**Estructura de datos y algoritmos I**

**Entrega final**

**Daniel Andrés Hernández Oyola**

**Jose Miguel Blanco Vélez**

**Dónovan Castrillón Franco**

**¿Qué es una estimación?**

Cuando se habla del concepto “estimar” se hace referencia a lo que va a ocurrir o lo que ya ocurrió en el pasado, y a pesar de que puede sonar un concepto muy matemático-estadístico es algo que está en nuestra cotidianeidad, como por ejemplo cuando decimos cuando nos podremos demorar en una tarea o cuando aproximamos el tiempo que nos tomará llegar a un sitio y no solo se puede estimar el tiempo, también podemos estimar distancias y resultados de cualquier índole en general.

A pesar de que sea un concepto que puede aplicarse en muchos campos hay que diferenciar cuando nos estamos refiriendo a una estimación de querer o inferir.

¿Por qué hay que diferenciar esto? Se debe a que cuando hablamos del querer son cosas que no tienen peso alguno en la estadística y no tiene manera de ser “calculado” con precisión alguna, en cambio cuando hablamos de inferir estamos hablando de un proceso que debe tener un resultado aparente a partir de parámetros anteriores con los cuales se han trabajado.

**¿Cómo se hace una estimación?**

Primero que todo debemos saber que dentro de la estimación estadística existen 2 tipos de estimación: la estimación puntual y la estimación por intervalo.

Para diferenciarlas veamos el siguiente ejemplo:

Imagínese que desea estimar la altura aproximada que tienen las personas que es encuentran en algún lugar, para esto usted puede decir simplemente un numero ,por ejemplo la altura promedio de las personas es de 1,70m en este caso se trata de una estimación puntual ya que se está mencionando únicamente un solo dato pero por otro lado usted podría decir que la altura promedio de las personas en el sitio está entre los 1,65m y 1,75m y aquí nos estaríamos refiriendo a una estimación por intervalo ya que como lo dice su nombre estamos dando un intervalo aproximado en el que se pueden o no encontrar la altura de las personas.

**Estimación en las ciencias de datos**

Cuando vamos a trabajar sobre una estimación, probabilidad, y en general de estadística en las ciencias de datos debe haber primero 5 actividades básicas:

1. **Diseño del análisis:** Esta actividad involucra el planeamiento de los detalles para obtener los datos que necesitamos y la generación de la hipótesis a ser evaluada.
2. **Exploración de datos:** En esta actividad nos dedicamos a jugar con nuestros datos, los describimos, los resumimos, realizamos gráficos para mirarlos desde distintos ángulos. Esta exploración nos ayuda a asegurarnos que los datos que obtuvimos son completos y que la etapa de diseño fue correcta.
3. **Armado del modelo:** En esta actividad intentamos armar un modelo que explique el comportamiento de nuestros datos y pueda llegar a hacer predicciones sobre los mismos. La idea es que el modelo pueda describir las propiedades fundamentales de nuestros datos.
4. **Realizar estimaciones**: Aquí vamos a intentar realizar estimaciones basadas en el modelo que armamos anteriormente. También vamos a intentar estimar el tamaño del error que nuestro modelo puede tener en sus predicciones.
5. **Contraste de la hipótesis**: Esta actividad es la que va a producir la decisión final sobre si las predicciones del modelo son correctas y ayudarnos a concluir si los datos que poseemos confirman o rechazan la hipótesis que generamos en la actividad

* También existen conceptos matemáticos que se deben tener en la cuenta dependiendo de las necesidades de nuestro proyecto en cuestión:
* Media aritmética: La media aritmética es el valor obtenido al sumar todos los datos y dividir el resultado entre el número total elementos.
* Varianza: La varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística.
* Desviación típica: La desviación típica es la raíz cuadrada de la varianza.
* Moda: La moda es el valor que tiene mayor frecuencia absoluta.
* Mediana: La mediana es el valor que ocupa el lugar central de todos los datos cuando éstos están ordenados de menor a mayor.

Y estos son solo algunos de los muchos conceptos estadísticos que existen.

Cada uno de estos tiene su manera de ser calculado, sin embargo, para facilitarnos las cosas a los desarrolladores en el lenguaje de programación Python existen distintas librerías que nos permite trabajar con probabilidad y estadística además de que son de libre acceso:

* [numpy](https://www.numpy.org/): El popular paquete matemático de [Python](https://python.org/), se utiliza tanto que mucha gente ya lo considera parte integral del lenguaje. Nos proporciona algunas funciones estadísticas que podemos aplicar fácilmente sobre los arrays de [Numpy](https://www.numpy.org/).
* [scipy.stats](https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html): Este submódulo del paquete científico [Scipy](https://docs.scipy.org/doc/scipy/reference/index.html) es el complemento perfecto para [Numpy](https://www.numpy.org/), las funciones estadísticas que no encontremos en uno, las podemos encontrar en el otro.
* [statsmodels](https://statsmodels.sourceforge.net/): Esta librería nos brinda un gran número de herramientas para explorar [*datos*](https://es.wikipedia.org/wiki/Dato), estimar modelos estadísticos, realizar pruebas estadísticas y muchas cosas más.
* [matplotlib](https://matplotlib.org/): Es la librería más popular en [Python](https://python.org/) para visualizaciones y gráficos. Ella nos va a permitir realizar los gráficos de las distintas distribuciones de datos.
* [seaborn](https://stanford.edu/~mwaskom/software/seaborn/): Esta librería es un complemento ideal de [matplotlib](https://matplotlib.org/) para realizar gráficos estadísticos.
* [pandas](https://pandas.pydata.org/): Esta es la librería más popular para análisis de [*datos*](https://es.wikipedia.org/wiki/Dato) y financieros. Posee algunas funciones muy útiles para realizar [estadística descriptiva](https://es.wikipedia.org/wiki/Estad%C3%ADstica_descriptiva) sobre nuestros datos y nos facilita sobremanera el trabajar con [series de tiempo](https://es.wikipedia.org/wiki/Serie_temporal).
* [pyMC](https://pymc-devs.github.io/pymc/index.html): [pyMC](https://pymc-devs.github.io/pymc/index.html) es un módulo de [Python](https://python.org/) que implementa modelos estadísticos bayesianos, incluyendo la [cadena de Markov Monte Carlo(MCMC)](https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo). [pyMC](https://pymc-devs.github.io/pymc/index.html) ofrece funcionalidades para hacer el análisis bayesiano lo más simple posible.

**¿Qué es la inteligencia artificial?**

La Inteligencia Artificial (IA) en unas pocas palabras es la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano.

La IA está presente en la **detección facial de los móviles,** en los **asistentes virtuales de voz** como Siri de Apple, Alexa de Amazon o Cortana de Microsoft y está **integrada en nuestros dispositivos cotidianos a través de**bots (abreviatura de robots) o **aplicaciones para móvil**

**¿Qué es machine learning?**

El machine learnig en una de las tantas disciplinas que se encuentra dentro del área de las ciencias informáticas. Esta está enfocada principalmente con el desarrollo de la inteligencia artificial, y que sirve para crear sistemas los cuales pueden aprender por sí mismos.

Esta es una tecnología que permite efectuar la automatización de tareas que son principalmente una serie de operaciones, esto con el fin de reducir la necesidad de que intervengan los seres humanos. Esto supone una gran ventaja cuando se trata de controlar una gran cantidad de información para que se realice de un modo mucho más efectivo.

Dentro del machine learning existen diferentes tipos de algoritmos, y estos se clasifican principalmente en 2: aprendizaje supervisado y aprendizaje sin supervisión

**Aprendizaje supervisado:** Cuando nos referimos a ente tipo de aprendizaje es cuando el operador le enseña a la maquina con ejemplo, que quiere decir esto, que se le da a la maquina tanto las entradas como las salidas esperadas y la maquina debe buscar un método con el cual pueda partir desde las entradas hasta dichas salidas esperadas.

Algunos algoritmos de aprendizaje supervisado son:

* Regresión linear
* Regresión polinomial
* Arboles de decisión
* Random forest

**Aprendizaje sin supervisión:**

En este caso el algoritmo de aprendizaje automático tiene que estudiar los datos para identificar algún tipo patrón o patrones. No existe ningún tipo de clave de respuesta o algún tipo de orden de parte del operador. La maquina por si sola debe determinar las relaciones que existan en los datos disponibles.

Algunos algoritmos de aprendizaje sin supervisión son:

* K-Means Clustering
* Hierarchical Clustering

Por temas de este proyecto vamos a profundizar más en el aprendizaje supervisado específicamente en los arboles de decisión.

**Arboles de decisión:**

Un árbol de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. Se pueden usar para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.

Los arboles de decisión generalmente comienzan con un único nodo raíz y se empieza a ramificar en sus posibles resultados los cuales generan nodos extras que a su vez generan otras posibilidades dando como consecuencia una forma similar a la de un árbol.

Dentro de los nodos existen 3 tipos distintos: de probabilidad, de decisión y nodos terminales.

Los arboles de decisión son muy populares ya que en ellos encontramos distintas ventajas:

* Son muy fáciles de entender
* Pueden ser útiles con o sin datos fehacientes, y cualquier dato requiere una preparación mínima
* Se pueden agregar nuevas opciones a los árboles existentes
* Su valor al seleccionar la mejor de numerosas opciones
* Se combinan fácilmente con otras herramientas de toma de decisiones

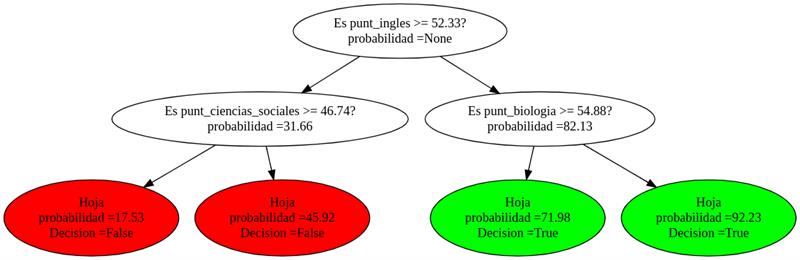
Sin embargo, en algunas ocasiones esto se pueden volver muy complejos lo cual hace que su compresión sea compleja. Cuando sucede esto se debería empezar a pensar en otro tipo de diagrama para la solución del problema.

**Proyecto:**

En el proyecto se prefirió usar la ganancia de la información en vez del Gini por motivos de practicidad esto resuelve la duda sobre si usar el Gini es necesario, la repuesta claramente es no. Existen distintas maneras de conseguir algo similar que sea igual de preciso.

Como demostración del resultado del proyecto estaremos usando el dataset con nombre “TRAIN(4)”. La profundidad que se decidió usar fue de 3 pero cabe aclarar que se puede usar cualquiera:

**Árbol generado**

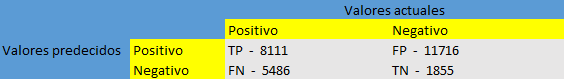


El color rojo es simboliza que la moda es falsa lo que traduce es que cualquiera que llegue a esta instancia es porque perdió, en cambio la verde es porque es verdadera lo cual significa que ganó.

En cada nodo existente se coloca la probabilidad que tiene que ganar la persona en ese punto.

La exactitud de este árbol generado es de un 72,98%

**Matriz de confusión**



Métodos auxiliares usados:

* partition
* targetcalculate
* entropía
* mejorDiv
* seleccionarDatos
* predicciones
* calcularMejorDiv
* crearMatriz
* correr

Bibliografía:

* Manzano Arrondo V. (2012-2014). Estimación estadística. Recuperado de: [EstimacionEstadistica.pdf (us.es)](http://asignatura.us.es/dadpsico/apuntes/EstimacionEstadistica.pdf)
* San Nicolás, N.L. (2009). Hacer estimaciones estadísticas. Recuperado de:

[Microsoft Word - 10 Hacer estimaciones (uanl.mx)](http://eprints.uanl.mx/12506/1/A10.pdf)

* (2015). Probabilidad y estadística con Python. Recuperado de:

[Probabilidad y Estadística con Python (relopezbriega.github.io)](https://relopezbriega.github.io/blog/2015/06/27/probabilidad-y-estadistica-con-python/)

* [Redacción APD](https://www.apd.es/author/redaccion/) (2019). ¿Qué es machine learning y cómo funciona? Recuperado de: [Qué es Machine Learning, cómo funciona y a qué se aplica | APD](https://www.apd.es/que-es-machine-learning/)
* [Redacción APD](https://www.apd.es/author/redaccion/) (2019). ¿Cuáles son los tipos de algoritmos de machine learning? Recuperado de: [¿Cuáles son los tipos de algoritmos del machine learning? | APD](https://www.apd.es/algoritmos-del-machine-learning/)
* Que es un diagrama de árbol de decisión. Recuperado de: [¿Qué es un diagrama de árbol de decisión? | Lucidchart](https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision)
* Ligdi Gonzales (2018). Clasificación de machine learning | #4 Curso de introducción al machine learning. Recuperado de: [(3) CLASIFICACIÓN DE MACHINE LEARNING | #4 Curso de Introducción a Machine Learning - YouTube](https://www.youtube.com/watch?v=R4SZGkLIp88&list=PLJjOveEiVE4Cbbx1dVjydfmPPpjl0pg86&index=4&ab_channel=AprendeIAconLigdiGonzalez)