Tecnológico de Monterrey

Reporte Reto Final

Alumnos:

Julio Osvaldo Hernández Bucio A017944366

Juan Antonio Melendres Villa A00369017

Equipo: 3

Materia: Ciencia y analítica de datos (Gpo 10)

Profesor: María de la Paz Rico Fernández.

Fecha: 18 de Noviembre de 2022

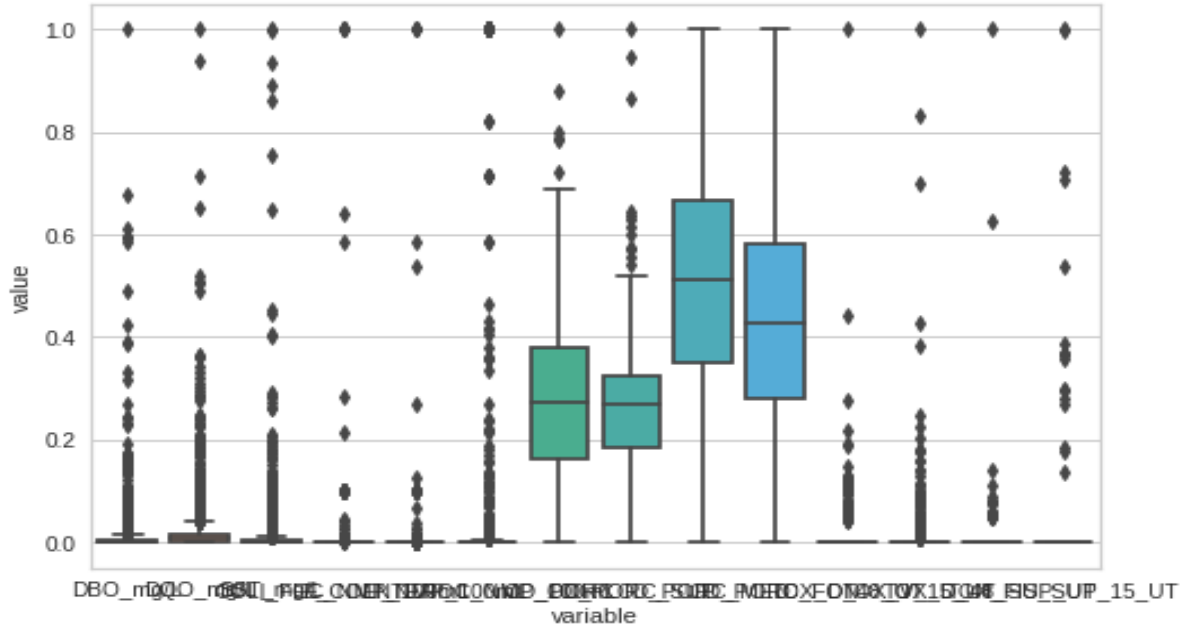
Datos utilizados

Para este reto se seleccionó la base de datos de aguas superficiales, esta contiene información relevante a los diferentes cuerpos de aguas superficiales en la república mexicana que cumplen con las características para formar parte de este grupo. Para el trabajo se utilizó un archivo en formato csv que contiene 3493 registros y un total de 53 columnas con información relevante para determinar si cierto cuerpo de agua puede clasificarse en calidad alta, media o baja, esto a partir del uso de la variable semáforo. Entre las columnas más importantes están las que describen las concentraciones de químicos en el agua (numéricas) en su mayoría presentan concentraciones de tipo mg/L y las categorías que indican la ubicación, el tipo de acuífero y finalmente el semáforo, que para fines del análisis es la variable de clasificación y a predecir.

Análisis y Limpieza

Para el análisis de variables, primero se buscaron las columnas que no tenían un valor significativo para el desafío, como las columnas clave, estado, ubicación, piscina, cuerpo de agua y tipo porque no representan información útil y se puede inferir. También eliminamos algunas columnas numéricas que contenían demasiados datos vacíos, dado que los datos numéricos son importantes para el análisis de nuestra base de datos, se eliminaron las columnas tox_d_48_fon_ut, calidad_tox_d_48_fon, tox_fis_fon_15_ut y calidad_tox_fis_fon_15. Convertimos el contenido de algunas columnas a valores numéricos, así como valores tipo "SÍ" y "NO" en 1 y 0 para manejar mejor los datos.

Boxplot nos ayuda a dibujar un gráfico, y ese gráfico a su vez determina qué columnas son más significativas que otras.

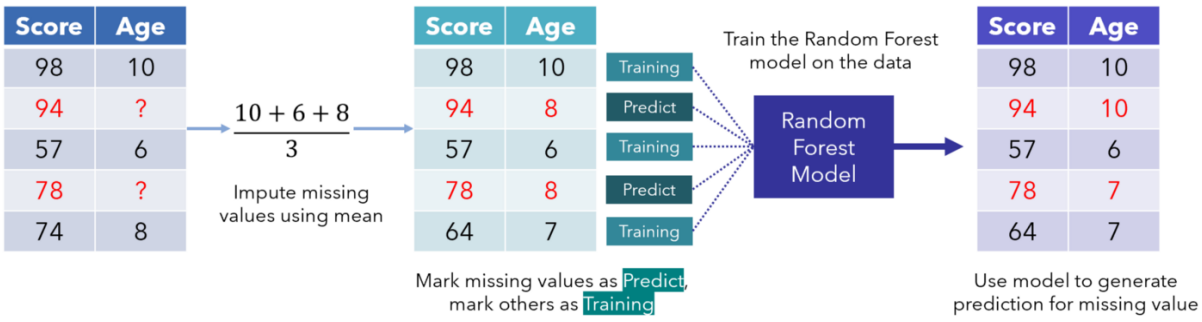


Se obtuvo una lista de variables numéricas finales antes del análisis: ['DBO_mg/L', 'DBO_mg/LMin', 'DQO_mg/L', 'DQO_mg/LMin', 'SST_mg/L', 'SST_mg/LMin', 'COLI_FEC_NMP_100mL', 'COLI_FEC_NMP_100mLMin', 'E_COLI_NMP_100mL', 'E_COLI_NMP_100mLMin', 'ENTEROC_NMP_100mL', 'ENTEROC_NMP_100mLMin', 'OD_PORC', 'OD_PORCMin', 'OD_PORC_SUP', 'OD_PORC_SUPMin', 'OD_PORC_MED', 'OD_PORC_MEDMin', 'OD_PORC_FON', 'OD_PORC_FONMin', 'TOX_D_48_UT', 'TOX_D_48_UTMin', 'TOX_V_15_UT', 'TOX_V_15_UTMin', 'TOX_D_48_SUP_UT', 'TOX_D_48_SUP_UTMin', 'TOX_FIS_SUP_15_UT', 'TOX_FIS_SUP_15_UTMin'].

Cabe mencionar que se utilizaron principalmente las bibliotecas pandas, numpy, sklearn y matplotlib. Pandas se utilizó principalmente para administrar el conjunto de datos, por lo que para el nuevo conjunto de datos se analizó si la mejor opción sería eliminar todos los registros con un valor nulo, este método no era el más adecuado porque se perdería un gran número de datos. Finalmente se decidió reemplazar los valores numéricos con el método de Machine Learning: Missing Forrest.

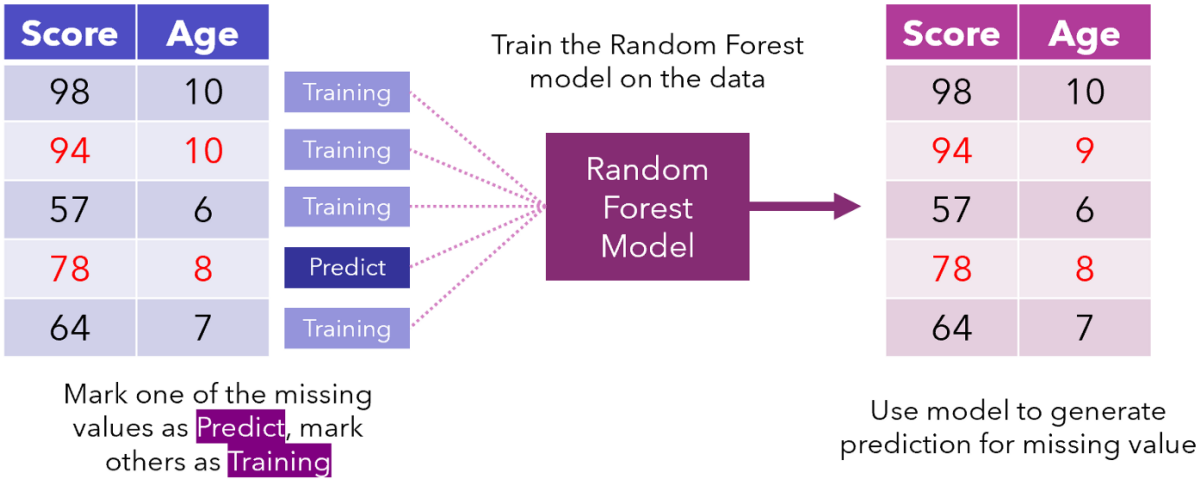
Un poco de este método se explica a continuación:

En primer lugar, los valores que faltan se completan mediante la imputación de mediana/moda. Luego, marcamos los valores que faltan como "Predecir" y los demás como filas de entrenamiento, que se introducen en un modelo de bosque aleatorio entrenado para predecir, en este caso, la edad según la puntuación. La predicción generada para esa fila luego se completa para producir un conjunto de datos transformado.



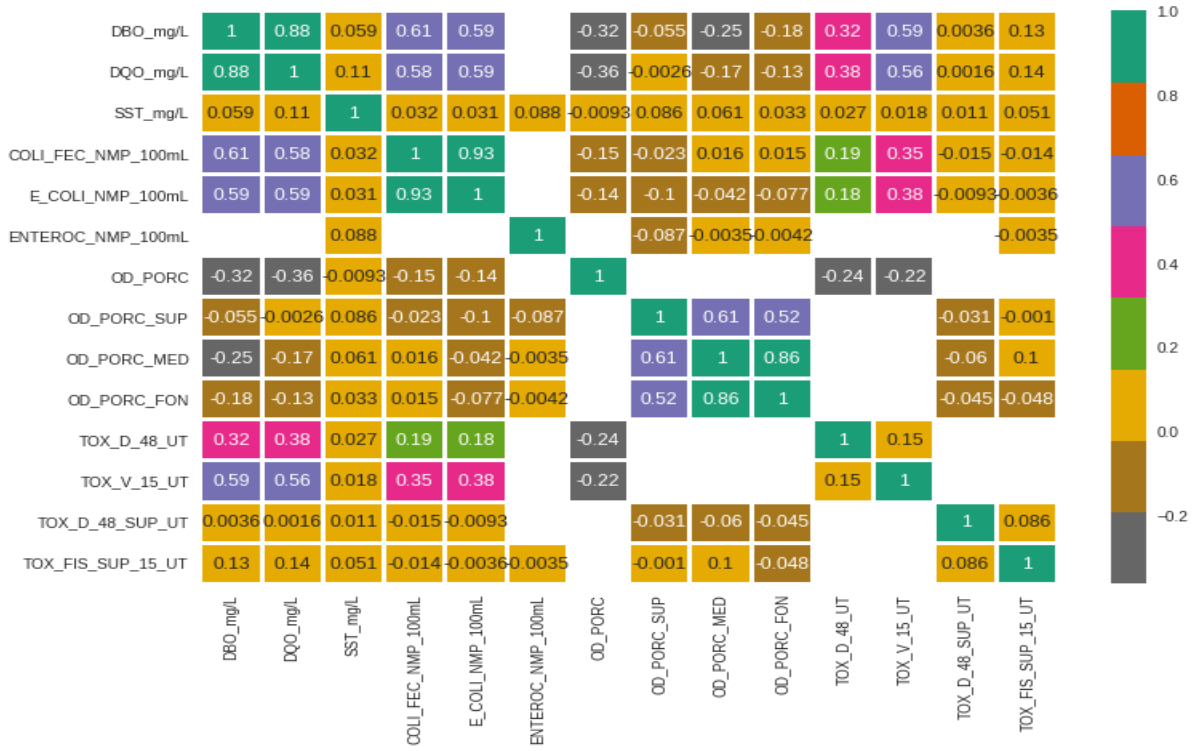
Este proceso de recorrer los puntos de datos faltantes se repite varias veces, y cada iteración mejora con datos cada vez mejores. Es como pararse sobre una pila de rocas mientras se agregan continuamente más para elevarse: el modelo usa su posición actual para elevarse aún más.

El modelo puede decidir en las siguientes iteraciones ajustar las predicciones o mantenerlas iguales.



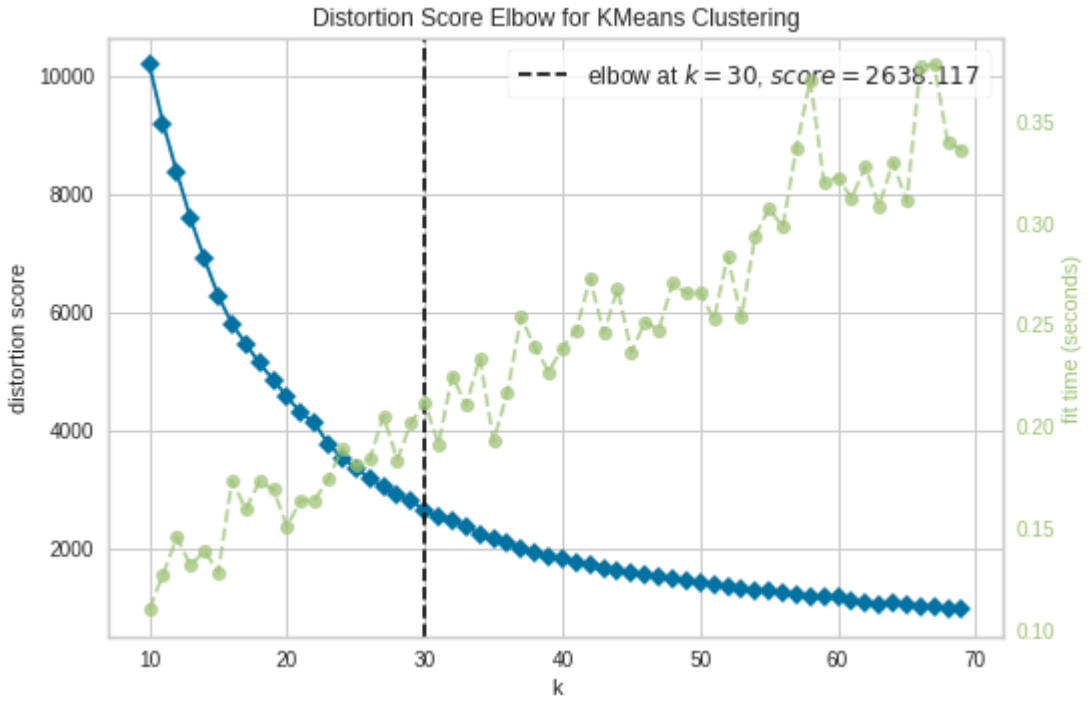
Las iteraciones continúan hasta que se cumplen algunos criterios de parada o después de que haya transcurrido un cierto número de iteraciones. Como regla general, los conjuntos de datos se imputan bien después de cuatro o cinco iteraciones, pero depende del tamaño y la cantidad de datos faltantes.

Se construyó una matriz de correlación simplemente para determinar si era necesario aplicar alguna técnica de reducción de dimensionalidad, aunque algunos valores mostraron bajas correlaciones, se decidió no reducir el número de componentes para no afectar el resultado final. No se muestran correlaciones significativas más allá de lo que realmente querían entender cuáles de estas columnas eran más importantes para los siguientes pasos.



K means

El siguiente paso lógico sería un análisis del vecino más cercano para determinar si existe una relación entre la ubicación y la calidad del agua. Lo primero que debe hacer es determinar el número óptimo de clústeres para agrupar los datos. Se hicieron diferentes pruebas con diferentes valores del número de racimos, pero finalmente se decidió que el número adecuado es el 30, como se puede ver en la siguiente imagen (prueba del codo), aquí puede el punto de inflexión central.

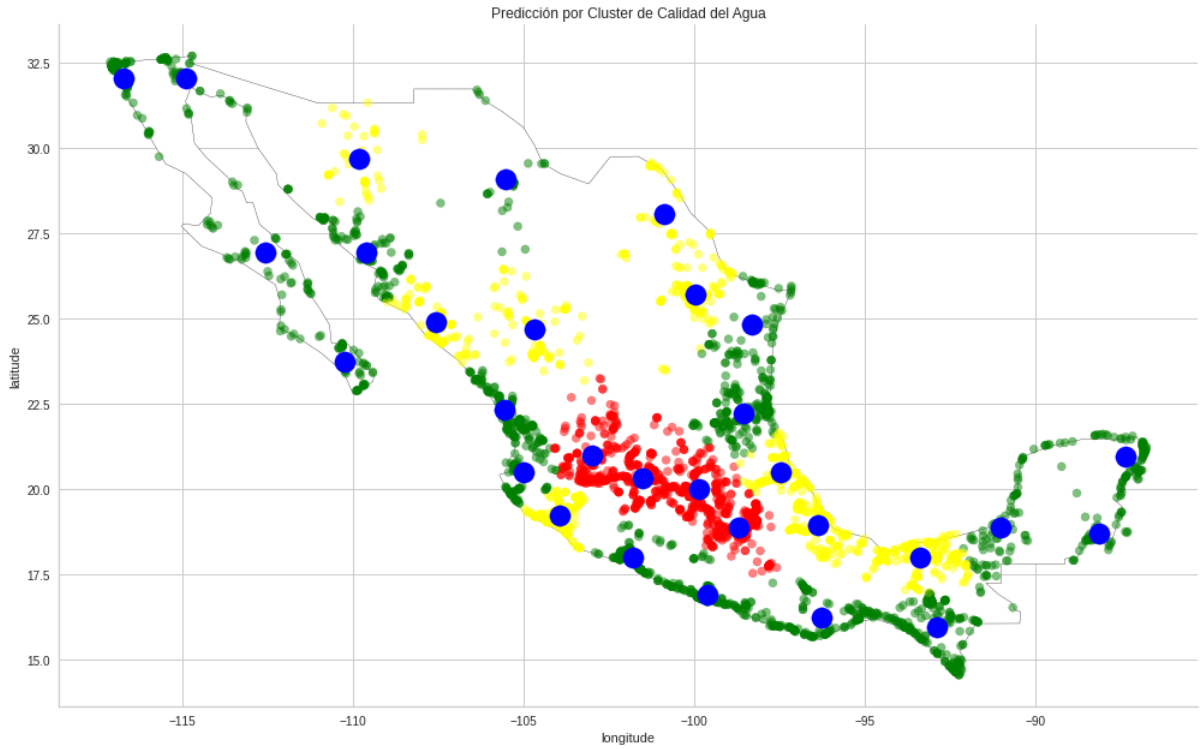


Con los datos determinados se utilizó la función Kmeans para los 30 clusters y fue posible encontrar las coordenadas de los centroides (latitud y longitud) y los puntos de cada región, todo lo cual fue compilado en un mapa soportado por la librería matplotlib.

En las siguiente imagen se muestra la localización de los 30 clusters esta imagen contiene los datos de longitud, latitud, coordenadas, y el numero de cluster.

	0	1	Coordinates	Numero de Cluster
0	19.983565	-99.857764	POINT (-99.85776 19.98356)	0
1	26.955628	-112.578254	POINT (-112.57825 26.95563)	1
2	15.932988	-92.892535	POINT (-92.89254 15.93299)	2
3	20.994410	-103.000208	POINT (-103.00021 20.99441)	3
4	24.888982	-107.582663	POINT (-107.58266 24.88898)	4
5	24.833880	-98.301737	POINT (-98.30174 24.83388)	5
6	20.944526	-87.335133	POINT (-87.33513 20.94453)	6
7	32.039858	-114.904824	POINT (-114.90482 32.03986)	7
8	18.924367	-96.372446	POINT (-96.37245 18.92437)	8
9	18.873404	-98.682674	POINT (-98.68267 18.87340)	9
10	29.685960	-109.814059	POINT (-109.81406 29.68596)	10
11	16.881812	-99.609470	POINT (-99.60947 16.88181)	11
12	19.206394	-103.942796	POINT (-103.94280 19.20639)	12
13	25.691611	-99.975702	POINT (-99.97570 25.69161)	13
14	22.222896	-98.536046	POINT (-98.53605 22.22290)	14
15	16.222521	-96.246602	POINT (-96.24660 16.22252)	15
16	24.686128	-104.671800	POINT (-104.67180 24.68613)	16
17	18.856073	-91.015804	POINT (-91.01580 18.85607)	17
18	20.331763	-101.508437	POINT (-101.50844 20.33176)	18
19	17.972711	-93.377103	POINT (-93.37710 17.97271)	19
20	32.042776	-116.743464	POINT (-116.74346 32.04278)	20
21	29.085653	-105.513873	POINT (-105.51387 29.08565)	21
22	22.342094	-105.558245	POINT (-105.55825 22.34209)	22
23	23.744402	-110.247810	POINT (-110.24781 23.74440)	23
24	28.055104	-100.884570	POINT (-100.88457 28.05510)	24
25	18.690294	-88.122346	POINT (-88.12235 18.69029)	25
26	17.981914	-101.808326	POINT (-101.80833 17.98191)	26
27	20.504083	-105.002706	POINT (-105.00271 20.50408)	27
28	20.501574	-97.457219	POINT (-97.45722 20.50157)	28
29	26.934362	-109.605568	POINT (-109.60557 26.93436)	29

En las siguiente imagen se muestra la localización de los 30 clusters y los cuerpos de agua con su valor de semáforo actual.

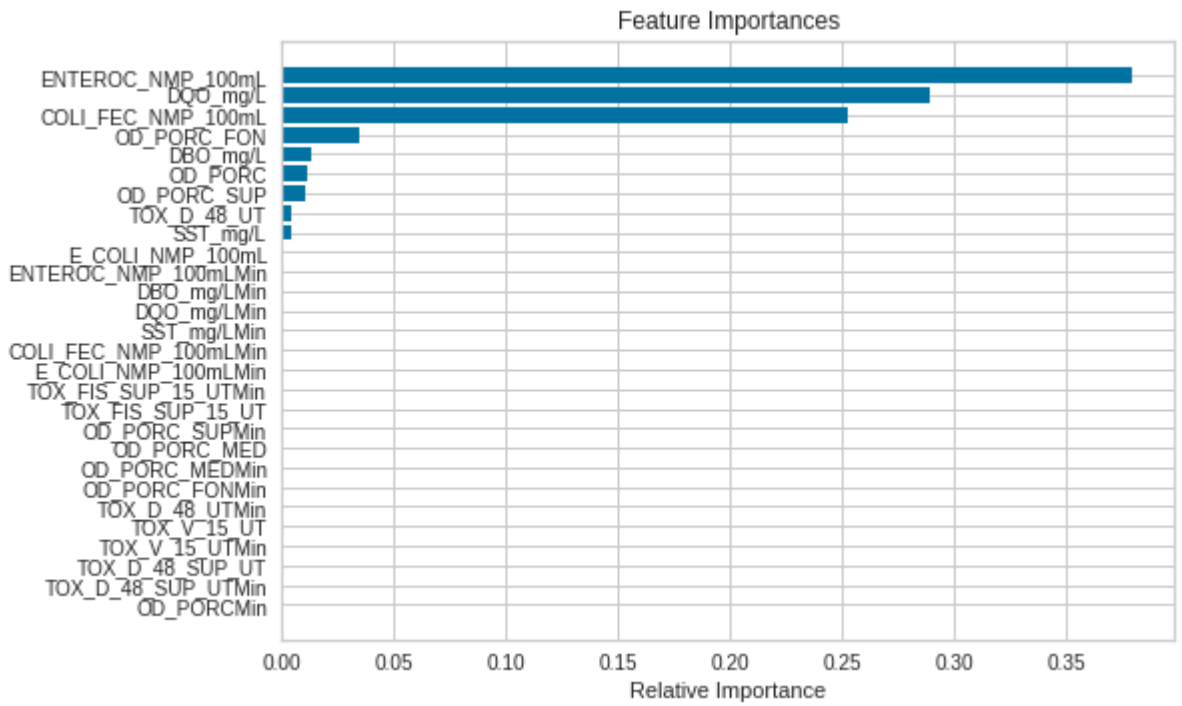


Los resultados fueron prometedores, es posible estimar la distribución equitativa de cuerpos de agua para cada grupo.

Se puede obvservar que la relación de la calidad del agua repecto al centro del país presenta cierto grado de contaminación, por lo que la calidad del agua es mala. Para las peninsulas y algunas costas en el pacifico y el golfo la calidad del agua es buena. A medida que algunos cuerpos de agua se acercan un poco al centro se observa que la calidad del agua es media.

Random forest y Árbol de decisión

Una vez que se obtuvieron los datos, el siguiente paso fue entrenar un modelo que pudiera predecir la variable de salida (semáforos) en función de las entradas. Lo primero que se hizo fue cambiar los valores de salida de “Verde, Amarillo y Rojo” a 0,1 y 2, para lo cual se utilizó el codificador de etiquetas. Además de lo anterior, es interesante encontrar las columnas más importantes, por lo que se utilizaría un modelo de árbol de decisión para dibujar las variables según su importancia. Como nota al margen, el conjunto de datos se dividió en conjuntos de entrenamiento y validación con proporciones de 70 % y 30 % antes de aplicar cualquier modelo de clasificación.

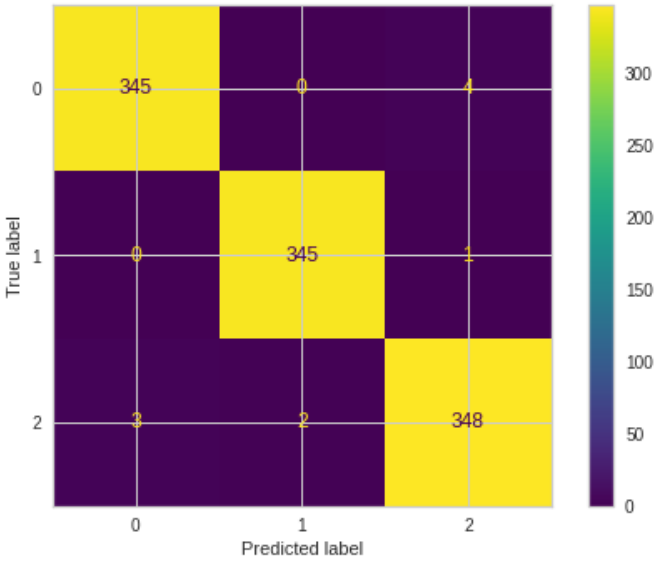


Con base en el gráfico de barras de importancia del atributo, se concluye que las sustancias más importantes son ENTEROC_NMP_100mLMin, COD_mg/L, COLI_FEC_NMP_100mL y OD_PORC_FON. Reconocemos que su presencia es un factor determinante en la variable resultado.

Los modelos fueron entrenados y probados con sus respectivos parámetros, extrayendo las predicciones a partir de los valores extraídos durante la distribución inicial. Las mediciones de rendimiento arrojaron los siguientes resultados.

LogisticRegression 0.7309160305343512
RandomForestClassifier 0.982824427480916
SVC 0.5152671755725191
VotingClassifier 0.9446564885496184

Además de la matriz de confusión.



Los resultados son prometedores, ha habido muchos éxitos, lo que minimiza tanto los falsos positivos como los falsos negativos. En realidad, podemos pensar que esto es una ilusión. Cómo podemos determinar si esto es sobreentrenamiento es probar el modelo con otro tipo de características, tal vez mostrando sus valores de cuerpos de agua de otros países para ver si las clasificaciones son realmente precisas o si los modelos están subentrenados. De la región de México a la formación. Sin embargo, podemos concluir que Random Forest funciona ligeramente mejor cuando se comparan medidas de precisión y recuperación.

La actividad fue muy interesante porque nos ayudó a obtener una comprensión integral de los modelos de clasificación y el significado de K-means. Sería interesante analizar otro tipo de etiquetas para trabajos futuros, y también podríamos confirmar los valores importantes para la clasificación. Comprender los factores que afectan la calidad del agua es el primer paso para crear estrategias de mejora.

Fuentes:

Ye, A. (2020, 31 de agosto). MissForest: The Best Missing Data Imputation Algorithm?. Towards Data Science.
https://towardsdatascience.com/missforest-the-best-missing-data-imputation-algorithm-4d01182aed3