PROYECTO FINAL

CLASIFICACIÓN DEL NIVEL DE CONTAMINACIÓN DE AGUAS SUBTERRÁNEAS EN MÉXICO A PARTIR DE UN MODELO DE PREDICCIÓN DE UN ÁRBOL DE DECISIONES (DECISION TREE)

CIENCIA Y ANALÍTICA DE DATOS

PROFESOR TITULAR: DRA. MARÍA DE LA PAZ RICO FERNÁNDEZ
PROFESOR TUTOR: ORLANDO FIGÓN CRUZ

DENISSE MARÍA RAMÍREZ COLMENERO
A01561497
EMMANUEL GONZÁLEZ CALITL
A01320739

18 DE NOVIEMBRE DE 2022

Tecnológico de Monterrey

PARTE 1

PARTE 1

LIMPIEZA

- Se eliminaron columnas con información repetida (variables con las palabras “CALIDAD” y “CUMPLE”) y la columna “SDT_mg/L”.
- Se reemplazaron los datos que tuvieran el signo “>” con su número máximo posible.
- Se imputaron los valores nulos (NaNs) con la mediana de dicha columna.

EXPLORACIÓN DE DATOS

- Se obtuvo información relevante de las variables numéricas con la ayuda de la función .describe() además de su mediana.
- Se graficaron los diagramas de caja y bigote (boxplot) para observar la distribución de los datos y la identificación de outliers.
- Se graficó un mapa de calor (heatmap) para obtener el grado de correlación de las variables numéricas.

```
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ALC_mg/L              1068 non-null   float64
 1   AS_TOT_mg/L           1068 non-null   float64
 2   CD_TOT_mg/L           1068 non-null   float64
 3   COLI_FEC_NMP/100_mL   1068 non-null   float64
 4   CONDUCT_mS/cm         1068 non-null   float64
 5   CR_TOT_mg/L           1068 non-null   float64
 6   DUR_mg/L              1068 non-null   float64
 7   FE_TOT_mg/L           1068 non-null   float64
 8   FLUORUROS_mg/L        1068 non-null   float64
 9   HG_TOT_mg/L           1068 non-null   float64
 10  MN_TOT_mg/L           1068 non-null   float64
 11  N_NO3_mg/L            1068 non-null   float64
 12  PB_TOT_mg/L           1068 non-null   float64
 13  SDT_M_mg/L            1068 non-null   float64
dtypes: float64(14)
memory usage: 116.9 KB
```

Figura 1. Variables numéricas sin datos nulos.

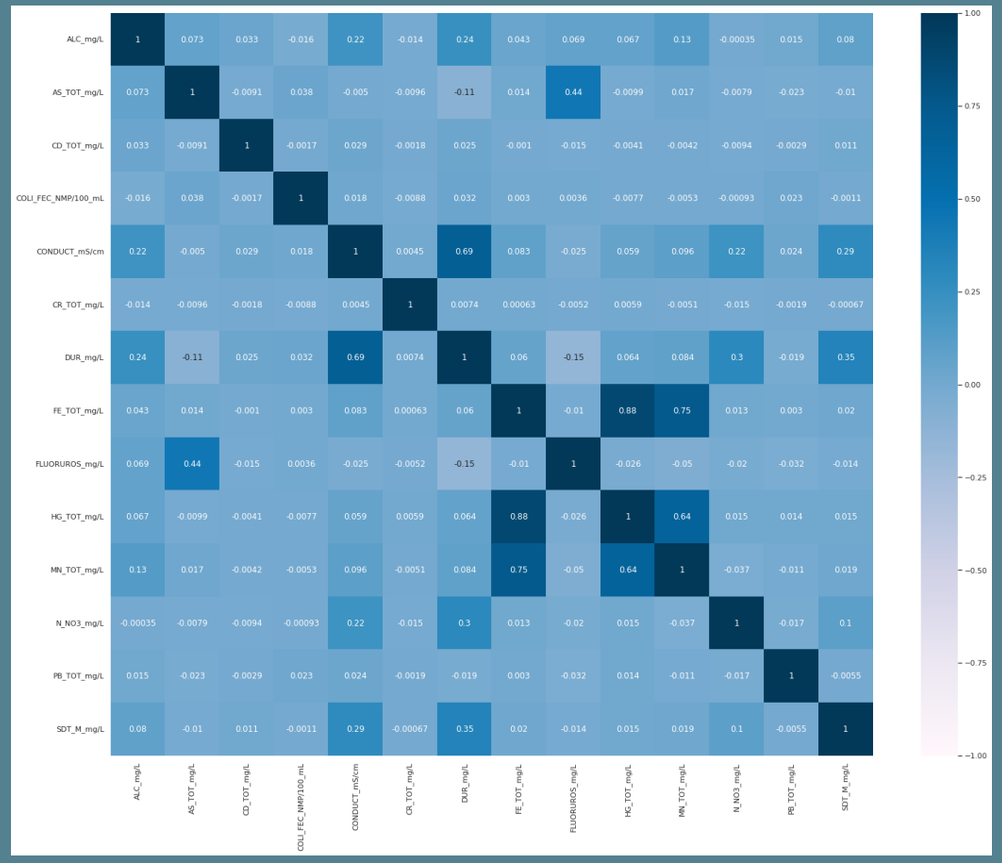
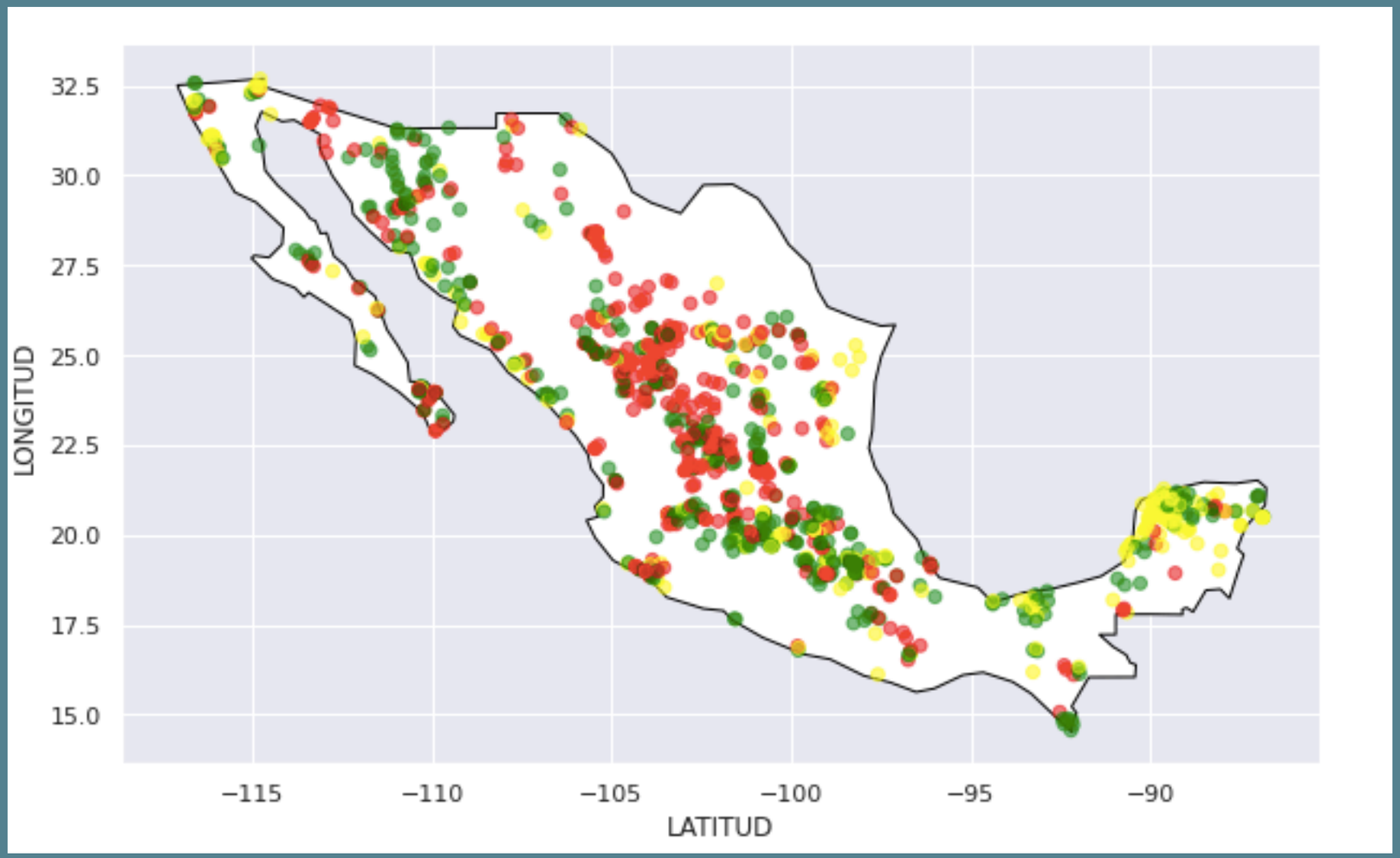


Figura 2. Mapa de calor de la correlación entre variables numéricas.

PARTE 1

ANÁLISIS, AGRUPAMIENTO Y VISUALIZACIÓN CON K MEANS

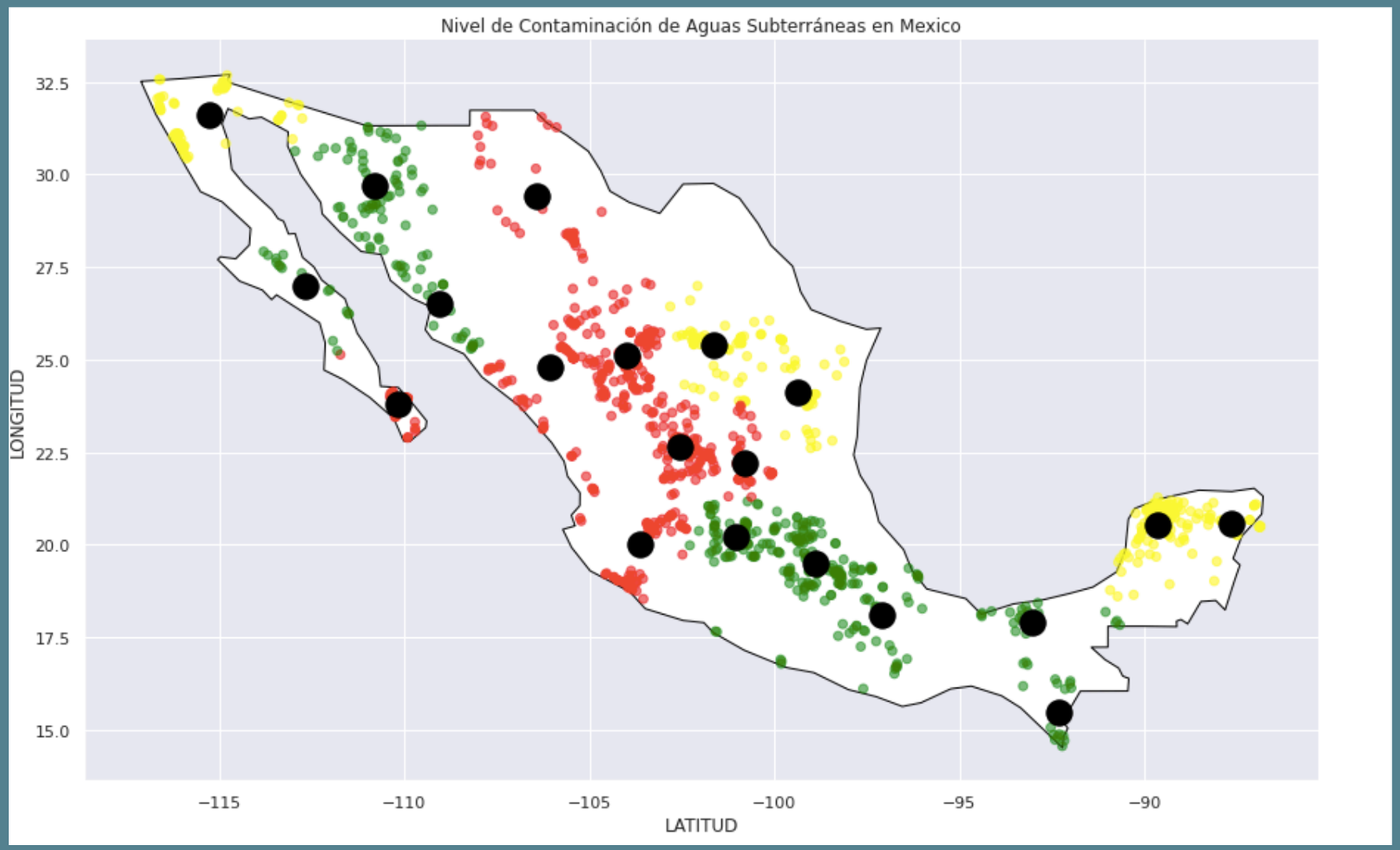
- Para el análisis de los datos se utilizaron herramientas de Geopandas y Shapely para gráficar las muestras de las aguas subterráneas en el mapa la república mexicana. La ubicación de cada muestra fue obtenida utilizando las columnas de “LATITUD” y “LONGITUD”. Los colores indican el nivel de contaminación del agua subterránea en ese sitio, (Verde: bajo, Amarillo: medio, Rojo: alto) de acuerdo a la variable de salida “SEMAFORO”.



PARTE 1

ANÁLISIS, AGRUPAMIENTO Y VISUALIZACIÓN CON K MEANS

- Luego de visualizar los datos en el mapa, se prosiguió a identificar patrones de las muestras según su nivel de contaminación por medio del método de agrupamiento K Means.
- Se seleccionó un número de 20 clusters para que estos se distribuyeran a lo largo del mapa y se obtuvo el nivel de contaminación predominante para cada cluster con la función .groupby().size(). Finalmente cada observación es agrupada en su k-cluster mas cercano y coloreada con el nivel de contaminación predominante en su cluster.



PARTE 2

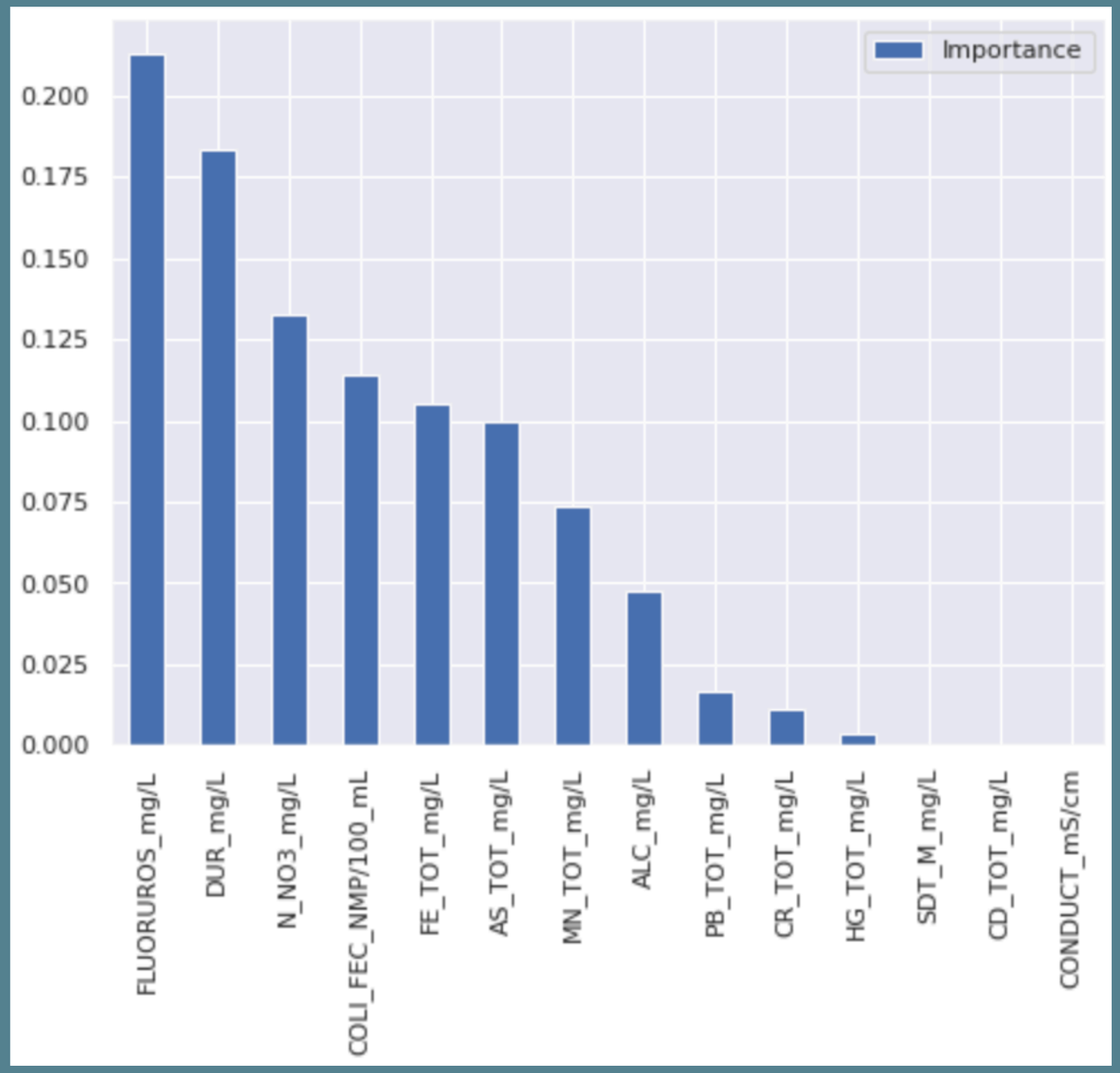
PARTE 2

CLASIFICACIÓN

- Se realiza un análisis general de las características más importantes por medio de un árbol de decisión, para saber qué variables de entrada son las que influyen de mayor manera a la calidad del agua (Semáforo)
- Se seleccionó un número de cinco variables con valores mayores a 0.1 de importancia para delimitar y usar razonablemente los recursos computacionales.

VARIABLES QUE MAYORMENTE INFLUYEN A LA CALIDAD:

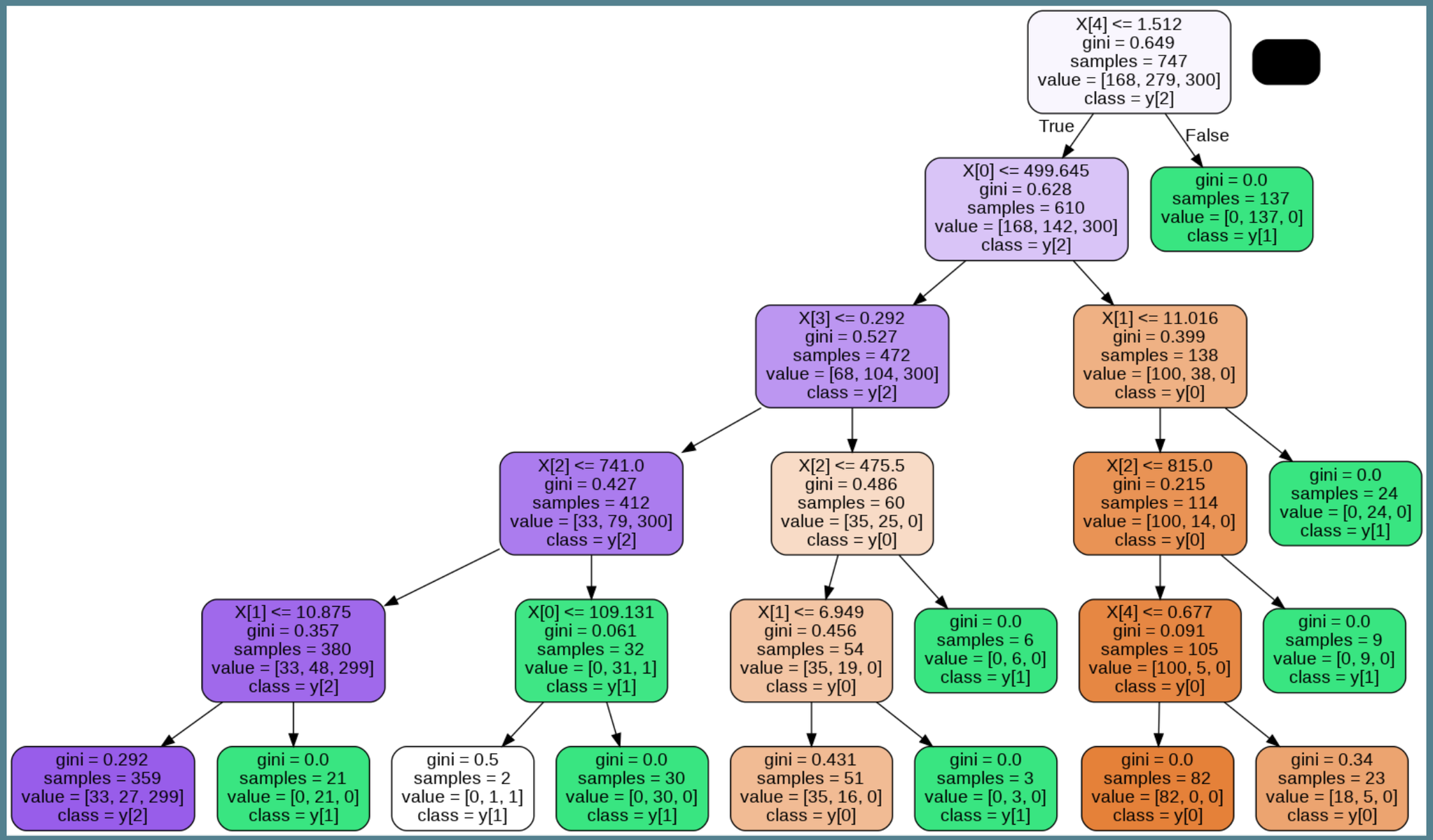
1. Fluoruros totales
2. Dureza total
3. Nitrogeno de nitratos
4. Coliformes Fecales
5. Hierro total



PARTE 2

CLASIFICACIÓN

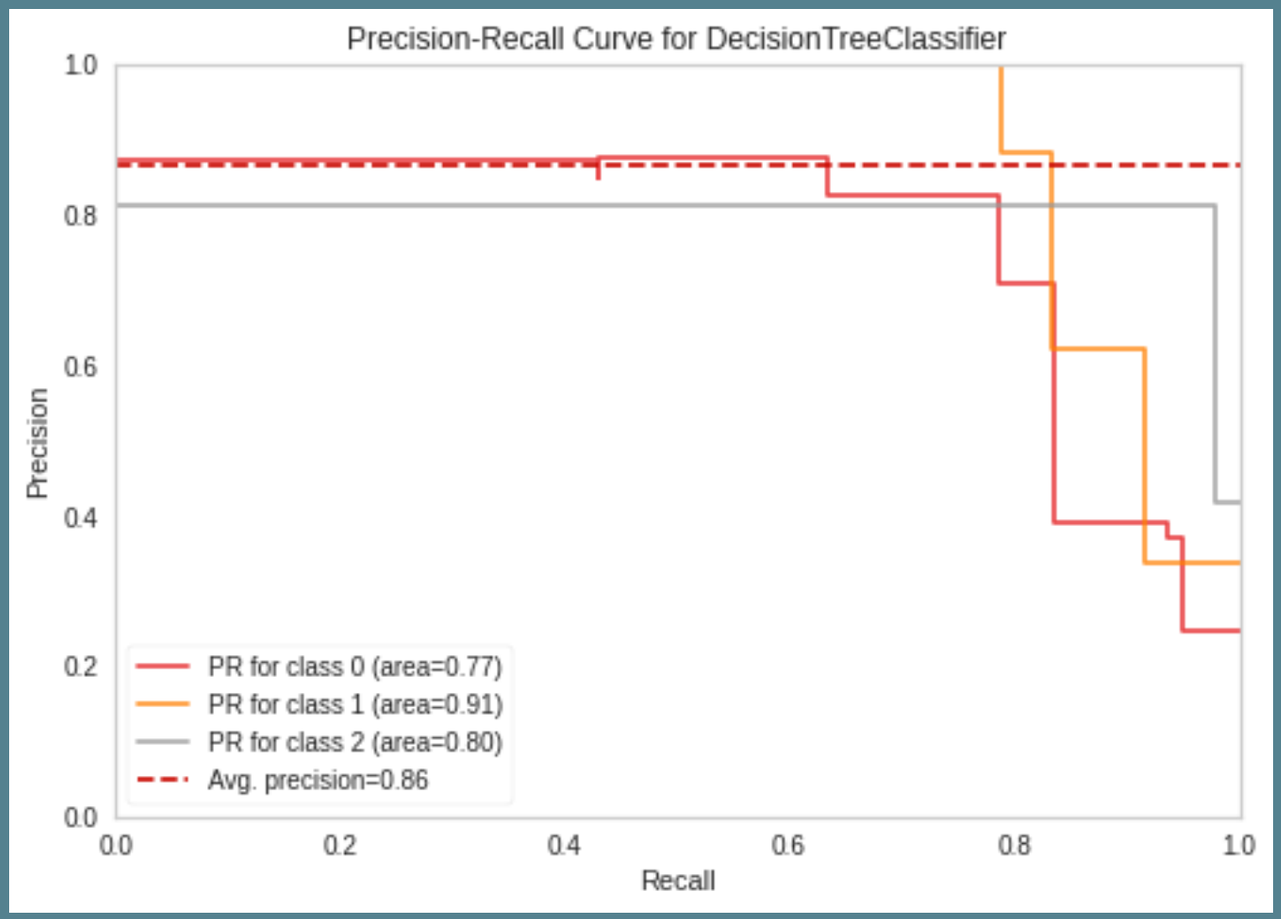
- Gracias a la visualización de nuestro árbol de decisiones podemos notar como cada uno de nuestros features importances contribuyen a la clasificación de las observaciones. Podemos notar que el semáforo en rojo, es la que el modelo clasifica con mayor facilidad, clasificándola desde la primera raíz del decision tree, mientras que para clasificar la clase 0 y 2 (amarillo y verde) le cuesta un poco más.

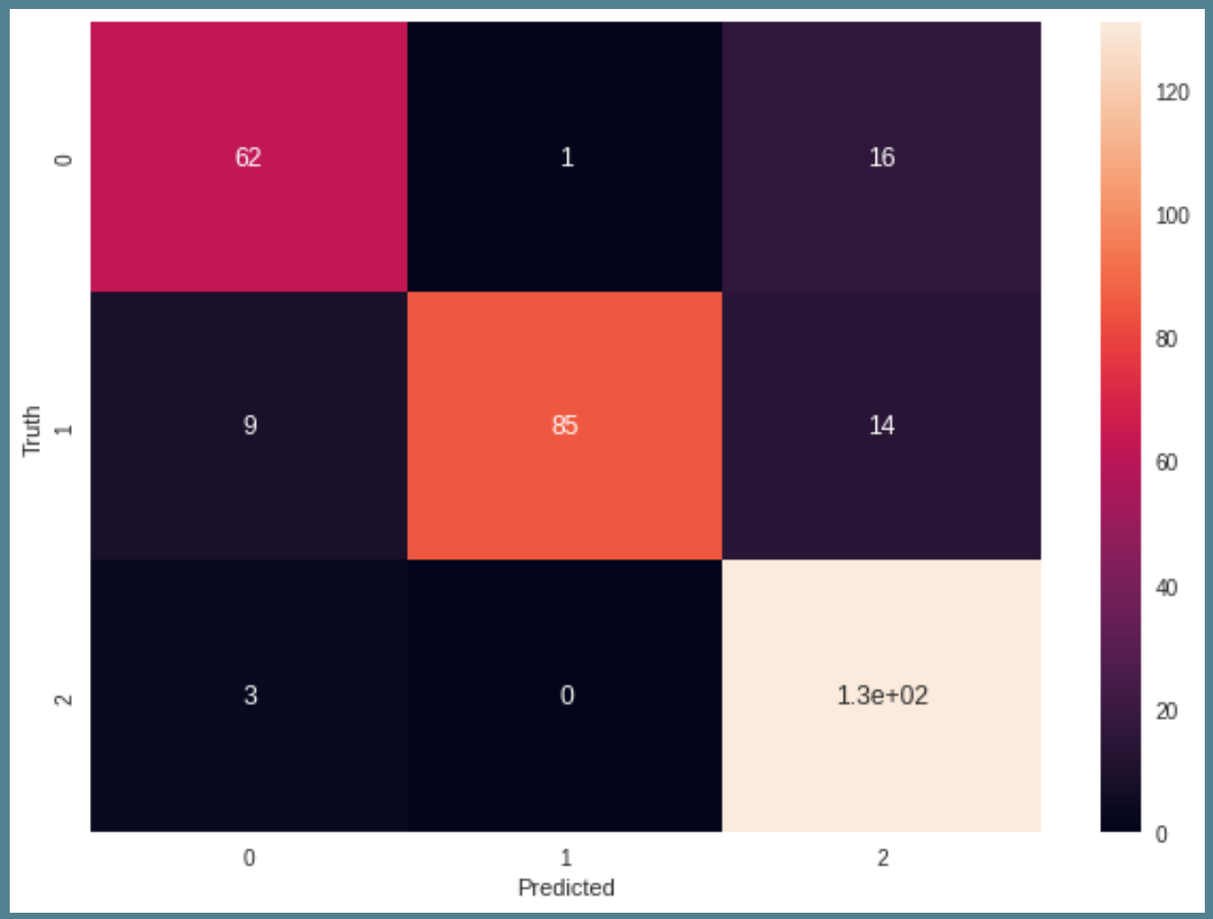


PARTE 2

RESULTADOS

- De la matriz de confusión:
  - De los 79 registros de la clase 0 (Amarillo), 62 los clasificó como tal. (recall = 62 / 62+1+16 = 0.78)
  - De los 108 registros de la clase 1 (Rojo), 85 los clasificó como tal. (recall = 85 / 85+9+14 = 0.79)
  - De los 134 registros de la clase 2 (Verde), 130 los clasificó como tal. (recall = 130 / 130+3 = 0.98)





CONCLUSIONES

Podemos decir que los resultados del reporte de clasificación y la curva PR concuerdan con los resultados de la matriz de confusión. Como comentamos anteriormente tanto Decision Tree como Random Forest tuvieron muy buenos resultados y ambos son buenos modelos para resolver este problema de clasificación, únicamente decidimos visualizar la curva y la matriz de confusión para Decision Tree simplemente porque consideramos que es un modelo más fácil de comprender.

Tecnológico de Monterrey