ACUAS SUBTERRANEAS

Equipo 70:

- Herbert Joadan Romero Villarreal (A01794199)
- Javier Pérez Sanagustín (A01794233)

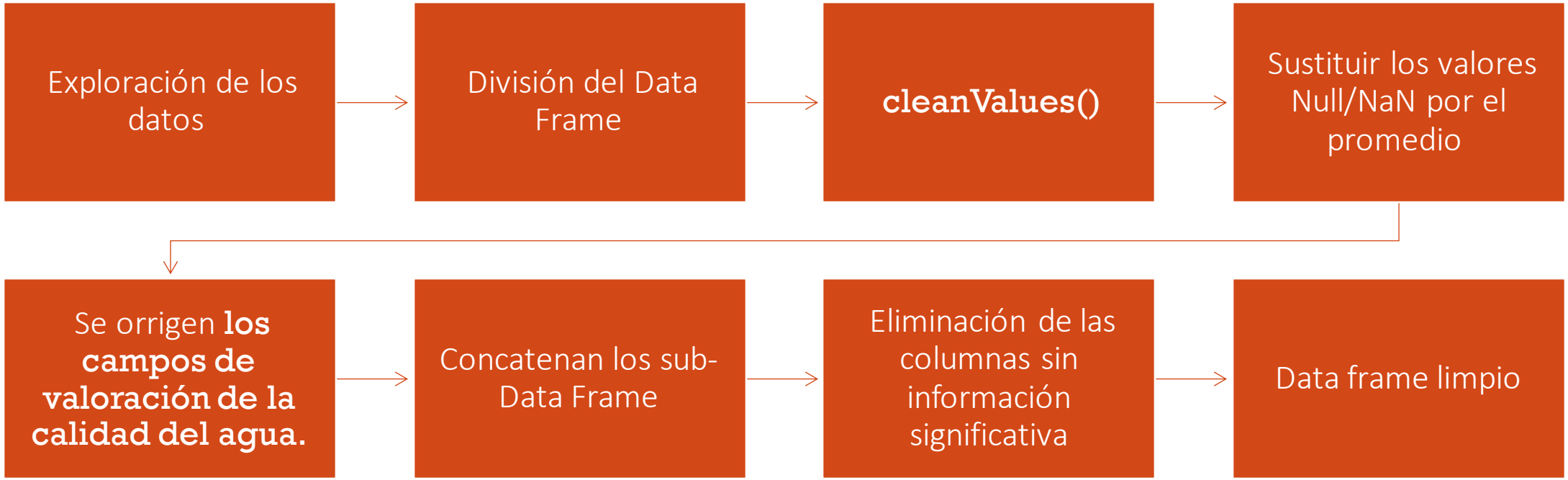
Materia: Ciencia y analítica de datos (Gpo 10)

Profesor titular: María de la Paz Rico Fernández

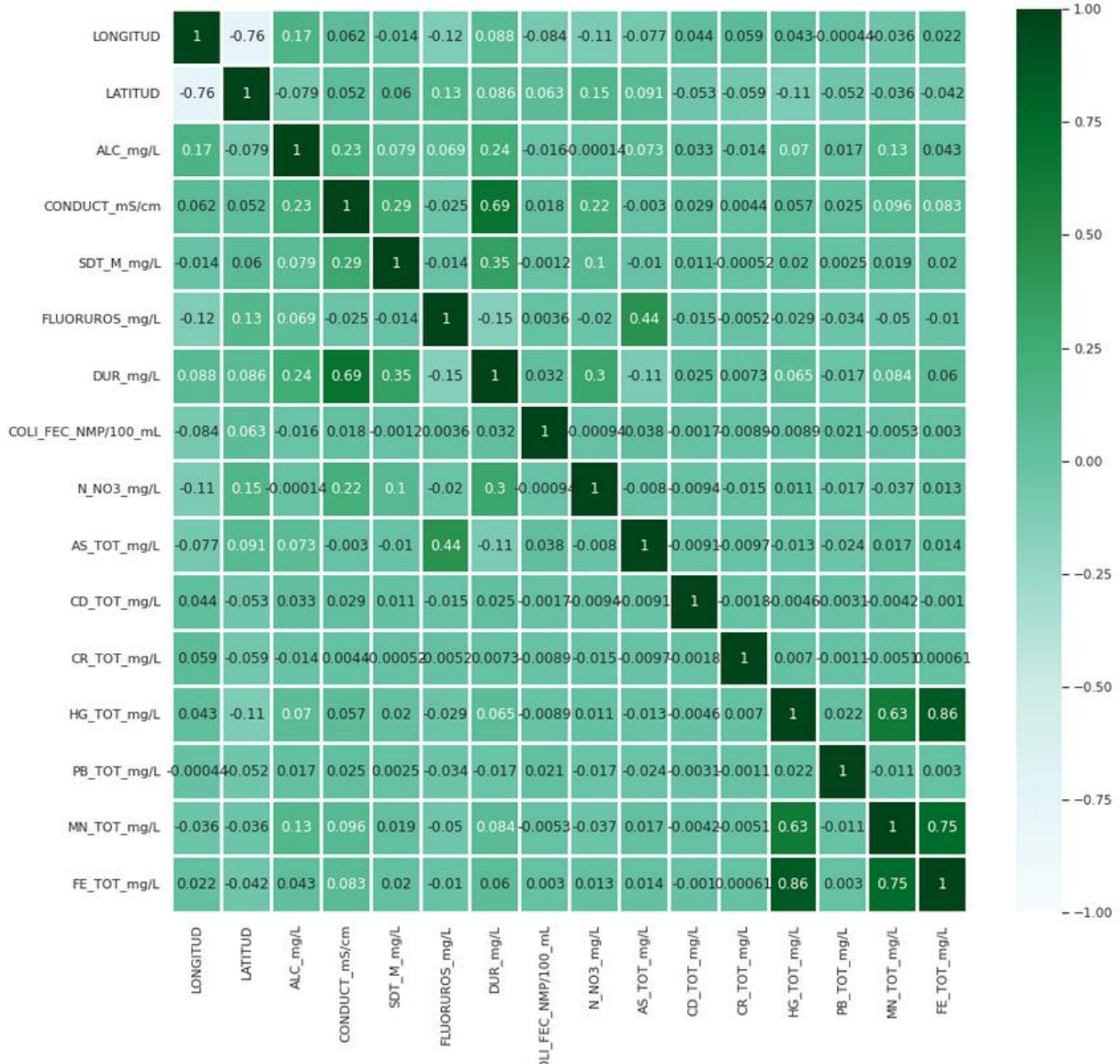
Profesor tutor: Bernardo Charles Canales

Fecha: 15 de noviembre del 2022

LIMPIEZA DE LA BASE DE DATOS



CORRELACIÓN DE VARIABLES

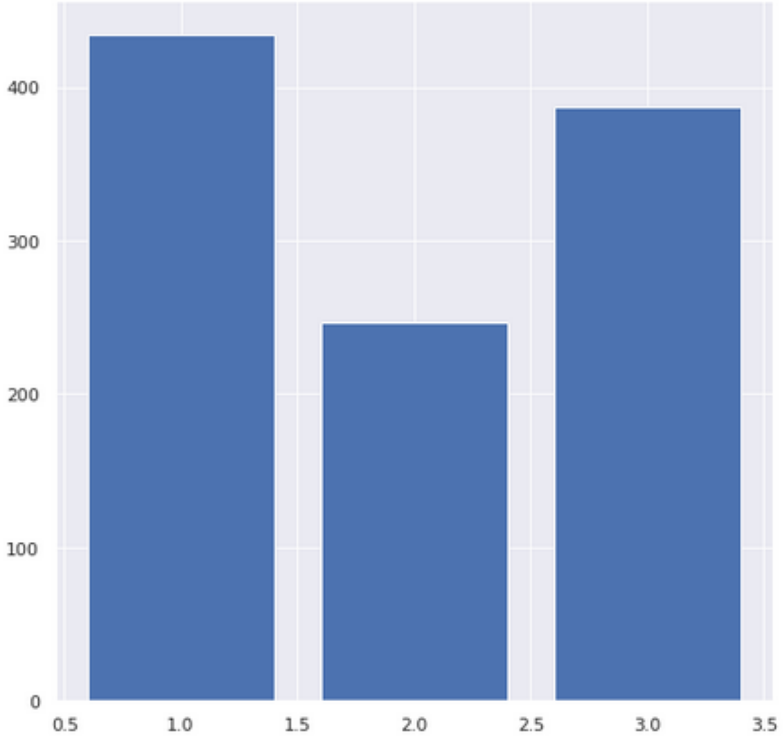


ANÁLISIS K-MEANS

Se construye el data frame con base a la longitud y latitud, para ver la relación entre las coordinadas y la calidad del agua

Disperción de los datos

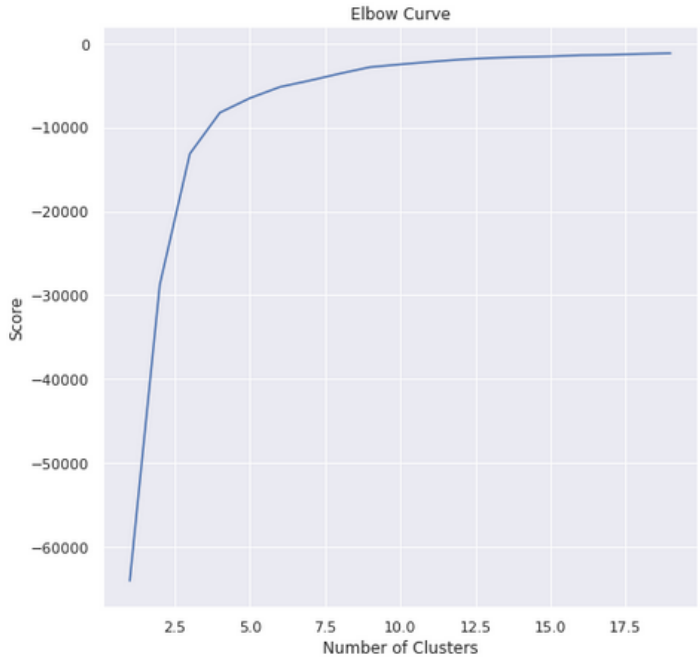
```
sns.set(rc={'figure.figsize':(8,8)})
y = df_sub_clean['SEMAFORO']
c = Counter(y)
plt.bar(c.keys(), c.values())
plt.show()
```



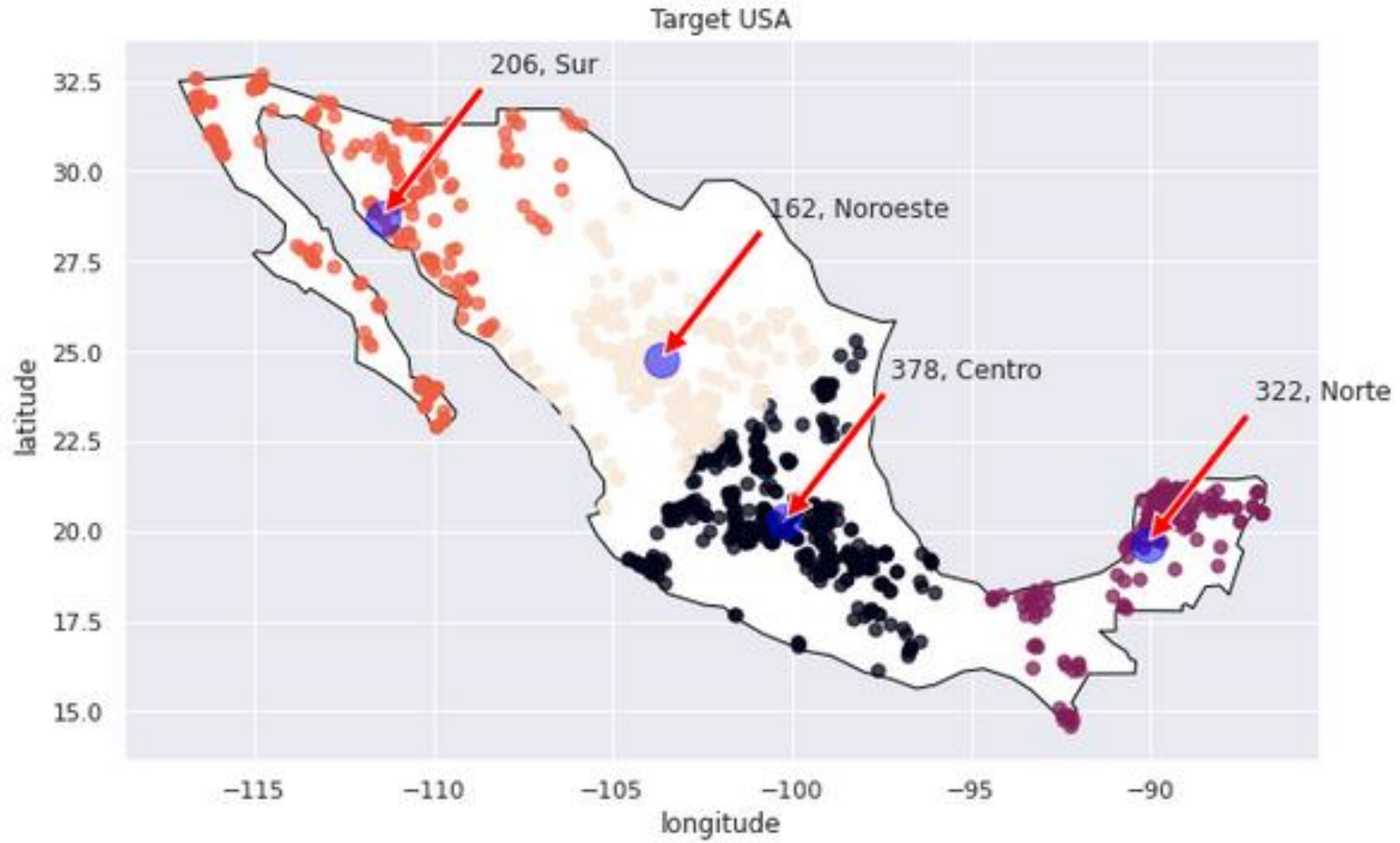
Número de clusters.

```
long_lat_sub = list(zip(gdf.LONGITUD, gdf.LATITUD))

# Grafica de codo
Nc = range(1, 20)
kmeans = [KMeans(n_clusters=i) for i in Nc]
kmeans
score = [kmeans[i].fit(long_lat_sub).score(long_lat_sub) for i in range(len(kmeans))]
score
plt.plot(Nc,score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

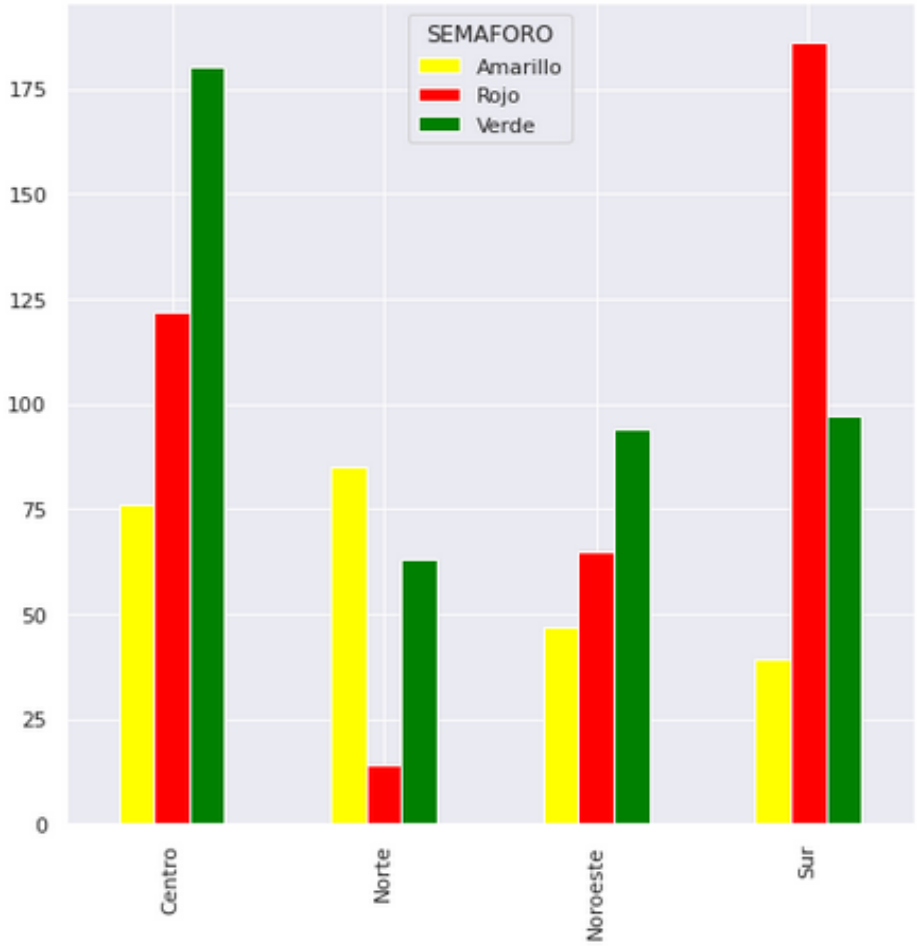


RESULTADOS DE AGRUPAMIENTO DE LATITUDES Y LONGITUDES CON K MEANS EN EL MAPA DE MÉXICO



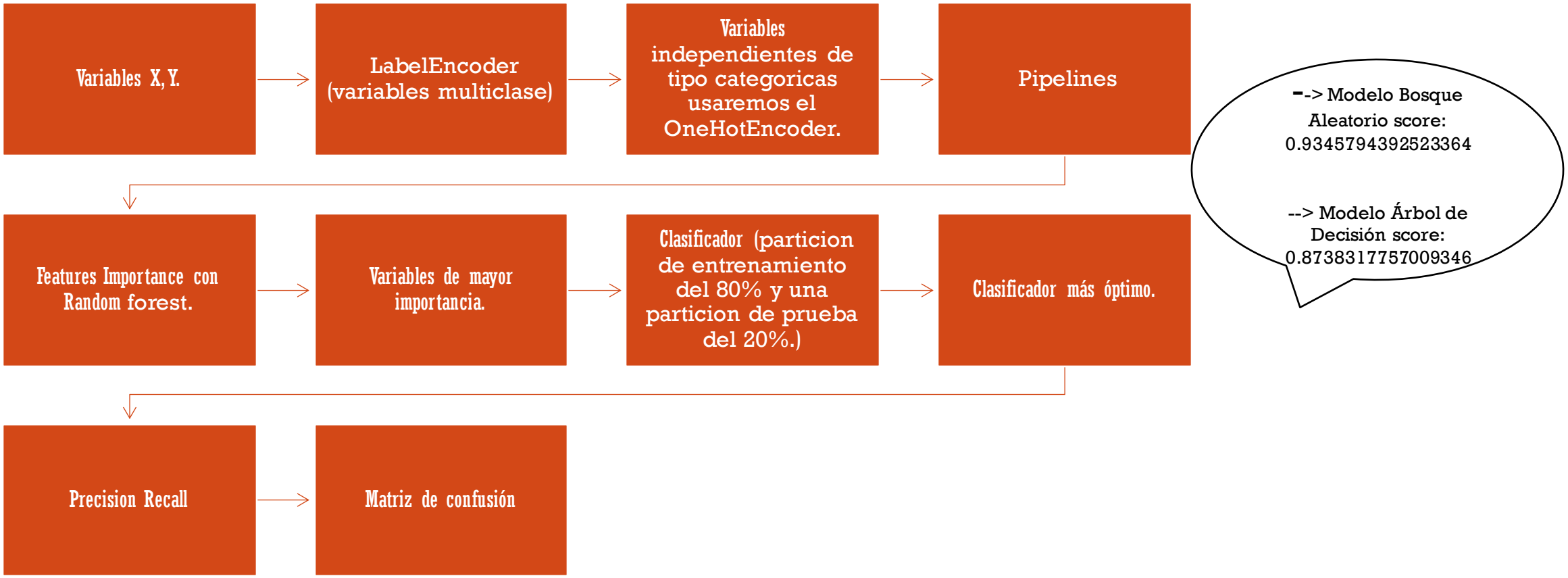
RESULTADOS

Con KMeans haremos el conteo de elementos en las distintas fases del semáforo y asociar la calidad del agua con cada cluster.



- La zona Centro tiene mejor registro de agua de buena calidad, pero también existiría un riesgo alto de que en una zona existiera agua de mala calidad.
- La zona Norte puede ser la zona con mejor calidad del agua, ya que encontramos abundancia de semáforo amarillo/verde y muy poca existencia de registros malos.
- La zona Noroeste presenta el mismo comportamiento que la zona Centro.
- La zona Sur es la zona con peor calidad del agua, encontrando una mayoría de registros en semáforo rojo.

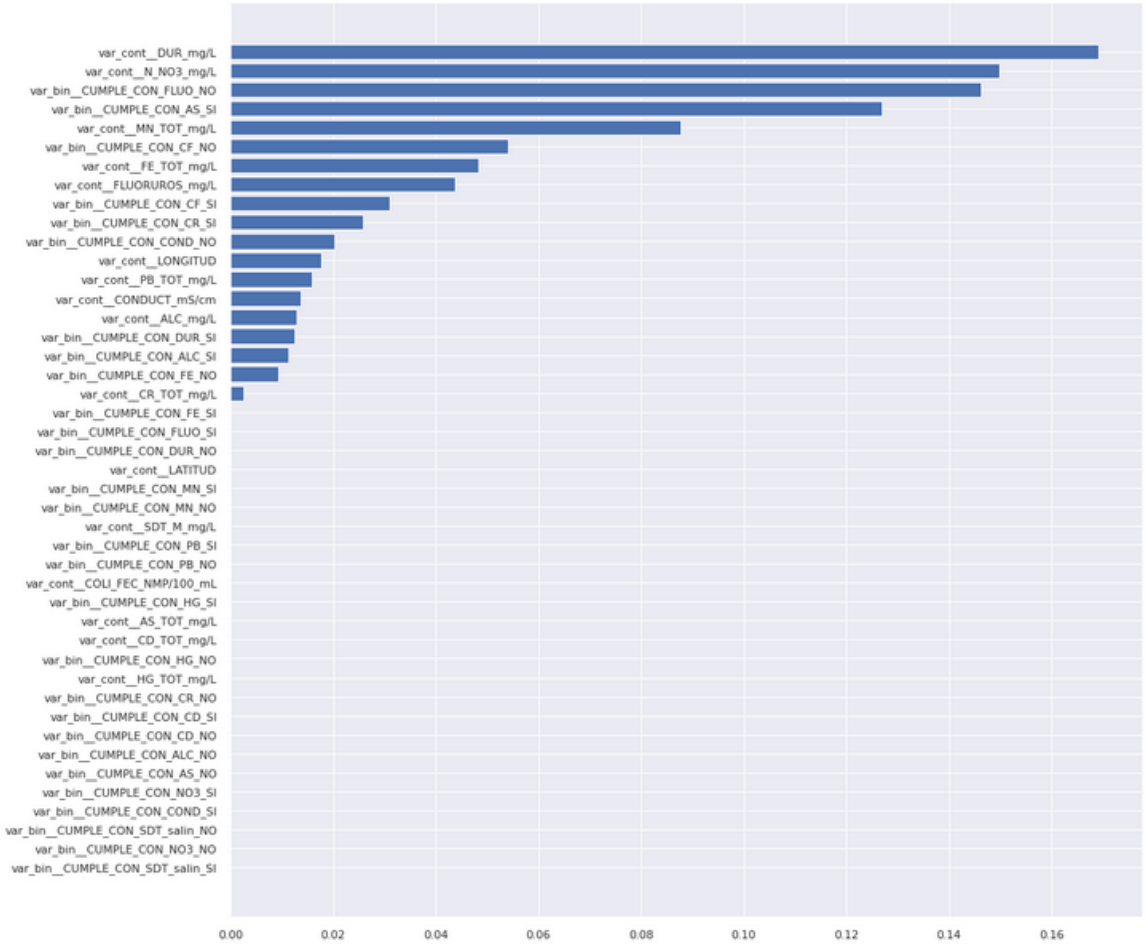
CLASIFICACIÓN



PRECISION-RECALL PARA RANDOM FOREST CLASSIFIER

Feature Importances, set de datos RandomForestClassifier

Variables dependientes de tipo binarias al pasar por el OneHotEnconder




```
from yellowbrick.classifier import PrecisionRecallCurve
# Falta cambiar los nombres de las clases
# Create the visualizer, fit, score, and show it
viz = PrecisionRecallCurve(
    RandomForestClassifier(n_estimators=10),
    per_class=True,
    cmap="Set1"
)
viz.fit(transform_columns.fit_transform(X_train), y_train)
viz.score(transform_columns.fit_transform(X_test), y_test)
viz.show()
```

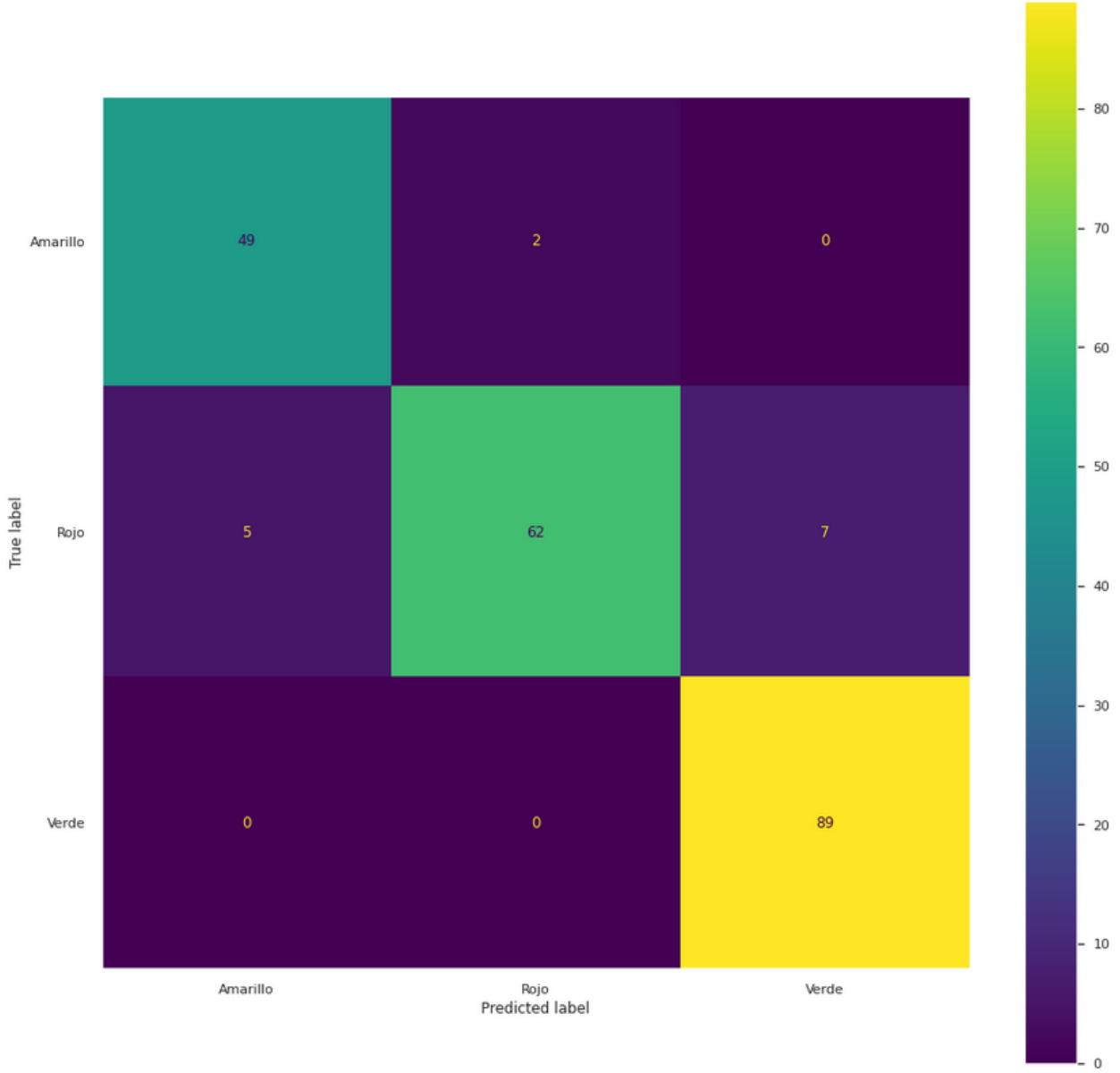
```
/usr/local/lib/python3.7/dist-packages/yellowbrick/classifier/prcurve.py:257: YellowbrickWarning: micro=True is ignored;specify per_class
=False to draw a PR curve after micro-averaging
  YellowbrickWarning,
```

Se tiene una precision promedio de 0.95

MATRIZ DE CONFUSIÓN

```
cm = confusion_matrix(y_test, y_pred, labels=pipes[0][1].classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=le.classes_)
disp.plot()
plt.grid(False)
plt.show()
```

De acuerdo con la Matriz de Confusión podemos corroborar que el número de Falsos Positivos y Falsos Negativos son muy pocos. Esto nos diría que el modelo está clasificando de manera acertada la mayoría de las veces tal cual lo indican las métricas.



CONCLUSIONES

De ambas entregas se concluye:

- La limpieza de la base de datos fue directa, se pudo ahorrar muchos procedimientos debido a la eliminación de columnas redundantes.
- Por medio de las variables numéricas, se detectaron los outliers, para detectar valores atípicos y proceder con la eliminación de columnas.
- Mediante K-Means, se encontró si existe relación entre la calidad del agua y su ubicación. Se demostró con el mapa de la república mexicana que:
- La zona Centro tiene mejor registro de agua de buena calidad, pero también existiría un riesgo alto de que en una zona existiera agua de mala calidad.
- La zona Norte puede ser la zona con mejor calidad del agua, ya que encontramos abundancia de semáforo amarillo/verde y muy poca existencia de registros malos.
- La zona Noroeste presenta el mismo comportamiento que la zona Centro.
- La zona Sur es la zona con peor calidad del agua, encontrando una mayoría de registros en semáforo rojo.
- Ya con la base de datos limpia, se procede a realizar el entrenamiento. El clasificador más óptimo resultó ser:
- Modelo Bosque Aleatorio score:0.9345794392523364
- Modelo Árbol de Decisión score: 0.8738317757009346

Por lo tanto, se usó el clasificador de Bosque Aleatorio.

- En la gráfica de Precision Recall se observa una precisión de 0.95. Lo que indica que el clasificador es óptimo.
- De acuerdo con la Matriz de Confusión, podemos corroborar que el número de Falsos Positivos y Falsos Negativos son muy pocos. Esto nos diría que el modelo está clasificando de manera acertada la mayoría de las veces tal cual lo indican las métricas, y, por lo tanto, las predicciones son más acertadas.
- Con ambas entregas se demostró los procesos, desde la limpieza total de la base de datos, hasta el proceso de entrenamiento y predicciones con el set de datos.