole_weight

- Shell_weight

Y el dataset se dividió en 70% para train y 30% para test.

III. Modelo

El modelo de Random Forest se implementó utilizando el Framwork de scikit-learn[4], utilizando la función de GridSearchCV para realizar una búsqueda exhaustiva sobre un conjunto de parámetros para obtener los hiperparametros óptimos para el modelo. Dentro de esta se configuro para integrar el proceso de Cross validation a través del proceso de k-fold cross validation para reducir la varianza de la precisión del modelo.

Esta implementación dio como resultado que los mejores parámetros son:

- Máxima profundidad: 6
- Numero de características: None
- Máximo numero de nodos hoja: 9
- Numero de árboles: 150

IV. Resultados

El promedio de puntaje de la validación cruzada del modelo fue -2.36297

Una Distancia media cuadrática mínima (RSME) de 2.46571

Un coeficiente de determinación R2 de 0.448363

Referencias:

1 Gluyas-Millán, María Georgina, & Talavera-Maya, Jesús. (2003). Composición por tallas y edades de las poblaciones de abulón Haliotis fulgens

y H. corrugata de la zona de Bahía Tortugas, Baja California Sur, México. Ciencias marinas, 29(1), 89-101. Recuperado en 11 de septiembre de 2023, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-38802003000100009&lng=es&tlng=es.

2 What is Random Forest? | IBM. (s. f.). https://www.ibm.com/topics/random-forest

3 Nash,Warwick, Sellers,Tracy, Talbot,Simon, Cawthorn,Andrew, and Ford,Wes. (1995). Abalone. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.

4 API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.