### Experimentos sobre los clasificadores

1. **Identificación componentes del Análisis y Diseño Experimental**
   1. **Unidad Experimental**

En el experimento planteado, las unidades experimentales per sé son los árboles de decisión trabajados que serían el manual y el automático hecho a base de librerías. Los árboles se someterán a un número determinado de pruebas, con distintos factores de entrada cómo la cantidad de datos para el entrenamiento y cantidad de datos para testear comparando así cuál árbol tiene una mejor *accuracy*.

### Variables de respuesta

La variable de respuesta escogida para este experimento es el porcentaje de asertividad y cantidad de datos filtrados de manera correcta. De acuerdo con la cantidad de datos para la prueba y principalmente la cantidad de datos usados para entrenar los modelos la cantidad de datos filtrados de manera correcta debería cambiar, y como consecuencia la *accuracy* también.

### Factores controlables

Los factores identificados que pueden ser controlados a voluntad para evidenciar variabilidad en la *accuracy* de los árboles en cuestión son:

* + - **Cantidad de datos para el entrenamiento:** Se harán pruebas con 250,200 y 150 datos para entrenar el modelo manual o automático
    - **Cantidad de datos para la prueba:** Se harán pruebas con los datos mencionados anteriormente agregando el factor de cantidad de datos para verificar o encontrar la *accuracy* del modelo estos serán 10, 20 y 30
    - **Clasificador (Árbol) manual o automático:** Los dos factores mencionados anterior mente se testearán en el clasificador manual y en el clasificador automático para determinar cuál es mejor
    - **Estructura del árbol de decisión:** No se harán pruebas sobre cómo afecta la *accuracy* los diferentes tipos de árboles que se pueden realizar pero que a su vez estos son equivalentes
    - **Formato numérico en que se procesan los datos:** No se realizaran pruebas para analizar cómo afecta el formato numérico que usa el árbol de decisión la *accuracy*.
    - **Librerías que se usan en el caso del clasificador automático:** Como solo se usó un tipo de librería para realizar el árbol de decisión automático, entonces no se harán pruebas para ver como el uso de una librería u otra afecta la *accuracy.*

### Factores no controlables

* **Decisión que debe tomar cuando una ocurrencia a recorrido todo el árbol y no se logra clasificar de manera precisa:** En otras palabras, cuando se quiere clasificar una ocurrencia y esta recorre todo el árbol y no se logra clasificar de manera concisa y el árbol determinara de manera aleatoria la clasificación de la ocurrencia, esto aplica principalmente para el clasificador automático realizado con librería.

### Factores estudiados

En este experimento tomamos los factores de estudio a los factores controlados ya que son los factores que influyen directamente en la variación de la *accuracy* en los clasificadores.

* + - Cantidad de datos para el entrenamiento.
    - Cantidad de datos para la prueba.
    - Clasificador (Árbol) manual o automático.

### Niveles

* + - Niveles utilizados según la cantidad de ocurrencias para entrenar los clasificadores

|  |  |
| --- | --- |
| **Número de elementos de entrenamiento** | **Nivel de elementos** |
| 250 | 1 |
| 200 | 2 |
| 150 | 3 |

* + - Niveles utilizados la cantidad de ocurrencias para testear los clasificadores

|  |  |
| --- | --- |
| **Número de elementos de prueba** | **Nivel de elementos** |
| 10 | 1 |
| 20 | 2 |
| 30 | 3 |

* + - Niveles utilizados según el clasificador que se usara en el experimento

|  |  |
| --- | --- |
| **Clasificador** | **Nivel de elementos** |
| Automático | 1 |
| Manual | 2 |

**g.** **Tratamientos**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cantidad de datos para entrenamiento** | **Cantidad de datos para la prueba** | **Clasificador automatico o manual** | **Tratamiento** | **Repeticion** | **Y (Resultado)(Accuracy arbol)** | **Z(Resultado)(Cantidad de acertados)** | **Promedio** |
| 250 | 10 | Automatico | 1 | 1 | 70% | 7 | 73% |
| 250 | 10 | Automatico | 2 | 70% | 7 |
| 250 | 10 | Automatico | 3 | 80% | 8 |
| 250 | 20 | Automatico | 2 | 1 | 70% | 14 | 75% |
| 250 | 20 | Automatico | 2 | 85% | 17 |
| 250 | 20 | Automatico | 3 | 70% | 14 |
| 250 | 30 | Automatico | 2 | 1 | 83% | 24 | 81% |
| 250 | 30 | Automatico | 2 | 73% | 21 |
| 250 | 30 | Automatico | 3 | 86% | 25 |
| 200 | 10 | Automatico | 4 | 1 | 70% | 7 | 77% |
| 200 | 10 | Automatico | 2 | 80% | 8 |
| 200 | 10 | Automatico | 3 | 80% | 8 |
| 200 | 20 | Automatico | 5 | 1 | 70% | 14 | 68% |
| 200 | 20 | Automatico | 2 | 65% | 13 |
| 200 | 20 | Automatico | 3 | 70% | 14 |
| 200 | 30 | Automatico | 6 | 1 | 76% | 22 | 71% |
| 200 | 30 | Automatico | 2 | 73% | 21 |
| 200 | 30 | Automatico | 3 | 63% | 18 |
| 150 | 10 | Automatico | 7 | 1 | 75% | 7 | 82% |
| 150 | 10 | Automatico | 2 | 80% | 8 |
| 150 | 10 | Automatico | 3 | 83% | 8 |
| 150 | 20 | Automatico | 8 | 1 | 60% | 12 | 67% |
| 150 | 20 | Automatico | 2 | 70% | 14 |
| 150 | 20 | Automatico | 3 | 70% | 14 |
| 150 | 30 | Automatico | 9 | 1 | 76% | 22 | 75% |
| 150 | 30 | Automatico | 2 | 76% | 22 |
| 150 | 30 | Automatico | 3 | 73% | 21 |
| 250 | 10 | Manual | 10 | 1 | 50% | 5 | 50% |
| 250 | 10 | Manual | 2 | 50% | 5 |
| 250 | 10 | Manual | 3 | 50% | 5 |
| 250 | 20 | Manual | 11 | 1 | 40% | 8 | 40% |
| 250 | 20 | Manual | 2 | 40% | 8 |
| 250 | 20 | Manual | 3 | 40% | 8 |
| 250 | 30 | Manual | 12 | 1 | 43% | 12 | 43% |
| 250 | 30 | Manual | 2 | 43% | 12 |
| 250 | 30 | Manual | 3 | 43% | 12 |
| 200 | 10 | Manual | 13 | 1 | 40% | 4 | 40% |
| 200 | 10 | Manual | 2 | 40% | 4 |
| 200 | 10 | Manual | 3 | 40% | 4 |
| 200 | 20 | Manual | 14 | 1 | 30% | 6 | 30% |
| 200 | 20 | Manual | 2 | 30% | 6 |
| 200 | 20 | Manual | 3 | 30% | 6 |
| 200 | 30 | Manual | 15 | 1 | 30% | 9 | 30% |
| 200 | 30 | Manual | 2 | 30% | 9 |
| 200 | 30 | Manual | 3 | 30% | 9 |
| 150 | 10 | Manual | 16 | 1 | 0% | 0 | 0% |
| 150 | 10 | Manual | 2 | 0% | 0 |
| 150 | 10 | Manual | 3 | 0% | 0 |
| 150 | 20 | Manual | 17 | 1 | 0% | 0 | 0% |
| 150 | 20 | Manual | 2 | 0% | 0 |
| 150 | 20 | Manual | 3 | 0% | 0 |
| 150 | 30 | Manual | 18 | 1 | 0% | 0 | 0% |
| 150 | 30 | Manual | 2 | 0% | 0 |
| 150 | 30 | Manual | 3 | 0% | 0 |

1. **Etapas del análisis y diseño de experimentos se han llevado a cabo hasta el momento**

A continuación, se enlista y se explica de manera resumida las etapas realizadas hasta el momento durante el experimento realizado cuyo objetivo es demostrar cuál de los dos clasificadores tiene mejor desempeño, es decir, una mejor *accuracy*.

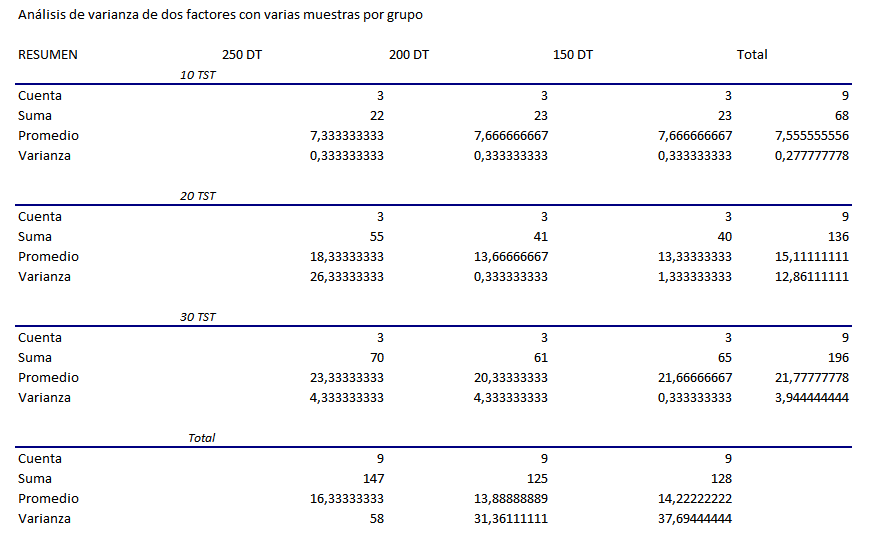
* **Etapa planeación y experimentación:** En esta etapa se delimitó el objeto del problema el cual fue el estudio de la *accuracy*. Se establecieron los factores de estudio, la variable de respuesta, niveles y realización del experimento con un número de repeticiones establecidas que pudieran arrojar resultados evidentes.
* **Etapa análisis:** En esta etapa se realiza la técnica estadística central en el análisis de experimentos llamado el análisis de varianza o ANOVA. Esta prueba se hace con el objetivo de obtener efectos o diferencias muestrales, entendiendo que estamos viendo que los resultados experimentales son observaciones muestrales y no poblacionales.
* **Etapa interpretación: En esta etapa a través del análisis de varianza se puede evidenciar que efectivamente se marca una tendencia que** el clasificador automático ofrece en términos generales una mejor *accuracy.*

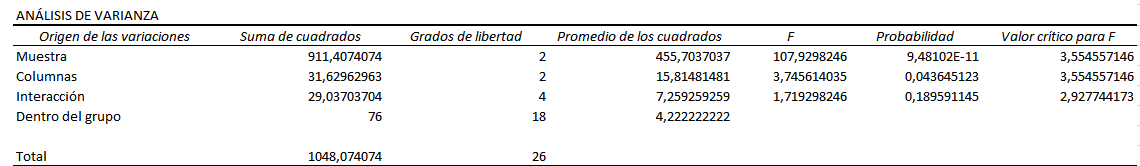
**Análisis ANOVA**

**En el siguiente análisis de datos se debe tener en cuenta lo siguiente:**

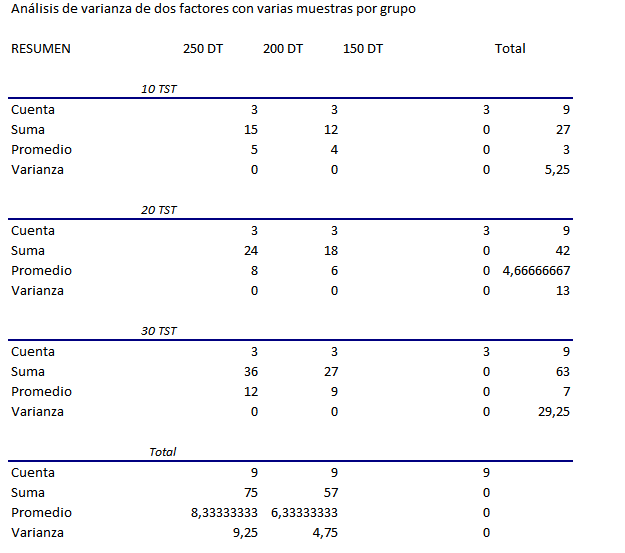
1. **Planteamiento de hipótesis:**
   1. **Hipótesis nula:** La media de la precisión de predicción de ambos modelos en los experimentos con 250,200 y 100 datos de población, son iguales
   2. **Hipótesis alternativa:** Se basa en afirmar que ambas implementaciones con las variaciones de tamaño de la población tendrán comportamientos distintos, por ende precisión de la predicción de cada modelo será distinta.

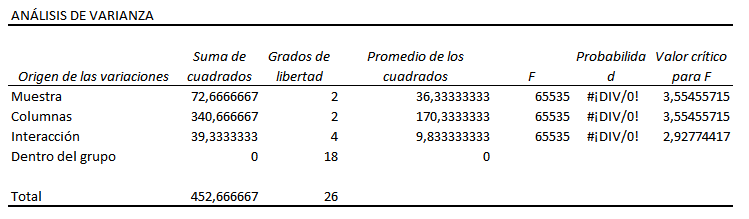
**Resultados análisis ANOVA utilización de librería para árbol de decisión**





**Resultados análisis ANOVA de árbol de decisión con implementación propia**





**Análisis resultados**

Según los datos obtenidos en las tres formas de ordenar elementos con respecto a los resultados de los experimentos, al ser F mayor al Valor Crítico para F, se rechazaría la hipótesis nula.

**Conclusiones**

* Tras haber realizado los cuasiexperimentos y observar los resultados, podemos concluir que la implementación por lo librería del árbol de decisión es mucho más preciso en cada cuasiexperimento realizado.
* Si se revisa la tabla de tratamientos se puede evidenciar que cada uno de los resultados del clasificador automático presenta una *accuracy* mejor que la del manual incluso, cuando se le asigna 150 datos para el entrenamiento logra sacar una *accuracy* aceptable, en cambio el clasificador manual con 150 datos no es capaz de generar el árbol de manera indicada por lo que da como resultado una *accuracy* del 0%.
* Se puede evidenciar que el clasificador automático en cada uno de los casos (150,200 y 250) logra una *accuracy* aceptable, una *accuracy* de un clasificador se considera aceptable si es superior al 65% en cambio el clasificador manual en ninguno de los casos logra una *accuracy* aceptable.
* Con respecto a la cantidad de datos para el entrenamiento en unos casos hace que la *accuracy* sea mayor, y en otros que sea menor, por lo que se puede decir que lo que realmente afecta la *accuracy* es la cantidad de datos seleccionados para el entrenamiento, como se evidencio previamente en el ANOVA.