Tecnológico de Monterrey

Maestría en Inteligencia Artificial Aplicada

Ciencia y Analítica de Datos

Dra. María de la Paz Rico

Clasificación-ensambles aguas subterraneas

Pablo Alejandro Bravo Vargas - A01793024

Antonio Saenz Ramirez - A01793884

Tecnológico de Monterrey

La base cuenta con la siguiente información:

- 56 columnas
- 1,068 registros

Limpieza de Datos

Se realizaron 2 tipos de limpieza:

- **Registros nulos:** el 23% de las columnas les faltaban datos, 1 columna se eliminó ya que no tenia ningun registro y el resto se sustituyeron por la mediana de los datos ya que el porcentaje de los faltantes eran menor del 1%.
- **Cambio de formato:** Los datos que queremos analizar contienen texto y los cambiamos por valores que pueden ser estudiados.

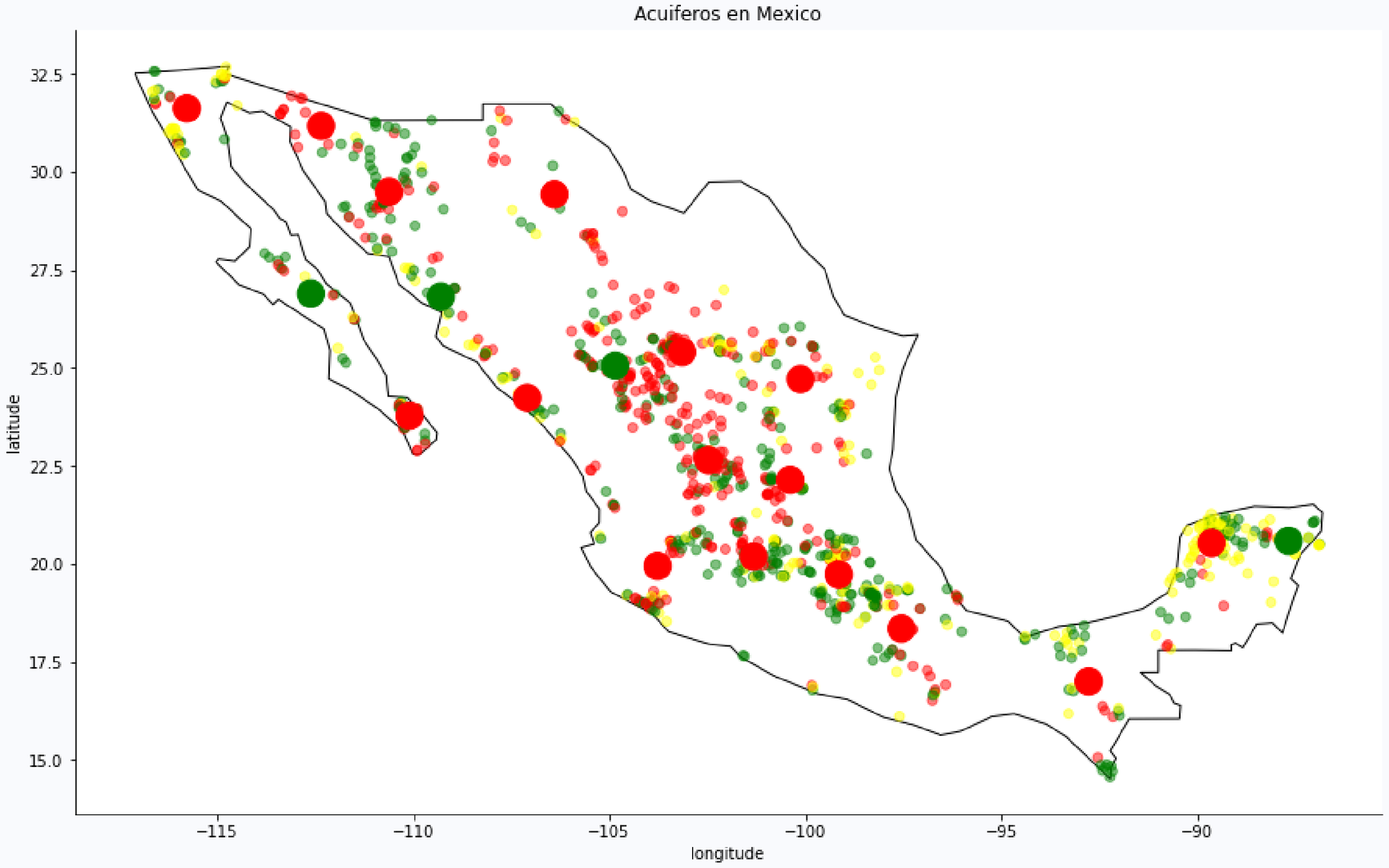
Tecnológico de Monterrey

Análisis de los Datos

Se obtuvo la matriz de correlación de los datos, en el cuál podemos observar que la mayoría de los datos presentan una correlación entre ellos.

Tecnológico de Monterrey

Análisis de los Datos

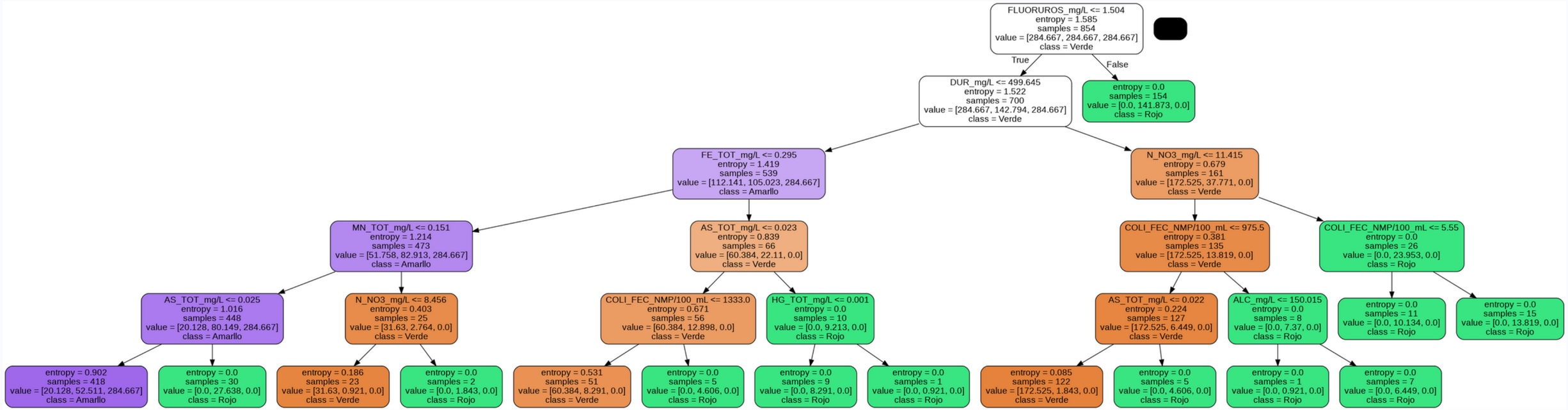


Se identificó la distribución de los diferentes mantos acuiferos subterranos de México y de acuerdo a su evaluación de la calidad del agua

Tecnológico de Monterrey

Clasificador "Tree Decision"

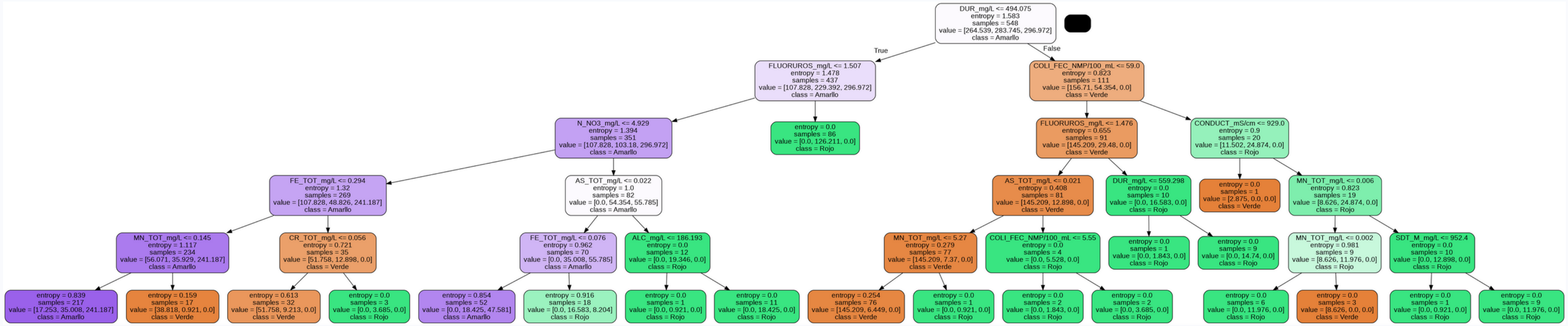
Creamos un clasificador de "Tree Decision" con 5 niveles el cuál tiene una precisión del 90.3%



Tecnológico de Monterrey

Clasificador "Random Forest"

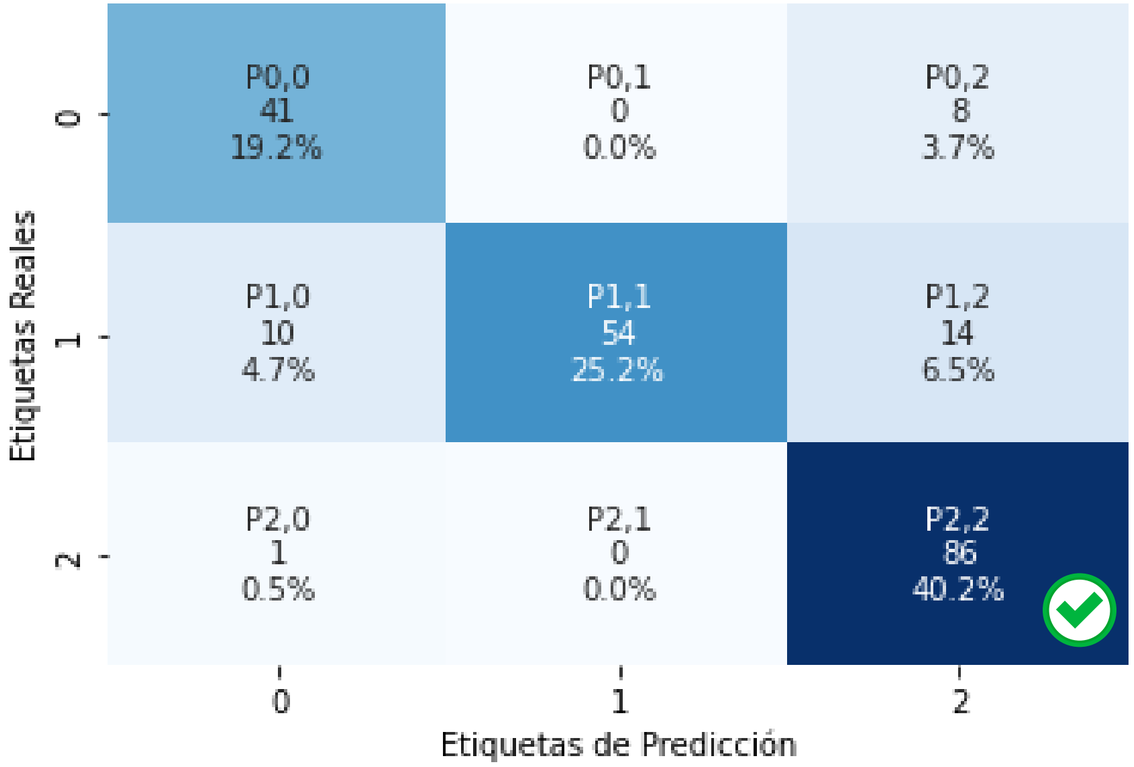
Creamos un clasificador de "Random Forest" con 5 niveles el cuál tiene una precisión del 96.7%



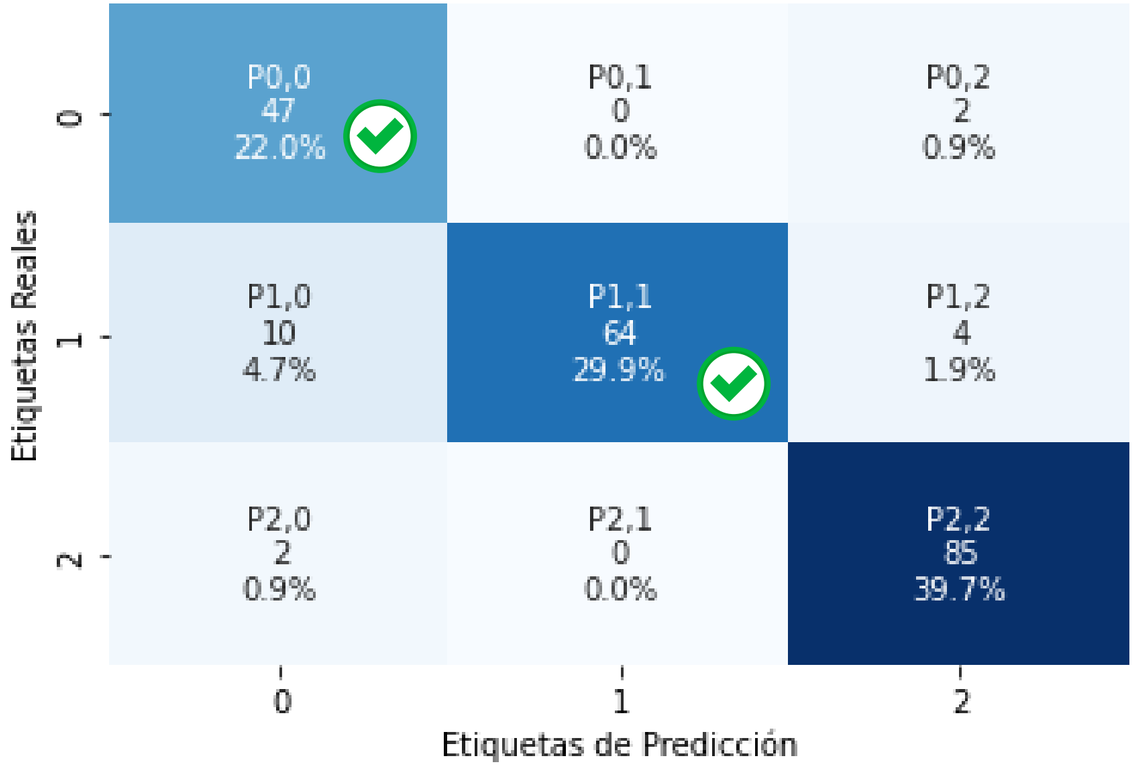
Tecnológico de Monterrey

Resultados: Matriz de Confusión

"Tree Decision"



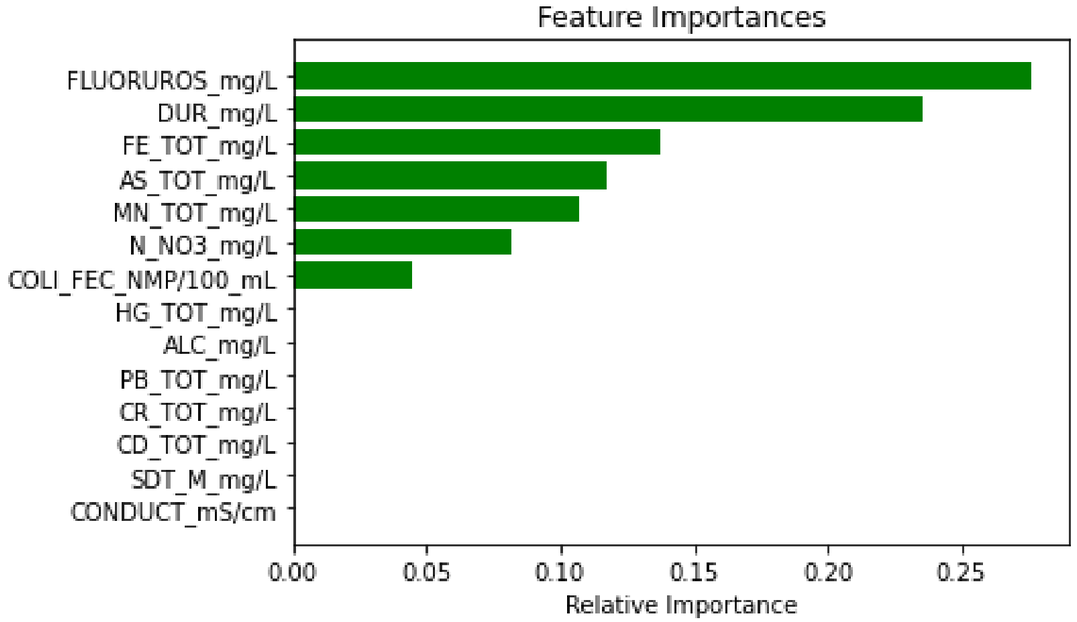
"Random Forest"



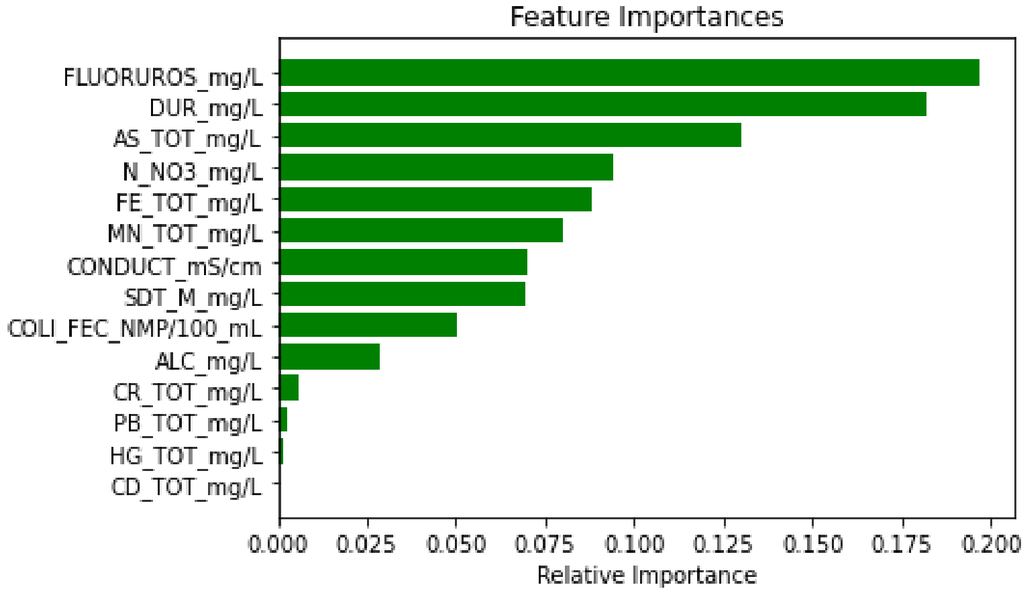
Tecnológico de Monterrey

Resultados: Importancia de las variables

"Tree Decision"



"Random Forest"



Tecnológico de Monterrey

Conclusión

Creemos que para este ejercicio los dos modelos que se utilizaron presentan una precisión aceptable, en el cual, considerando el mismo número de niveles (en nuestro ejercicio consideramos 5), el clasificador con mayor exactitud el el Random Forest con 96.72% de exactitid vs 90% del Tree Decision.

Esto se debe a que el Random Forest es la combinación de varios Tree Decision, aunque en este ejemplo el número de datos era pequeño se pudo trabajar con los dos clasificadores con la misma velocidad, posiblemente si la base fuera más grande, la mejor opción podría llegar a ser Tree Decision perdiendo un poco de precisión pero mejorando el rendimiento.