Modelo de Predicción de Especies de Peces

André Chávez Contreras
Universidad de Xalapa
Estructura de datos
Xalapa, México
ux23ii263@ux.edu.mx

***Abstract*—This document presents the implementation of a learning algorithm applied to a training dataset of fish species. The goal of the model is to accurately predict the species, weight or length of a fish based on various input features. The purpose of this project is to implement ID3 decision trees or Bayes theorem into a training dataset.**

***Index Terms*—ID3 decison trees, Bayes theorem, learning algorithms, training dataset**

I. Introducción

Este proyecto explora la implementación de algoritmos de aprendizaje (Árboles de decisión ID3, Teorema de Bayes) y aplica un modelo de aprendizaje a un conjunto de datos sobre especies de peces, con el objetivo de predecir con precisión la especie, el peso o la longitud de un pez. Mediante la implementación de un algoritmo de aprendizaje, se busca obtener resultados precisos y un modelo de clasificación eficaz, el cual puede ser adaptado para otros conjuntos de datos y resolver problemas de clasificación más grandes.

A. ¿Qué tan efectivo es el uso de algoritmos de aprendizaje para clasificar?

El propósito de conocer la efectividad de los algoritmos de aprendizaje para clasificar es comprender su capacidad para identificar patrones y hacer predicciones precisas en datos complejos, estos algoritmos pueden tener un impacto significativo en distintas áreas que necesiten la clasificación y predicción de datos con distintas características [1].

B. Objetivos

- Desarrollar modelos de aprendizaje automático que prediga la especie, peso y longitud de los peces a partir de un conjunto de datos.
- Saber si las predicciones de los modelos son parecidas a las especies reales de peces y puede mantener proporciones realistas.
- Conocer si los algoritmo implementados dan los mejores resultados para la clasificación de peces o existen algoritmos más optimos.

C. Justificación

Consideramos importante el desarrollar diferentes modelos de aprendizaje automático para predecir y clasificar, ya que de ser implementado en diversas industrias puede llegar a ser realmente útil y efectivo. Además la implementación de algoritmos de aprendizaje en este proyecto no solo contribuye al conocimiento del área, sino que también proporciona herramientas valiosas para aprender y entender de mejor manera los algoritmos de inteligencia artificial; al conocer y ejecutar distintos algoritmos y compararlos se pueden identificar las técnicas más efectivas para resolver y obtener ciertos resultados [2].

II. Marco teórico

A. Árboles de decisión ID3

El algoritmo de los árboles de decisión ID3 se basa en construir árboles de decisión para clasificar datos, basándose en la ganancia de información de cada atributo. Se utiliza en problemas de clasificación, donde las decisiones se representan como nodos y las posibles respuestas como ramas [3].

B. Teorema de Bayes

El teorema de Bayes permite calcular la probabilidad de un evento basado en información previa, en aprendizaje automático se utiliza para actualizar las probabilidades de una hipótesis conforme se añaden nuevos datos [4].

C. Algoritmos de aprendizaje

Los algoritmos de aprendizaje son métodos y secuencias que permiten a las computadoras aprender a partir de datos base. Estos algoritmos son escenciales y muy efectivos para encontrar patrones, clasificar y predecir resultados.

D. Dataset

Es el conjunto de datos utilizado para aplicar un modelo de aprendizaje automático, contiene ejemplos con datos clasificados de manera correspondiente, lo cual permite que el modelo aprenda a hacer predicciones a partir de ellos.

E. Machine learning

Machine learning es una rama de la inteligencia artificial que permite a sistemas aprender de los datos para identificar patrones y hacer predicciones sin programación explícita. [5].

F. Overfitting

El overfitting es un problema de machine learning el cual ocurre cuando un modelo de aprendizaje automático se ajusta exageradamente bien a los datos de entrenamiento, capturando anomalías y el ruido. Esto puede ocasionar que el modelo haga predicciones erróneas [6].

III. Metodología

A. Interpretación de los datos

Para realizar los modelos, se utilizó un conjunto de datos que contiene información sobre las características de diversas especies de peces. Las variables de la base de datos son: Especie, Peso (g), Longitud (cm) y la proporción entre peso y longitud. Estas cuatro variables abren las puertas a la implementación de algoritmos de aprendizaje automático de clasificación; el análisis de estos datos permite encontrar las relaciones entre las variables. Por ejemplo, la proporción entre el peso y la longitud puede ser un factor clave para determinar o predecir la especie, conforme el algoritmo se ajusta a los datos, el modelo encontrará y optimizará sus parámetros para maximizar su capacidad de reconocer patrones y elaborar predicciones precisas.

Para obtener un mejor entendimiento de los datos proporcionados por la base de datos, se elaboró la siguiente gráfica, la cual muestra en el eje x la longitud del pez en cm, y en su eje y se aprecia su peso en g (gramos), la gráfica también muestra en diferentes colores las especies, esta gráfica es de utilidad para poder entender de mejor manera los comportamientos y los patrones que tiene que reconocer el modelo para poder predecir de manera acertada.

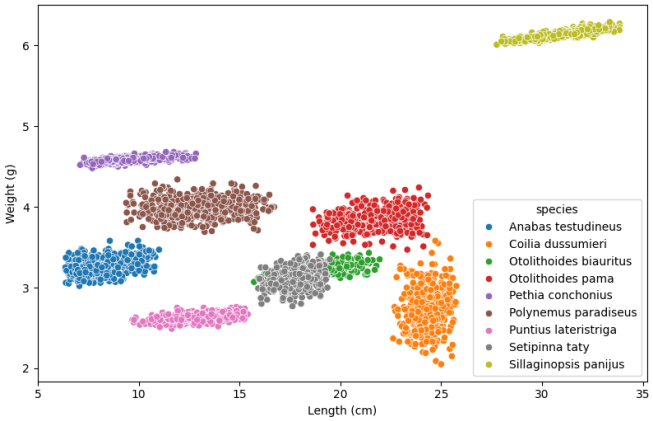

Fig. 1. Length vs Weight

B. Modelos de aprendizaje

Se implementaron los dos modelos de aprendizaje, ID3 y NB, donde se encuentra que para el teorema de Bayes debido a su capacidad para manejar problemas de clasificación utilizando distribuciones de probabilidad. Este enfoque considera la distribución de las características de cada clase (especie) y calcula probabilidades condicionales basadas en una distribución gaussiana. La precisión del modelo fue evaluada utilizando el conjunto de prueba, alcanzando resultados prometedores.

Cuando se implementa el algoritmo de Naive Bayes, el modelo calcula las probabilidades condicionales para cada clase basándose en la distribución de las características, como media, varianza y un valor previo (prior) que representa la probabilidad de cada clase en el conjunto de datos. Este enfoque permite clasificar nuevas observaciones en la clase más probable según los datos aprendidos.

Una técnica implementada para evitar el sobre-ajuste es la regularización de datos, esto nos permite trabajar con un escalado de datos proporcionales y con rangos más parejos, en este caso la diferencia de unidades en el eje x es distinta al del eje y, por lo que es óptimo implementar la regularización de datos [6].

Con el modelo de ID3 podemos ver que este enfoque divide el espacio de características, es decir que divide el conjunto de datos en subconjuntos, que se les llama ramas, las cuales están basadas en la maximización de la ganancia de información, para así evaluar las características de cada clase, seleccionando la mas relevante para clasificar las observaciones.

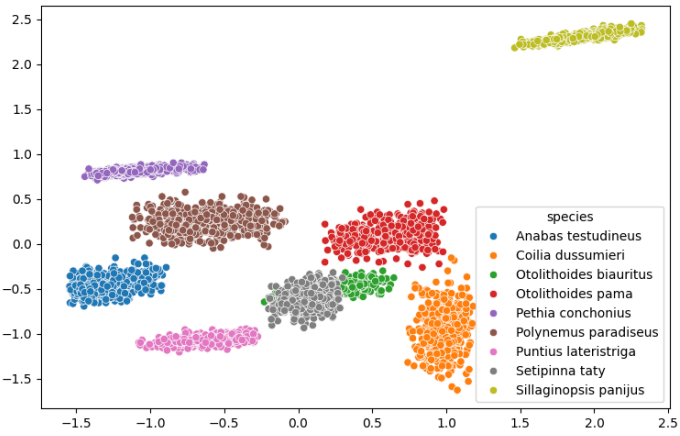

Fig. 2. Standarized values

- **Ejecución de los algoritmos:**

  En esta etapa, los modelo utiliza la totalidad del conjunto de datos para realizar cálculos probabilísticos y clasificar las muestras. En el caso de NB lo que hace, es que para cada clase en los datos, se asume una distribución gaussiana para las características, lo que permite calcular la probabilidad condicional de cada clase dada una muestra. Además, se considera la probabilidad a priori de cada clase, basada en la proporción de muestras pertenecientes a dicha clase en el conjunto completo de datos.
  Este enfoque emplea el 100% de los datos disponibles para modelar y predecir directamente las clases de nuevas muestras, eliminando la necesidad de separar los datos en subconjuntos de entrenamiento y prueba.

  En contraste, ID3 no utiliza probabilidades condicionales ni distribuciones gaussianas, ID3 construye un árbol

dividiendo los datos en subconjuntos, donde utiliza la entropía para medir la impureza de los datos, y a partir de esto selecciona las características que mejor separan las clases.

- **Predicción:**

  Para cada muestra del conjunto de prueba, el modelo evalúa la probabilidad de que la muestra pertenezca a cada clase, combinando la probabilidad a priori de la clase y la probabilidad condicional de las características, calculada con la distribución gaussiana. Finalmente, se asigna a la muestra la clase con mayor probabilidad.

  ID3 evalúa las características de la muestra siguiendo el árbol de decisiones previamente construido, eligiendo el camino que conduce a la clase predicha.

Algorithm 1 Clasificación con Naive Bayes

Entrada: Base de datos completa $D = \{(X, y)\}$, donde X son las características y y son las etiquetas
Salida: Modelo y clases predichas para cualquier nueva muestra

Ejecución del algoritmo:
for cada clase c en y **do**
    Calcular la media μ_c y la varianza σ_c^2 de las características en X para la clase c
    Calcular la probabilidad *prior* $P(c)$ como el número de muestras de la clase c dividido entre el total de muestras
end for
Predicción para una nueva muestra x:
for cada clase c **do**
    Calcular la probabilidad condicional $P(x|c)$ usando la fórmula de la distribución gaussiana:

$$P(x|c) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right)$$

end for
Calcular la probabilidad posterior $P(c|x)$ para cada clase c como:

$$P(c|x) = P(x|c) \cdot P(c)$$

Asignar a x la clase con la mayor probabilidad posterior

Algorithm 2 Algoritmo ID3 (Árbol de Decisión)

Entrada: Conjunto de datos D con características X y etiquetas y **Salida:** Árbol de decisión construido

Función principal:
Construir árbol de decisión:
Si el conjunto de datos D tiene una única clase, **retornar** un nodo hoja con la clase.
Si el conjunto de características X está vacío, **retornar** un nodo hoja con la clase mayoritaria de D.
De lo contrario, calcular el mejor atributo A para dividir el conjunto de datos D:
    Usar la métrica de entropía para seleccionar el atributo A con la mayor ganancia de información.
Crear un nodo de árbol con el atributo A.
Para cada valor v en el dominio de A:
    Crear un subárbol:
        Dividir D en subconjuntos D_v donde $A = v$.
        Llamar recursiva-mente a *Construir árbol de decisión* sobre D_v para construir el subárbol.
Asignar los subárboles al nodo correspondiente.
Retornar el árbol de decisión construido.

Para el desarrollo de este proyecto, se implementaron los modelos Naive Bayes e ID3 utilizando el lenguaje de programación Python. Ambos modelos fueron entrenados con un conjunto de datos de especies de peces, que incluye características como la longitud, el peso y la relación peso-longitud.

El código completo, incluyendo los scripts utilizados para entrenar los modelos y realizar las predicciones, está disponible en el siguiente repositorio público de GitHub: https://github.com/Andrecha13/Proyecto-final-Estructura-de-datos

IV. Resultados

A. Predicción de especie

En la primera prueba realizada, se utilizó un conjunto de características que correspondía a una muestra del conjunto de datos: una longitud de 10.66, un peso de 3.45 y una relación peso-longitud de 0.32. El modelo Naive Bayes predijo correctamente que la especie correspondiente era *Anabas testudineus*, asignando una probabilidad del 100% a esta especie. Las demás especies tuvieron una probabilidad de 0, lo que indica una clasificación sin ambigüedades.

Este resultado resalta la precisión y efectividad del modelo Naive Bayes al momento de predecir especies a partir de características como la longitud, el peso y la relación peso-longitud, especialmente cuando los datos de entrada coinciden con patrones bien establecidos en el modelo.

Para una segunda prueba, se utilizaron características diferentes: longitud de 7, peso de 6.3 y relación peso-longitud de 0.9. En este caso, el modelo Naive Bayes predijo la especie *Polynemus paradiseus*, asignando una probabilidad de 1.00 a esta especie y 0 a las demás. Este resultado demuestra la capacidad del modelo para distinguir entre especies cuando las características de entrada corresponden a un patrón específico dentro de los datos.

En cuanto a la comparación con el modelo ID3, los resultados fueron consistentes con los obtenidos mediante Naive Bayes. En ambas pruebas, ID3 predijo con alta precisión las especies *Anabas testudineus* y *Polynemus paradiseus*, respectivamente, otorgando una probabilidad de 100% a la especie correcta y 0 a las demás. Este comportamiento indica que ambos modelos están realizando clasificaciones efectivas en función de las características proporcionadas.

Sin embargo, en pruebas adicionales, se observaron diferencias en el comportamiento de los modelos al manejar valores extremos o inusuales. Por ejemplo, al usar una longitud de 25, un peso de 3 y una relación peso-longitud de 0.12, ambos modelos predijeron correctamente la especie *Coilia dussumieri*, asignando una probabilidad del 100% a esta especie. Esto reafirma la capacidad de los modelos para manejar datos fuera del rango común pero aún en un espacio reconocible.

No obstante, al probar con características más extremas como una longitud de 5, un peso de 10 y una relación peso-longitud de 2.0, el modelo Naive Bayes presentó problemas en el cálculo de probabilidades, generando valores `NaN`. Este resultado destaca la importancia de considerar un preprocesamiento adicional o reglas para manejar valores fuera del rango entrenado. Finalmente, con características más pequeñas, como una longitud de 1, peso de 0.1 y una relación peso-longitud de 0.1, ambos modelos predijeron correctamente la especie *Anabas testudineus*, asignando nuevamente una probabilidad del 100%.

Estos resultados evidencian que, aunque los modelos Naive Bayes e ID3 funcionan correctamente dentro del rango esperado, es fundamental implementar controles adicionales para casos extremos y garantizar una robustez general en las predicciones.

V. Conclusiones

En este estudio, se evaluaron dos algoritmos de clasificación, Naive Bayes e ID3, con el objetivo de predecir la especie de peces a partir de características como longitud, peso y relación peso-longitud. Ambos algoritmos demostraron una capacidad notable para realizar predicciones certeras en los casos de prueba proporcionados. Sin embargo, se observaron diferencias clave en su desempeño, especialmente en términos de costo computacional.

El algoritmo Naive Bayes resultó ser una opción preferida en este caso, debido a su bajo costo computacional, lo que permitió realizar predicciones de manera rápida y eficiente. Por otro lado, el algoritmo ID3, aunque también funcionó correctamente en las predicciones, presentó un costo computacional mayor, lo cual podría ser una desventaja en situaciones con grandes volúmenes de datos o cuando se requiere una rápida clasificación.

Es importante destacar que la base de datos utilizada en este estudio tiene clases muy bien definidas, lo que contribuyó a la alta precisión de las predicciones realizadas por ambos algoritmos. Las características de los peces en la base de datos son claras y diferenciables, lo que facilita que los modelos realicen predicciones sumamente certeras para los casos incluidos en el conjunto de datos.

Sin embargo, una limitación de estos modelos es que, al ser entrenados con un conjunto específico de datos, podrían presentar problemas al intentar predecir especies no registradas en la base de datos original. Al introducir nuevos peces o especies no previamente observadas, los algoritmos pueden generar predicciones erróneas, ya que no tienen información suficiente para generalizar a partir de estos nuevos casos. Esto subraya la importancia de contar con una base de datos diversa y actualizada para garantizar la efectividad de los modelos en escenarios del mundo real, donde pueden aparecer muestras no contempladas durante el entrenamiento.

VI. Trabajos futuros

- Implementar el modelo con librerías de Python para observar el comportamiento de los modelos con mayor precisión.
- Entrenar los modelos con un porcentaje de los datos totales para evitar el sobreajuste.
- Evaluar la precisión de las pruebas del modelo (Cross-validation).

References

[1] L. Gonzales, "Ventajas y desventajas de los algoritmos de clasificación," 2019.

[2] A. Priyam, G. R. Abhijeeta, A. Rathee, and S. Srivastava, "Comparative analysis of decision tree classification algorithms," *International Journal of current engineering and technology*, vol. 3, no. 2, pp. 334–337, 2013.

[3] M. M. Caballero, "Id3 y arbolpedia: Un enfoque clásico en clasificación," *Data Science*, 2023.

[4] ciberseguridad, "Teorema de bayes en el aprendizaje automático: una guía importante," 2021.

[5] E. Alpaydin, *Machine learning*. MIT press, 2021.

[6] T. Dietterich, "Overfitting and undercomputing in machine learning," *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 326–327, 1995.