

Proyecto integrador I

Proyecto final. Entrega IV

Profesor: Juan Manuel Reyes

Integrantes

Juan Fernando Martinez

Gonzalo De Varona

Alejandro Garcia

Santiago Rodas Rodriguez

Santiago de Cali - Junio 2021

**Etapas del diseño de experimento**

**1. Planeación y realización  
A) Entender y delimitar el problema u objeto de estudio.**

Para realizar el experimento, primero necesitamos recolectar la información necesaria acerca de las dos implementaciones del árbol de decisión utilizadas en el proyecto:

1. **Implementación propia:** inicialmente obtuvimos el indice de gini de cada columna para poder saber que columnas clasificaba mejor los datos a comparación de las demás, sin embargo la mayoria de las columnas eran de variables continuas entonces se comparó entre cada columna que valores clasificaban mejor los datos para poder construir el árbol de decisión, inclusive se puede presentar varias veces en el árbol de decisión que se utilice una misma columna varias veces para clasificar los datos.
2. **Implementación usando la librería de ML .Net de Microsoft:** estalibrería de C# es mucho más fácil de usar a comparación de hacer la implementación propia, su proceso es simplemente seleccionar el dataset, la variable dentro de este que se va a usar para la clasificación, introducir la cantidad de tiempo destinado al entrenamiento de datos y ya. La gran desventaja de esta librería es que funciona como una especie de “caja negra” entonces no se específica muy bien que pasos siguen los algoritmos que usa la librería, sin embargo esta es capaz de hacer predicciones, generar un modelo y obtener un porcentaje de precisión con base en un dataset dado. Un punto positivo para esta implementación es que se presentan resultados y modelos con base a distintos algoritmos y tecnicas de machine learning, entonces no se limita a generar un árbol de decisión como la otra implementación.

Ambas implementaciones cumplen con la tarea de generar un árbol de decisión con base a nuestro dataset, una implementación es más sencilla de hacer que la otra pero esto no indica que ambas sean igual de precisas, que es lo que realmente importa en este experimento, determinar cual de las dos implementaciones es la más precisa.

\*Nota: Para ambas implementaciones se hizo una discretización de las variables continuas, agrupando valores en intervalos.

**B) Elegir la(s) variables(s) de respuesta que será medida en cada punto del diseño y verificar que se mide de manera confiable.**

Tomando en cuenta los resultados que esperamos tener, y el diseño que queremos llevar a cabo, decidimos tomar el tiempo de entrenamiento del árbol, el porcentaje de la precisión del modelo usando el dataset de entrenamiento, y el porcentaje de la precisión del modelo usando el dataset de prueba como las variables de respuesta, ya que es de esta forma es que se puede medir la precisión y eficiencia de los modelos producidos por los árboles de decisión generados.

En la implementación con ML .Net cuando se presentan los resultados de los modelos generados está incluido el porcentaje de precisión del modelo usando el dataset de prueba y de entrenamiento para cada algoritmo usado por la librería y también el tiempo de entrenamiento del modelo.

En la implementación propia, se selecciona la parte del dataset no usada para entrenar los datos y se introduce cada datapoint al árbol de decisión generado, si el árbol clasifica correctamente el datapoint entonces se cuenta dentro de los casos exitosos y se hace casos de éxito sobre casos totales eso multiplicado por 100 para obtener la precisión de las predicciones hechas usando el dataset de prueba, solo basta con cambiar el dataset al de entrenamiento y hacer este mismo proceso para obtener la precisión de las predicciones hechas usando el dataset de entrenamiento. Para obtener el tiempo de entrenamiento del modelo este es cronometrado en el programa.

**C) Determinar la unidad experimental.**

La unidad experimental es un conjunto de ejecuciones donde se construye el árbol de decisión usando las 2 implementaciones usadas, con ML .Net y con nuestra propia implementación. Donde además serán llevadas a cabo en diferentes computadores con diferentes especificaciones, lo cual puede arrojar distintos resultados en cada caso.

**D) Determinar cuáles factores deben estudiarse o investigarse, de acuerdo a la supuesta influencia que tienen sobre la respuesta.**

Los factores estudiados elegidos para realizar el experimentos son los siguientes 3:

1. Tamaño del dataset de entrenamiento: Es de suma importancia la cantidad de datos del dataset al momento de obtener un modelo y generar un árbol de decisión ya que esto puede aumenatar su preción de forma directa, sin embargo no es algo que suceda para todos los casos posibles.
2. Hardware del computador: Procesador y cantidad de memoria ram, el hardware es evidente que influye en el tiempo que tardan los ordenamientos. La memoria ram usualmente toma valores entre 4 y 32GB siendo los más comunes entre 4-16 GB. Los modelos de procesadores son bastante variados con Intel y AMD las empresas más grandes que los fabrican.

Factores controlables no estudiados: Modo de consumo energético del sistema (Balanceado, ahorrador de energía, máximo rendimiento), se fija en máximo rendimiento.

Los factores no controlables que inciden en el proyecto son: el desgaste de los componentes internos de los computadores donde se realizará el experimento, la totalidad de tareas de los computadores que se están ejecutando en el momento que se realice el experimento (sólo algunas tareas se pueden controlar pero algunas tareas internas no), frecuencia a la cual trabaja el procesador, al ser dinámica.

**E) Seleccionar los niveles de cada factor, así como el diseño experimental adecuado a los factores que se tienen y al objetivo del experimento.**

Para las variables de estudio escogidas se escogen los siguientes niveles que resultan en un total de 16 tratamientos para cada implementación.

1. Tamaño del dataset de entrenamiento, de los tamaños posibles se escogieron 4:

a. 10806 (50%)  
b. 14048 (65%)  
c. 17290 (80%) d. 20532 (95%)

1. Hardware del computador, el procesador y RAM del computador de cada miembro del equipo.
   1. (Santiago) Intel Core i5-8250U y 8GB de RAM.
   2. (Gonzalo) Intel Core i5-8250U y 16 GB de RAM.
   3. (Juan) Intel Core i7-7700HQ y 16 GB de ram RAM
   4. (Alejandro) Intel Core i5-8250U y 8GB de RAM.

Este experimento se realizará con 3 repeticiones para cada tratamiento por cada implementación, pero el usuario tendrá la opción de realizar una mayor cantidad de repeticiones. Claramente, tomando en cuenta cada uno de los aspectos anteriormente nombrados. Además, como las condiciones pueden ser distintas, se espera que cada repetición sea diferente. Es decir, un caso puede ser tomar el dataset con 10806 datos, 8 GB de RAM, procesador Intel Core i5-8520U. Mientras tanto, otro caso puede ser tomar el dataset con 17290 datos, 16 GB de RAM, procesador Intel Core i7-7700HQ.

No se usó el 100% del dataset para el entrenamiento del modelo puesto que hay que dejar una cantidad de datos del mismo para posteriormente hacer un testeo del modelo.

De esta manera, no solo se piensa obtener resultados de alta precisión, sino también podremos saber la resolución de cada una de las medidas alcanzadas.

**F) Planear y organizar el trabajo experimental.**

En este diseño experimental se tomarán en cuenta los siguientes aspectos:

1. Las personas involucradas en el diseño y análisis del experimento serán todos los integrantes.
2. Las personas involucradas en la codificación y programación del experimento serán todos los integrantes.
3. Las personas involucradas en la toma de resultados e interpretación de los datos serán todos los integrantes.

Se piensa realizar en conjunto, ya que cuando los miembros colaboran, su capacidad para el éxito se multiplica. Sin embargo, los equipos exitosos no se crean de la noche a la mañana. En lugar de eso, mediante un proceso que implica la gestión de proyectos, el desarrollo del flujo de trabajo y la creación de procedimientos, los equipos encuentran el éxito, un paso a la vez. Además, cuando los empleados y miembros del equipo trabajan juntos en un entorno alentador, los beneficios son innegables. El hecho de forjar la comunicación dentro del grupo y otras condiciones abiertas ayudará a crear un sentido de cohesión, inversión y confiabilidad entre todos los miembros del equipo. Esto conduce a una colaboración y una gestión general de proyectos más efectivas.

**G) Realizar el experimento.**

En este subpunto se toma las siguientes decisiones:

1. En caso de un imprevisto o de un problema, este inmediatamente será notificado a la siguiente persona: Juan Fernando Martinez.
2. Los datos serán puestos en un archivo Excel, donde podremos visualizarlos por medio de gráficas comparativas y/o algunas herramientas de análisis de datos preestablecidas.
3. Todo se visualizará en el repositorio principal: link en GitHub.

**2. Análisis**

El análisis se llevará a cabo mediante el uso de indicadores de estadística descriptiva para comparar los diferentes tratamientos. Las comparaciones a realizar se limitan de acuerdo a lo siguiente:

1. La exactitud de medición: El cronómetro permite medir con una exactitud de milisegundos, y aquellas mediciones que tardaron menos de este valor no pueden ser analizadas con el mismo nivel detalle que las demás
2. Resultados exactos: En el caso de que los resultados puedan tomarse como constantes para un tratamiento dado. Está también ligado a lo anterior, por ejemplo cuando los tiempos son constantes y son 0.
3. Niveles para los cuales hay resultados invariables: Cuando los niveles analizados son limitados por los dos puntos anteriores y varios tienen resultados aparentemente equivalentes.

**Caso especial, los datos de entrenamiento tomados del dataset son menores a 10806:**

En el campo de machine learning no siempre lo mejor es agregar más datos al dataset de entrenamiento del modelo pero siempre se debe tener una cantidad minima considerable para poder crear el modelo, en este caso consideramos que esa minima cantidad era 10806 datos, consideramos que menos de eso, la precisión del modelo era muy baja por tanto este era practicamente inutil.