

UNIVERSIDAD CATOLICA DE HONDURAS “NUESTRA SEÑORA REINA DE LA PAZ”

CAMPUS “JESUS SACRAMENTADO”



12 de abril de 2020

Mayra GUEVARA, ANGEL SORTO, GERMAN ANTONELLI, CRISTIAN CLAROS, LUIS

ING HECTOR SABILLON

TEMA:

ARBOLES Y BOSQUES ALEATORIOS

Índice

[2 INTRODUCCION 2](#_Toc37623984)

[3 OBJETIVOS 3](#_Toc37623985)

[4 Árboles de Decisión (Decision Trees(DTs)) 4](#_Toc37623986)

[4.1 Construcción de un árbol de decisión 4](#_Toc37623987)

[4.1.1 Elementos 4](#_Toc37623988)

[4.1.2 Conceptos 4](#_Toc37623989)

[4.2 Ventajas y desventajas 5](#_Toc37623990)

[4.2.1 Ventajas 5](#_Toc37623991)

[4.2.2 Desventajas 5](#_Toc37623992)

[4.3 ¿Cómo decide un árbol donde ramificarse? 6](#_Toc37623993)

[4.4 A tener en cuenta cuando usamos Árboles de Decisión 6](#_Toc37623994)

[4.4.1 Regularización 6](#_Toc37623995)

[5 Bosques Aleatorios 7](#_Toc37623996)

[5.1 Bosques Aleatorios Clasificación 7](#_Toc37623997)

[5.2 ¿Cómo funciona el algoritmo? 7](#_Toc37623998)

[5.3 Importancia de las características 7](#_Toc37623999)

[5.4 Bosques Aleatorios versus Árboles de Decisión 8](#_Toc37624000)

[5.5 Ventajas 8](#_Toc37624001)

[5.6 Desventajas 8](#_Toc37624002)

[6 Los Roles y Responsabilidades 9](#_Toc37624003)

[7 Reuniones y Discusiones Sobre El Proyecto A Desarrollar 10](#_Toc37624004)

[8 TABLA DE EVALUCION 11](#_Toc37624005)

[9 CONCLUSIONES 11](#_Toc37624006)

[10 ANEXOS 12](#_Toc37624007)

[10.1 Imagen reunión vía zoom 10 de abril del 2020 12](#_Toc37624008)

[10.2 Imagen reunión vía zoom 11 de abril del 2020 12](#_Toc37624009)

[10.3 Imagen reunión vía zoom 10 de abril del 2020 13](#_Toc37624010)

[11 BIBLIOGRAFIA 14](#_Toc37624011)

# INTRODUCCION

**Machine Learning** es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. **La máquina que realmente aprende es un algoritmo** que revisa los datos y es capaz de predecir comportamientos futuros.

En el siguiente proyecto aprenderemos que son los arboles de decisión, sus ventajas y desventajas así mismo que son los bosques aleatorios, sus ventajas y desventajas. Tratando de poner en practica este tema en un dataset.

# OBJETIVOS

* Investigar que son arboles de decisión.
* Investigar que son bosques aleatorios.
* Implementar en un ejemplo con un dataset el tema asignado para comprender que mane funciona.
* Investigar las ventajas y desventajas de usar arboles de decisión Vrs bosques aleatorios.

# Árboles de Decisión (Decision Trees(DTs))

Un árbol de decisión es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Dicho de otra manera, son un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. El objetivo es crear un modelo que prediga el valor de una variable objetivo mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos.

## Construcción de un árbol de decisión

### Elementos

Los árboles de decisión están formados por nodos, vectores de números, flechas y etiquetas.

* Cada **nodo** se puede definir como el momento en el que se ha de tomar una decisión de entre varias posibles, lo que va haciendo que a medida que aumenta el número de nodos aumente el número de posibles finales a los que puede llegar el individuo. Esto hace que un árbol con muchos nodos sea complicado de dibujar a mano y de analizar debido a la existencia de numerosos caminos que se pueden seguir.
* Los **vectores** de números serían la solución final a la que se llega en función de las diversas posibilidades que se tienen, dan las utilidades en esa solución.
* Las **flechas** son las uniones entre un nodo y otro y representan cada acción distinta.
* Las **etiquetas** se encuentran en cada nodo y cada flecha y dan nombre a cada acción.

### Conceptos

Cuando tratemos en el desarrollo de árboles utilizaremos frecuentemente estos conceptos:

* **Costo:** Se refiere a dos conceptos diferentes: el costo de medición para determinar el valor de una determinada propiedad (atributo) exhibida por el objeto y el costo de clasificación errónea al decidir que el objeto pertenece a la clase X cuando su clase reales Y.
* **Sobreajuste (Overfitting):** Se produce cuando los datos de entrenamiento son pocos o contienen incoherencias. Al tomar un espacio de hipótesis H, se dice que una hipótesis h ∈ H sobreajusta un conjunto de entrenamiento C si existe alguna hipótesis alternativa h' ∈ H tal que h clasifica mejor que h' los elementos del conjunto de entrenamiento, pero h' clasifica mejor que h el conjunto completo de posibles instancias.
* **Poda (Prunning):** La poda consiste en eliminar una rama de un nodo transformándolo en una hoja (terminal), asignándole la clasificación más común de los ejemplos de entrenamiento considerados en ese nodo.
* **La validación cruzada**: Es el proceso de construir un árbol con la mayoría de los datos y luego usar la parte restante de los datos para probar la precisión del árbol.

Los árboles de decisión pueden usarse para resolver problemas tanto de[clasificación como de regresión](https://iartificial.net/clasificacion-o-regresion/).

| **Regresión** | **Clasificación** |
| --- | --- |
| Variable dependiente es continua | Variable dependiente es categórica |
| Valores de los nodos terminales se reducen a la media de las observaciones en esa región. | El valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han “caído” en esa región. |

## Ventajas y desventajas

### Ventajas

* Fácil de entender.
* Útil en exploración de datos: identificar importancia de variables a partir de cientos de variables.
* Menos limpieza de datos: outliers y valores faltantes no influencian el modelo (A un cierto grado)
* El tipo de datos no es una restricción
* Es un método no paramétrico (i.e., no hay suposición acerca del espacio de distribución y la estructura del clasificador)

### Desventajas

* Sobreajuste
* Pérdida de información al categorizar variables continuas
* Precisión: métodos como SVM y clasificadores tipo ensamblador a menudo tienen tasas de error 30% más bajas que CART (Classification and Regression Trees)
* Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol. Por lo tanto, la interpretación no es tan directa como parece.

## ¿Cómo decide un árbol donde ramificarse?

* La decisión de hacer divisiones estratégicas afecta altamente la precisión del árbol.
* Los criterios de decisión son diferentes para árboles de clasificación y regresión.
* Existen varios algortmos para decidir la ramificación.
* La creación de subnodos incrementa la homogeneidad de los subnodos resultantes. Es decir, **la pureza del nodo se incrementa respecto a la variable objetivo.**
* Se prueba la división con todas las variables y se escoge la que produce subnodos más homogéneos.

## A tener en cuenta cuando usamos Árboles de Decisión

### Regularización

Los árboles de decisión de [scikit-learn](https://iartificial.net/librerias-de-python-para-machine-learning/" \l "scikit-learn) no están regularizados por defecto. Esto significa que, para obtener el menor error posible, pueden crear tantos nodos como necesiten. Esto puede resultar en un modelo muy complejo que funcione muy bien en el conjunto de entrenamiento, pero muy mal con datos nuevos.

La [**regularización**](https://iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/)consiste en limitar de alguna forma las capacidades del modelo para obtener un modelo de aprendizaje automático que sea más simple y  [generalice](https://iartificial.net/generalizacion-en-machine-learning/) mejor.

En scikit-learn podemos usar varios hiper-parámetros para configurar cómo regularizamos los árboles de decisión. los más usados son:

* **max\_depth**: la profundidad máxima del árbol.
* **min\_samples\_split**: número mínimo de muestras necesarias antes de dividir este nodo. También se puede expresar en porcentaje.
* **min\_samples\_leaf**: número mínimo de muestras que debe haber en un nodo final (hoja). También se puede expresar en porcentaje.
* **max\_leaf\_nodes**: número máximo de nodos finales

# Bosques Aleatorios

Los Bosques Aleatorios es un algoritmo de Machine Learning flexible y fácil de usar que produce, incluso sin ajuste de parámetros, un gran resultado la mayor parte del tiempo. También es uno de los algoritmos más utilizados, debido a su simplicidad y al hecho de que se puede usar tanto para tareas de clasificación como de regresión.

Los Bosques Aleatorios es un algoritmo de [aprendizaje supervisado](https://ligdigonzalez.com/todo-sobre-aprendizaje-supervisado-en-machine-learning/)que, como ya se puede ver en su nombre, crea un bosque y lo hace de alguna manera aleatorio. Para decirlo en palabras simples: el Bosque Aleatorio crea múltiples [árboles de decisión](https://ligdigonzalez.com/arboles-de-decision-regresion-teoria-machine-learning/)y los combina para obtener una predicción más precisa y estable. En general, mientras más árboles en el bosque se vea, más robusto es el bosque.

## Bosques Aleatorios Clasificación

Técnicamente es un método de conjunto, basado en el enfoque de dividir y conquistar, de árboles de decisión generados en un conjunto de datos dividido al azar. Los árboles de decisión individuales se generan utilizando un indicador de selección de atributos, como la ganancia de información, la relación de ganancia y el índice Gini, para cada atributo. Cada árbol depende de una muestra aleatoria independiente. En un problema de clasificación, cada árbol vota y se elige la clase más popular como resultado final. Es más simple y más potente en comparación con otros algoritmos de clasificación no lineal.

## ¿Cómo funciona el algoritmo?

Funciona en cuatro pasos:

* Construir un árbol de decisión para cada muestra y obtener un resultado de predicción de cada árbol de decisión.
* Realizar una votación por cada resultado previsto.
* Seleccionar el resulta de la predicción con más votos como predicción final.

## Importancia de las características

Otra gran cualidad del algoritmo de Bosques Aleatorios es que es muy fácil medir la importancia relativa de cada característica en la predicción. La librería de Python Sklearn proporciona una gran herramienta para esto, que mide la importancia de las características observando en qué medida los nodos de los árboles, que utilizan esas características, reducen la impureza en todos loar árboles del bosque. Calcula esta puntuación automáticamente para cada característica después del entrenamiento y escala los resultados, de modo que la suma de toda la importancia es igual a 1.

A través de la importancia de las características, puedes decidir qué características deseas eliminar, porque no contribuyen lo suficiente o nada al proceso de predicción. Esto es importante, porque una regla general en Machine Learning es que cuantas más funciones tengas, más probable es que tu modelo sufra de sobreajuste y viceversa.

## Bosques Aleatorios versus Árboles de Decisión

* Los Bosques Aleatorios son un conjunto de múltiples árboles de decisión.
* Los árboles de decisión profundos pueden sufrir de sobreajuste, pero los bosques aleatorios evitan el sobreajuste creando árboles en subconjuntos aleatorios.
* Los árboles de decisión son computacionalmente más rápidos.
* Los Bosques Aleatorios son difíciles de interpretar, mientas que un árbol de decisión es fácilmente interpretable y puede convertirse en reglas.

## Ventajas

* Los Bosques Aleatorios se consideran un método muy preciso y robusto debido al número de árboles de decisión que participan en el proceso.
* No sufre el problema del sobreajsute. La razón principal es que toma el promedio de toas las predicciones, lo que anula los sesgos.
* El algoritmo puede utilizarse tanto en problemas de clasificación como de regresión.
* Los Bosques Aleatorios también pueden manejar los valores que faltan. Hay dos maneras de manejarlos: usando valores medianos para reemplazar variables continuas, y calculando el promedio ponderado por proximidad de los valores faltantes.
* Puede obtener la importancia relativa de las características, lo que ayuda a seleccionar las características más importantes para el clasificador.

## Desventajas

* Los Bosques Aleatorios son lentos en generar predicciones porque tienen múltiples árboles de decisión. Cada vez que hace una predicción, todos los árboles en el bosque tienen que hacer una predicción para la misma entrada dad y luego realizar una votación sobre ella. Todo este proceso lleva mucho tiempo.
* El modelo es difícil de interpretar en comparación con un árbol de decisión, donde se puede tomar una decisión fácilmente siguiente la ruta del árbol.

# Los Roles y Responsabilidades

|  |  |
| --- | --- |
| **Integrante** | **Responsabilidad** |
| Mayra Guevara | * Investigación del tema correspondiente, y aportar con ejemplo de un dataset. * Desarrollo de manual de desarrollo, del proyecto. |
| Ángel Sorto | * Investigación del tema correspondiente, y aportar con ejemplo de un dataset. * Host En las reuniones de Zoom. * Encargado De escribir contenido en la libreta oficial. |
| Cristian Claros | * Investigación del tema correspondiente, y aportar con ejemplo de un dataset. * Desarrollo de manual de desarrollo, del proyecto. |
| German Antonelli | * Investigación del tema correspondiente, y aportar con ejemplo de un dataset. * Encargado en Escribir contenido en la libreta oficial. |
| Luis | * Investigación del tema correspondiente, y aportar con ejemplo de un dataset. |

# Reuniones y Discusiones Sobre El Proyecto A Desarrollar

Nota: (\*) Significa asistió a la reunión.

**Nota:** Se decidió trabajar en dos libretas, una donde íbamos ir contribuyendo de manera individual para la investigación de términos y ejemplos del tema correspondiente, así mismo la libreta a presentar tiene el tirulo de LIBRETA OFICIAL.

|  |  |  |
| --- | --- | --- |
| Reuniones | Asistencia | Discusiones |
| Reunión por Chat grupal vía Whatsapp(6 de abril 2020) | |  |  |  |  |  | | --- | --- | --- | --- | --- | | M | A | G | C | L | | \* | \* | \* | \* | \* | | Discutimos quien haría el repositorio y sus respectivas ramas para trabajar. Se acordó que Mayra Guevara sería la encargada |
| Reunión por Chat grupal vía Whatsapp(7 de abril 2020) | |  |  |  |  |  | | --- | --- | --- | --- | --- | | M | A | G | C | L | | \* | \* | \* | \* | \* | | Monitoreo de trabajo individual y comunicando quien haría push y merge , para que no tuviéramos conflictos con el progreso del trabajo |
| Reunión vía ZOOM (10 de abril del 2020)  Hora: 3:00pm – 5:00pm | |  |  |  |  |  | | --- | --- | --- | --- | --- | | M | A | G | C | L | | \* | \* | \* |  | \* |   Cristian: Reporto su internet no funcionaba. Se le puso al día de la reunión y temas discutidos vía grupo WhatsApp. | Imagen de la reunión en el anexo.  Reunión para verificar avances en la libreta donde íbamos a poner nuestros propios ejemplos  Y planear reunión para el 11 de abril del 2020 para empezar a escribir en la libreta oficial. |
| Reunión Vía zoom (11 de abril del 2020)  Hora: 3:00pm - 6:00pm | |  |  |  |  |  | | --- | --- | --- | --- | --- | | M | A | G | C | L | | \* | \* | \* | \* |  | | Estructuración de la libreta oficial.  Discutimos todos juntos que dataset vamos a presentar para explicar el tema correspondiente a nuestro grupo.  Decidimos que es buena idea realizar un powerpoint .para conceptos. |
| Reunión Vía zoom (11 de abril del 2020)  Hora: 5:00pm - 11:30pm | |  |  |  |  |  | | --- | --- | --- | --- | --- | | M | A | G | C | L | | \* | \* | \* | \* |  | | Trabajo en la libreta final a presentar, y agregar contenido faltante al proyecto. |

# TABLA DE EVALUCION

|  |  |
| --- | --- |
| **Intrigantes** | **Puntaje / 10** |
| Mayra Guevara | 9 |
| Ángel Sorto | 10 |
| Cristian Claros | 9 |
| German Antonelli | 9 |
| Luis Motiño (Coordinador) |  |

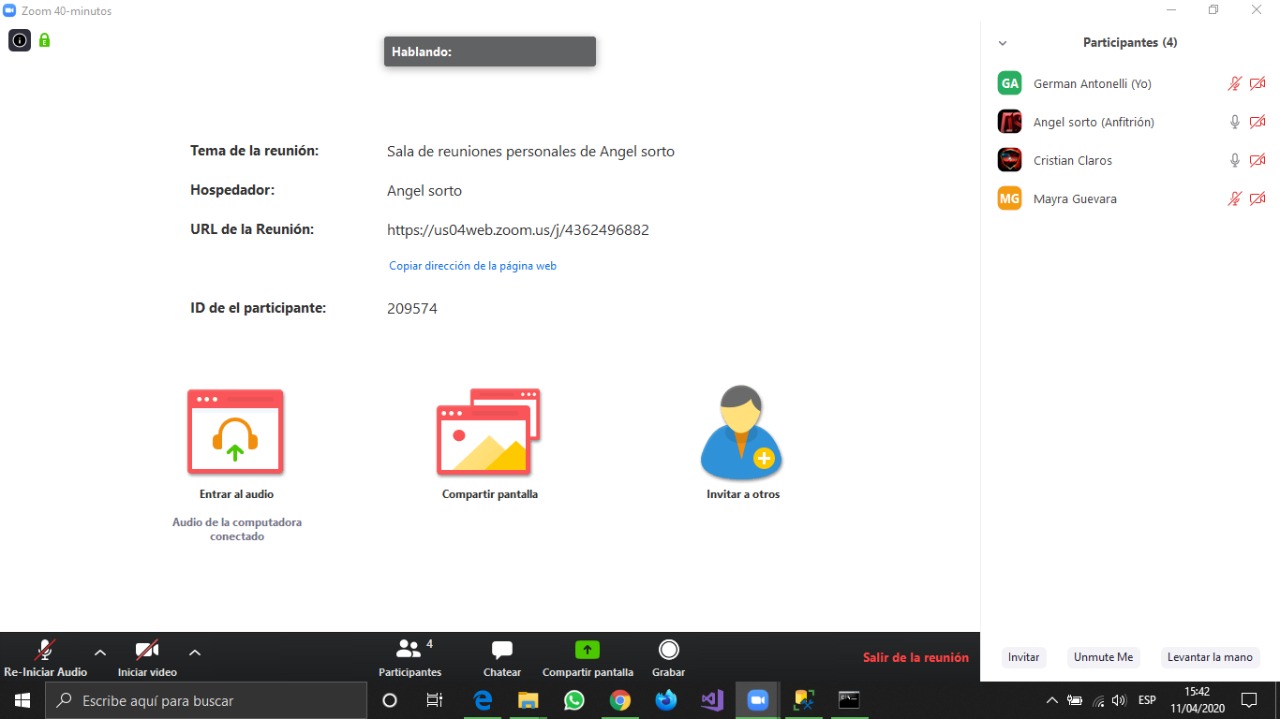
# CONCLUSIONES

* Investigamos que son arboles de decisión.
* Investigamos que son bosques aleatorios.
* Implementamos en un ejemplo con un dataset el tema asignado para comprender que mane funciona.
* Investigamos las ventajas y desventajas de usar arboles de decisión Vrs bosques aleatorios.

# ANEXOS

## C:\Users\Mayra Guevara\Downloads\WhatsApp Image 2020-04-10 at 3.41.20 PM.jpegImagen reunión vía zoom 10 de abril del 2020

## Imagen reunión vía zoom 11 de abril del 2020



## Imagen reunión vía zoom 10 de abril del 2020C:\Users\Mayra Guevara\Downloads\WhatsApp Image 2020-04-12 at 10.44.05 PM.jpeg

# BIBLIOGRAFIA

* <https://iartificial.net/arboles-de-decision-con-ejemplos-en-python/>
* <https://bookdown.org/content/2031/arboles-de-decision-parte-i.html>
* <https://www.youtube.com/watch?v=TqsD5q6aqt0>
* <https://ligdigonzalez.com/bosques-aleatorios-regresion-teoria-machine-learning/>
* <https://ligdigonzalez.com/bosques-aleatorios-clasificacion-teoria-machine-learning/>
* <https://ligdigonzalez.com/bosques-aleatorios-clasificacion-scikit-learn-machine-learning/>
* <https://archive.ics.uci.edu/ml/datasets/spambase>
* <https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html>
* <https://es.coursera.org/lecture/big-data-procesamiento-analisis/modelizacion-con-arboles-de-regresion-s3e5-ipynb-01npW>

# LINK DEL REPOSITORIO