BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

COMPLEJO REGIONAL CENTRO - SAN JOSÉ CHIAPA

Análisis de Datos Históricos para Clasificar Contaminación en el Agua del Estado de Puebla

T E S I S

Que para obtener el título de

Licenciado en Ingeniería en Sistemas y Tecnologías de Información Industrial

P R E S E N T A

Gustavo Jafet Yáñez García

ASESORES DE TESIS

Dra. María Claudia Denicia Carral
Dra. Ana Luisa Ballinas Hernández

Puebla, Pue., Mayo 2024

Agradecimientos

Durante la investigación y elaboración de este trabajo de tesis se pasaron por muchas etapas y modificaciones. Surgieron dudas, errores y problemáticas, sin embargo, no estuve solo, personas que ofrecieron su apoyo, conocimiento y experiencia en el tema me acompañaron en este proceso, por eso este apartado es para ellos.

Quiero agradecer a mi madre, su ayuda, cariño y trato me ayudó en muchas ocasiones donde tuve dificultades y que gracias a ella pude superar.

Quiero agradecer a mi padre por acompañarme, aconsejarme, respetar mi espacio, guiándome no sólo en mi preparación profesional sino en el hombre en el que quiero aspirar a ser.

Quiero agradecer a mi hermana, siendo tan diferentes, pero tan unidos, tomaremos caminos separados, pero siempre nos encontraremos, aunque ella es mayor buscaré siempre ayudarla en lo que pueda y protegerla, así como tú haces por mí. Sin olvidar a mi sobrino que recientemente se integró a la familia, apenas comenzando su camino pero que trataré de guiarlo en lo que yo crea correcto, enseñarle y ayudarlo en cualquier cosa que él necesite.

Quiero agradecer a mis asesoras de tesis la doctora María Claudia Denicia Carral y la doctora Ana Luisa Ballinas Hernández, sin ellas dos este trabajo de tesis no sería posible, compartir sus conocimientos conmigo sobre la materia, no limitándose a sólo el trabajo de tesis sino aclarando mis dudas y dándome sus consejos al área profesional y personal.

Quiero agradecer a mis profesores, compañeros y amigos de universidad, las risas, los nervios, la frustración y paz, los éxitos y fracasos, problemas y soluciones que experimenté desde el inicio hasta el final de mi carrera no tienen precio en mi desarrollo, una etapa universitaria será algo que jamás olvidaré, el estudio es el mayor regalo que mis padres pudieron ofrecerme y las personas que estuvieron involucradas las que hicieron más especial ese regalo.

Resumen

La problemática de la contaminación de los recursos naturales no renovables afecta a todas las personas, el agua es uno de estos recursos que más problemas de contaminación presenta, existen distintos métodos para poder clasificar el agua con contaminación a partir de la medición de diferentes parámetros como son parámetros físicos, químicos, metales, entre otros. La Comisión Nacional del Agua (CONAGUA) es el organismo en México encargado de medir distintos parámetros en los cuerpos de agua, para esta tesis se utilizaron datos provenientes de esta fuente para analizar la contaminación del agua en el estado de Puebla. Se utilizó minería de datos para clasificar los cuerpos de agua como contaminados y no contaminados a partir de la cantidad de oxígeno disuelto en el agua y de la combinación de otros parámetros como el pH, la temperatura, coliformes fecales, entre otros. Los resultados obtenidos demuestran que utilizando algoritmos como árboles de decisión se obtiene un F-measure superior al 78%, lo que demuestra que los métodos de clasificación automática unidos a los procesos de preparación y limpieza de datos de la minería de datos son una buena alternativa para clasificar la contaminación del agua.

Contenido

Agradecimientos I

Resumen II

Lista de Imágenes VI

Lista de Ecuaciones VI

Lista de Tablas VII

Lista de Tablas Anexadas VII

1. Introducción 1

1.1 Planteamiento del problema 1

1.2 Objetivos 1

1.3 Pregunta de investigación y/o Hipótesis 2

1.4 Justificación 2

1.5 Alcances y limitaciones 2

1.6 Organización de la tesis 3

2. Marco Histórico 5

2.1 Antecedentes históricos 5

2.2 Estado del arte 5

3. Marco Teórico 8

3.1 Modelo CRISP-DM 8

3.2 Algoritmos para clasificación de minería de datos 13

3.2.1 Arboles de decisión 13

3.2.2 K-NN 14

3.2.3 Naive Bayes 16

3.3 Hidrología de Puebla 17

3.3.1 Tipos de cuerpos de agua 20

3.3.2 Contaminantes más frecuentes 21

3.3.3 Contaminantes menos frecuentes 22
3.4 Métodos de evaluación para medir la contaminación del agua. 23
4. Marco metodológico. 26
4.1 Conociendo los datos 26
4.2 Limpieza de los datos 31
4.2.1 Homogenizar datos 31
4.3 Selección de parámetros a descartar 32
4.4 Casos Experimentales 34
4.5 Modelo de procesamiento 34
4.5.1 Pre-procesamiento 34
4.5.2 Procesamiento 34
4.6 Resultados 36
4.6.1 Experimento parámetros físicos 39
4.6.2 Experimento parámetros químicos 39
4.6.3 Experimento parámetros biológicos 40
4.6.4 Experimento parámetros metales pesados 41
4.6.5 Experimento parámetros físicos-químicos 41
4.6.6 Experimento parámetros físicos-metales pesados 42
4.6.7 Experimento parámetros químico-metales pesados 42
4.6.8 Experimento parámetros biológicos-metales pesados 43
4.6.9 Experimento parámetros biológicos-químicos 43
4.6.10 Experimento de todos los parámetros juntos 44
5. Conclusiones y trabajo a futuro 45
5.1 Conclusión 45
5.2 Trabajo a futuro 45
6. Bibliografía 47

7. Anexos 51

Lista de Imágenes

Imagen 1. Diagrama del modelo CRISP-MD 8

Imagen 2. Diagrama de flujo en Árbol de decisión 13

Imagen 3. Mapa de la Red Hidrográfica Digital de Puebla, Escala 1:250 000 (Obtenido de: INEGI - CONAGUA, 2007) 18

Imagen 4. *Hidrología del estado de Puebla por zonas.(Obtenida de: INEGI 2006* (Tulcingo, 2018 - 2021) *)* 19

Imagen 5. Montañas (Obtenida de Pixabay, bajo licencia CC) 20

Imagen 6. Pantano (Obtenida de: Pixabay, bajo licencia CC) 20

Imagen 7*. Cueva (Obtenida de: Pixabay, bajo licencia CC)* 20

Imagen 8. *Lago (Obtenida de: Pixabay, bajo licencia CC)* 20

Imagen 9. *Páramo Paisaje (Obtenida de: Pixabay, bajo licencia CC)* 20

Imagen 10. Contaminantes con mayor número de registros por año 30

Imagen 11. Contaminantes con menor número de perdida por meses 30

Imagen 12. Pre-procesamiento de los datos con operadores en RapidMiner 31

Imagen 13. Selección de todos los elementos que sean contaminantes 32

Imagen 14. Vista estadística de contaminantes en RapidMiner 33

Imagen 15. Limpieza de datos 35

Imagen 16. Vista interior en operador de Validation en RapidMiner 36

Imagen 17. Modelo de procesamiento de datos 36

Lista de Ecuaciones

Ecuación 1. Naive Bayes 16

Ecuación 2. Reemplazo de valores para Naive Bayes 16

Lista de Tablas

Tabla 1. Resumen del estado del arte.....7

Tabla 2. Tipos de cuerpos de agua y su descripción 20

Tabla 3. Contaminantes más frecuentes.....22

Tabla 4. Contaminantes menos frecuentes.....23

Tabla 5. Métodos para medida de calidad del agua.....25

Tabla 6. Parámetros usados para el registro de los contaminantes 27

Tabla 7. Contaminantes más frecuentes organizados por año 28

Tabla 8. Contaminantes más frecuentes registrados por mes 29

Tabla 9. Clases para los datos en operador de Discretización 35

Tabla 10. Número de registros que tienen un porcentaje de pérdida mayor al 90% 37

Tabla 11. Porcentaje de acierto para los experimentos y sus contaminantes. 38

Lista de Tablas Anexadas

Anexo Tabla 1. Nombre de sitios divididos en rangos 52

Anexo Tabla 2. Todos los contaminantes 60

Anexo Tabla 3. Experimentos con contaminantes y respectivos porcentajes de perdida..64

Anexo Tabla 4. Experimentos planteados para estimar la calidad del agua.....66

1. Introducción

1.1 Planteamiento del problema

La contaminación del agua genera grandes problemas en la humanidad, tales como, la escasez de recursos naturales, la reducción y disminución de la calidad de los productos agrícolas destinados a la alimentación, enfermedades cutáneas ocasionadas por un mal control en la calidad del agua, así como problemas digestivos en la preparación de alimentos. Esta contaminación se origina por diferentes factores, tales como, empresas que no son reguladas en el proceso de desechar residuos, escurrimientos de material contaminante en los campos de riego cercanos a los cuerpos de agua, actos irresponsables por parte de ciudadanos para deshacerse de sus desperdicios rápidamente. Existen organismos gubernamentales encargados de hacer toma de muestreo en diversos puntos de los distintos cuerpos de agua en el estado de Puebla. Se han obtenido una gran cantidad de registros en varios años a través de sitios web de forma gratuita, sin embargo, estos datos sin un procesamiento adecuado no son de utilidad debido a que requieren técnicas como las de minería de datos para poder indagar en las causas y efectos de la contaminación en el agua.

1.2 Objetivos

Los objetivos que se buscan en este trabajo son los siguientes:

Objetivo General.

Aplicar técnicas de analítica de datos en registros históricos de la contaminación del agua en el estado de Puebla para clasificar cuerpos de agua contaminados mediante el uso de algoritmos de minería de datos.

Objetivos Específicos.

- Recolectar datos históricos sobre contaminación de cuerpos de agua del estado de Puebla a través del uso de repositorios públicos.
- Aplicar métodos de pre-procesamiento para limpiar y normalizar los datos para una correcta aplicación de los procesos de minería de datos.

- Aplicar distintos modelos de clasificación a los datos para identificar los patrones de los contaminantes del agua.
- Analizar los resultados obtenidos por los algoritmos de clasificación para determinar cuál es el mejor y las directrices futuras de la investigación.

1.3 Pregunta de investigación y/o Hipótesis

La pregunta de investigación para este trabajo de tesis es la siguientes:

1. ¿Cómo puede identificarse la contaminación del agua a través de modelos de minería de datos a partir de registros históricos de los cuerpos de agua del estado de Puebla?

A partir de esta pregunta se deriva la hipótesis:

Los métodos de minería de datos aplicados a datos históricos del agua del estado de Puebla son eficaces para clasificar los cuerpos de agua contaminados.

1.4 Justificación

El avance tecnológico en el estado de Puebla requiere de avances al mismo nivel en diferentes áreas, tales como, la agronomía, transporte, industria y la demografía. Siendo esta última la que más relevancia tiene ya que la población hace que el resto de las áreas se pongan en operación en el desarrollo de nuevos productos, estudios para mejorar dichos productos, optimización de procesos, construcción de viviendas por el aumento en la población. Todas estas acciones tienen un gran impacto en el agua debido a que cambian su composición, extracción, concentración de diferentes parámetros como indicadores de su calidad. Esto crea un bucle donde se extrae el agua, se emplea en alguna actividad y los residuos acaban contaminando los mantos acuíferos y deteriorando todo el ciclo.

1.5 Alcances y limitaciones

En este trabajo se realizará un análisis de los datos relacionados con la contaminación de los cuerpos de agua de Puebla, los datos son obtenidos de la Comisión Nacional del Agua (CONAGUA) y son reportes de las mediciones de distintos tipos de contaminantes presentes en ríos, lagos y lagunas recolectados a lo largo de varios años. Los alcances de este trabajo se resumen en los siguientes puntos:

- Se realizará un análisis de los datos de la CONAGUA para describir los datos a partir de sus características por medio de estadística descriptiva.
- Se aplicarán algoritmos de minería de datos para clasificar los cuerpos de agua como contaminados y no contaminados a partir de la selección de variables que son de incidencia en la contaminación de un cuerpo de agua.

Por lo tanto, el alcance principal de este trabajo será la selección de características como un número de registros aceptable, que no sobrepase el 30% de pérdida en los distintos contaminantes para una correcta clasificación de la contaminación del agua en Puebla, aunque no se descarta que esté método pueda ser replicable en otros países.

En cuanto a las limitaciones del trabajo, se pueden nombrar las siguientes: el espacio, ya que solo se analizaron datos del estado de Puebla; la temporalidad porque se analizaron datos de un periodo de tiempo determinado; los datos, ya que algunos registros no están completos porque se utilizaron datos de una fuente externa, por lo que la recolección de datos no pudo ser controlada. Finalmente, el trabajo se limitó al estudio de tres algoritmos de minería de datos, aun cuando existe una gran variedad de ellos.

1.6 Organización de la tesis

En el capítulo 1 se presenta el planteamiento del problema, los objetivos, preguntas de investigación, hipótesis, justificación y por último alcances y limitaciones. Todos estos elementos sirven de guía para la investigación y para las metas que busca cumplir el trabajo.

En el capítulo 2 se recopila información de trabajos relacionados con la búsqueda de patrones mediante el uso de minería de datos identificando el ámbito al que van dirigidos, así como los objetivos logrados.

El capítulo 3 presentan los fundamentos teóricos, una descripción del área de estudio, las principales características específicas de los elementos analizados, conceptos sobre análisis de datos y algunos métodos para medir la calidad del agua.

El capítulo 4 se presenta la metodología desarrollada. Inicia con un análisis y explicación de los datos utilizados, el planteamiento de algunos casos experimentales, los resultados

obtenidos tras el procesamiento de los datos, la aplicación de métodos de minería de datos seleccionados para cada uno de los casos experimentales planteados.

El capítulo 5 se muestras las principales conclusiones del presente trabajo, así como el trabajo a futuro derivado de la investigación.

2. Marco Histórico

2.1 Antecedentes históricos

El agua es un recurso natural no renovable y muy importante del cual dependen actividades como la agronomía, ganadería, actividades industriales que incluyen fabricación de productos, manufactura de prendas, forja de herramientas que faciliten las tareas cotidianas. Además, se usa en diversas actividades de las personas tales como, la higiene personal, cuidado de la vegetación, cocina, limpieza, entretenimiento, medicina, entre otros (Fuertes, 2018). Desafortunadamente, este recurso está siendo contaminado con elementos fisicoquímicos, químicos, metales y bacteriólogos que son medidos mediante equipo especializado que implican sensores de pH, imanes, pruebas químicas reactivas ante la presencia de contaminantes específicos u otros no identificados. La contaminación se puede medir de forma específica con componentes anteriormente mencionados, pero también se pueden medir con los sentidos, es decir, se usa el sentido de la vista para visualizar la claridad que presenta el agua. Cuando este vital líquido es más claro y cristalino se considera como agua “limpia”, sin embargo, si se ve turbia o con tonalidades oscuras o marrones, pasará a ser juzgada como agua “contaminada”. Cuando presenta un olor extraño o no muy agradable es también un indicativo que el agua se encuentra contaminada. Finalmente, si al ser consumida se descubre un sabor extraño, podría ser un indicativo muy evidente de la contaminación del agua.

2.2 Estado del arte

La contaminación del agua, en sus diferentes cuerpos, es una gran preocupación para diversos países del mundo. Por ello, se crearon normas con las cuales considerar el agua apta para el consumo o uso humano junto con sus distintos métodos de medición a los contaminantes. Cabe mencionar que los parámetros de calidad del agua de un país no pueden ser aplicados a los demás países, ni tampoco sus métodos para procesar la información ya que son contextos diferentes. A pesar de ello, los métodos y procesos de analítica de datos que se utilizan son de utilidad y sirven de base para aplicarlo al análisis de la contaminación del agua en nuestro país, México. En la siguiente tabla se presentan algunos trabajos usados para el tratamiento de datos para la búsqueda de patrones de los contaminantes presentes en distintos cuerpos de agua **(***¡Error! No se encuentra el origen de la referencia.***)**. A pesar de esto, la principal similitud entre trabajos es en la métrica con la cual se estima que el agua está contaminada o no. Esta métrica es la saturación de

oxígeno disuelto debido a que algunos contaminantes interfieren con la composición del agua reflejándose en todas las métricas relacionadas con el oxígeno.

Cita	Objetivos	Técnica de análisis de datos	Datos utilizados	Resultados obtenidos
(Jatnika, 2021)	Analizar los datos de distintos cuerpos de agua buscando los niveles de contaminación de cada uno	Método K-Means y clustering de los datos	Los datos fueron recabados de 34 provincias de Indonesia donde previamente se había registrado contaminación en el agua con anterioridad	De los 34 lugares resultó que 4 de ellos tenían los valores más altos registrados y los 30 restantes tenían los valores más bajos registrados y usando el método DBI se evaluó los algoritmos de agrupamiento proyectando como “buena calidad” al agua
(Sameer & Ashok K., 2021)	Reconocimiento de la calidad del agua en el río Yamuna	Clustering de conglomerados jerárquicos.	Con una longitud de 1,376 kilómetros y 3 metros de profundidad se dificulta el abarcar todo del rio por ello se dividió en 19 puntos locaciones y cuatro puntos de muestreo repartidos por el río	Al considerar la parte temporal de los registros se identificó que la temporada veraniega e invernal son donde más altos son los parámetros de contaminación, asimismo que en algunos sitios hay más contaminación que en otros
(Haixia, Jianhong, & Xiangwei, 2021)	Búsqueda de características sobre la contaminación en el agua marina.	Minería de datos, monitoreo contante plasmado en diversos gráficos y método K-means	Los datos son obtenidos contantemente con equipo especializado para el monitoreo	Dado que sólo se analizaron dos tipos de contaminantes en el agua no hay resultados concretos si bien eso no sería considerado como un resultado favorable se obtuvo un procedimiento de análisis de contaminantes en agua marina.

Cita	Objetivos	Técnica de análisis de datos	Datos utilizados	Resultados obtenidos
(Aslam, y otros, 2022)	Adecuar repositorios de datos con máquinas de aprendizaje automático para la gestión de la calidad del agua.	Algoritmos de aprendizaje automático	La zona de estudio es un cuerpo de agua que se encuentra ubicada en la frontera entre Pakistán y China	El cruce entre los algoritmos de aprendizaje automático y minería de datos dejaron con un la tasa de éxito mayor al 90 por ciento con los diversos contaminantes presentes en el cuerpo de agua
(Theyazn H. H, Mohammed, Hasan, & Mashael, 2020)	Predecir la calidad del agua con algoritmos de inteligencia artificial	Modelo de redes neuronales	Datos históricos de diferentes estados de India entre los años 2005 y 2014 con una cantidad de 1679	Culminó con una herramienta preparada para la predicción en la calidad del agua, sin embargo no está finalizada por lo que requiere mayor entrenamiento.
(Mosle Hmoud & Fawaz Waselallah, 2021)	Modelar y predecir la calidad del agua usando inteligencia artificial Siendo una continuación de un estudio anterior	Modelo de redes neuronales, tablas para estimar la confusión del dato y métodos de evaluación	Datos históricos de 662 fuentes de agua como ríos y lagos en la India	Pulido en los procesos de una herramienta de predicción en la calidad del agua, agregando funcionalidad de indicar el tipo de datos que aún no han sido comprendidos y más métodos de evaluación

Tabla 1. Resumen del estado del arte

3. Marco Teórico

En esta sección se presentan los fundamentos teóricos que son de utilidad para una fácil comprensión del trabajo realizado en el siguiente capítulo. Además, se presentan los conceptos sobre las metodologías empleadas y en que consiste cada fase.

3.1 Modelo CRISP-DM

Ofrece un gran apoyo a los proyectos de una forma generalizada, mostrando el flujo de la información desde el ingreso de la información hasta el punto en que es transformada en conocimiento para culminar en una etapa de acción la cual consiste en aplicar el conocimiento buscando obtener beneficios (IBM, Guía de CRISP-DM, 2021).

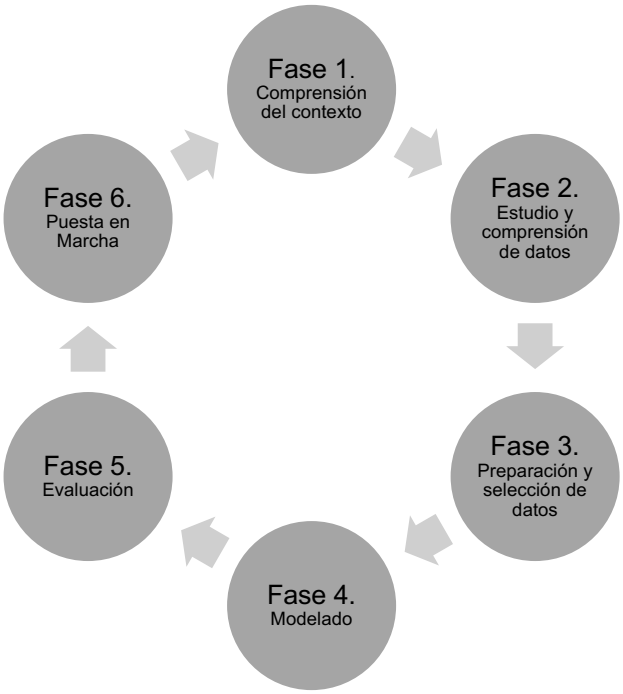

Imagen 1. *Diagrama del modelo CRISP-MD*

Comprensión de contexto

Frecuentemente este paso es hecho cuando el proyecto es nuevo, no se tiene ninguna base con la cual trabajar, cuando es un ámbito nuevo, en la integración de una nueva empresa a un mercado al que no conocían previamente, entre otros. Al ser el primer escalón es necesario llevarlo a cabo de forma correcta. Para ello, se tienen las siguientes recomendaciones: ser meticuloso y dar el tiempo que sea necesario para que este primer paso se haga en su totalidad sin dejar nada de lado. La desventaja de no hacerlo correctamente y no darle el tiempo necesario puede llevar al proyecto a una pérdida total tanto de dinero como de tiempo. Para reducir la probabilidad de fracaso se pueden considerar los siguientes elementos:

1. *Autoconocimiento*: esto puede ser algo extenso de realizar ya que se debe enlistar los recursos con los que se tiene a disposición, las capacidades de expansión que se tienen en caso de requerirlo, además de tener en cuenta el rendimiento que es fácil estimar respondiendo a las siguientes preguntas: ¿Dispongo de los recursos necesarios para llevar a cabo el proyecto?, ¿Dispongo del personal suficiente?, ¿Cuáles son mis principales factores de riesgo?, ¿tengo un plan de contingencia y en qué consiste?, ¿cuento con el rendimiento necesario para la minería de datos?.
2. *Objetivos para el proyecto*: plantear objetivos de forma prematura servirán en etapas más avanzadas del proyecto ya que podrán ser acoplados a los objetivos finales, planes, metas establecidas en los tres niveles temporales los cuales son: corto, mediano y largo plazo o simplemente descartarlos por no tener un valor en las etapas porvenir.

Estudio y comprensión de datos

De nueva cuenta la comprensión de la información de los cuales se disponen es de extrema importancia ya que tiene relación con los tres tipos de datos existentes para esta etapa. El primer tipo son los datos existentes, en caso de que no sean satisfactorios siempre se contempla la opción de adquirir nuevos datos los cuales pueden ser obtenidos de distintas maneras que pueden ser:

- Encuestas web: se aconseja este tipo de datos sean de opción múltiple, tales como números (encuestas de calidad) en vez de respuestas abiertas como

texto. Esto es debido a que el primer caso es cuantificable y totalmente medible en comparación del texto que queda a interpretación del lector. Este proceso requiere frecuentemente de mucho tiempo de recolección.

- Monitoreo de variables: para estos datos la mayor inversión es económica ya que deberán tomarse el tiempo en la fabricación o compra del equipo necesario con la capacidad de hacer la toma de medidas, configuración y almacenamiento de los datos.
- Fuentes externas o terceros: este tercer tipo de datos es una combinación de las dos anteriores ya que se pueden buscar datos ya recabados por terceras personas o instituciones los cuales son compartidos para uso libre.

Sin importar el origen de los datos, pueden siempre ser utilizados como complemento para los datos originales, siempre y cuando estén en el mismo rango, tipo de datos, entre otras consideraciones.

Para comprender los datos es necesario conocer de que naturaleza son. La primera prueba puede ser visual, sin embargo, no es tan exacta, ya que en caso de ser registros que rondan los miles puede ser muy cansado para el encargado de esta tarea, no hay conteos ya que sólo es una prueba visual, ni ningún otro tipo de medidas en esta prueba salvo una excepción que sería identificando “los que parecen más vacíos”, aunque no puede dar una idea mínima del estado o la apariencia de estos y no es confiable. La segunda prueba, presente durante toda la etapa del proyecto, puede sufrir cambios o actualizaciones en cualquier momento, se trata de un diccionario de datos, conteniendo información como: tipo de dato, número en que se presenta dicho dato y significado de códigos, los datos son identificados con códigos que son en esencia abreviaciones de los datos, ahorrando espacio y memoria. La tercera prueba es hacer una tabla de pesos de cada dato, el uso de una herramienta de visualización de datos especializada es de gran apoyo, dichas tablas posicionan a los datos según los que mayor número de datos contienen con respecto a los datos faltantes y la última prueba es la escritura uniforme de los datos, explorando los datos intentando hallar algún carácter alfanumérico o de puntuación que no coincida con la escritura habitual de los datos.

Preparación y selección de datos

Durante esta etapa se requiere una herramienta de análisis y minería de datos ya que estas tienen múltiples funcionalidades asimismo que serán de gran ayuda para las dos etapas posteriores a esta.

Los programas para el análisis y minería de datos tienen los llamados “filtros”. Estos operadores son la primera tarea para la preparación o limpieza de los datos. La configuración de filtros puede variar entre los diferentes programas y por las necesidades del proyecto. Algunas de las funciones que se siguen son: selección de atributos en específico, en caso de la simbología alfabética, de puntuación o espacios en blanco que no concuerden con el contenido de los datos y de igual forma los registros en blanco o “elementos vacíos”; selección de un rango en específico, los rangos son aplicados para dos situaciones, los elementos de mayor valor o los valores de menor valor.

Continuando con la selección de datos, se eligen los datos por columnas y por filas. un buen punto de inicio es tomar los valores que presentaron el menor número de valores perdidos. La segunda tarea consiste en dar pesos a los datos según los objetivos del proyecto, el hecho de tener exactamente los mismos datos en dos proyectos distintos hace que los pesos de los parámetros sean divergentes, respondiendo a la pregunta: ¿qué quiero conseguir con estos datos? y teniendo en cuenta que es mejor dividir objetivos grandes en objetivos más pequeños y específicos para obtener mejores resultados. Para la elección de los parámetros con mayor relevancia es prudente crear un nuevo repositorio de datos con las correcciones hechas durante esta etapa. Por último, la incorporación de nuevos datos a esta depuración amplia nuestra probabilidad de éxito y aumenta nuestro conocimiento.

Modelado

En el transcurso de esta etapa aún se pueden realizar cambios en las fases anteriores, es decir, las fases de preparación y selección de datos, y la de comprensión de datos y del contexto. Las herramientas de análisis y minería de datos presentan ayuda a comprender el comportamiento de los datos mediante resultados visuales en forma de graficas relacionadas con los objetivos del proyecto. A continuación, se presentan algunos ejemplos:

- Histogramas: el uso de barras en este tipo de grafico es habitual y dependiendo de la altura de las barras corresponde la frecuencia durante ese periodo de tiempo. Sirven

para mostrar una visualización de las variables o conjuntos de datos orientada en los patrones o comportamientos de estos a lo largo del tiempo como un historial. Por otra parte, hay histogramas que muestran dos variables simultáneamente, con la restricción de que estén relacionadas.

- Graficas de caja: los elementos llamados “cajas” en estas graficas se posicionan sobre los datos registrados. Estos datos permiten ver su “preferencia” como un comportamiento que se debe tomar en consideración. Sin embargo, hay un elemento en estas graficas que es algo imperceptible y es una cruz de un tamaño bastante pequeño la cual apunta a valores asimétricos o que están fuera de un conjunto o de un comportamiento.
- Graficas de dispersión: los valores son ubicados en forma de puntos dentro de la gráfica y muestran el comportamiento de las variables, la mayor aportación es el ser combinado con otros tipos de graficas resultando con una presentación visual más extendida.
- Grafica de pastel: el aporte es la comparación entre las distintas variables que fueron seleccionadas para que las contenga y el segundo es el dimensionar la magnitud de cada una de las variables.

Todos estos modelos son determinados según las necesidades del proyecto. Una forma de seleccionar la mejor grafica consiste en hacer las siguientes preguntas, ¿cuál es el objetivo con el modelado? y ¿cuál o cuáles de todas las gráficas me sirve para mostrar la mayor cantidad de información sin ser confusa? Finalmente, toda esta etapa de modelado debe ir acompañada con un reporte detallado con toda la información sobre el proceso realizado, información estadística y descripción de las gráficas.

Evaluación

Como penúltima fase se encuentra la evaluación, la forma de evaluar es de dos formas, la primera es estimando el rendimiento del proceso de minería de datos y la segunda es valorar si se satisfacen los resultados obtenidos con los objetivos planteados al inicio del proyecto. Con el propósito de cumplir lo anteriormente mencionado, se deberán responder las siguientes preguntas: ¿Los modelos responden o se acercan a los objetivos planteados?, ¿qué conocimientos nuevos se obtuvieron?, ¿qué no se conoce aún?, ¿se pueden plantear nuevos objetivos con el nuevo conocimiento obtenido?, ¿se alcanzaron las metas esperadas?, ¿qué sorpresas hubo durante esta etapa?, ¿el plan de contingencia deber ser modificado?, entre otras.

Puesta en marcha

Después de estudiar el sector al cual está destinado el proyecto se deben analizar sus resultados, pero antes es imprescindible crear tareas para cada uno de los objetivos planteados. Lo óptimo es desplegarlo a corto, mediano y largo plazo. Hacer monitoreo de la eficacia de cada una de las tareas permite un mayor control sobre estas, reacción a tiempo en caso de una contingencia y reducción de los daños de la contingencia.

3.2 Algoritmos para clasificación de minería de datos

En esta sección se presentan los algoritmos de minería de datos utilizados en este trabajo, la selección corresponde a la investigación en el estado del arte y en los algoritmos disponibles según las herramientas empleadas.

3.2.1 Arboles de decisión

Es un algoritmo aplicado al aprendizaje automático supervisado condicional usado tanto en actividades de clasificación como de regresión teniendo una estructura de jerarquía de forma descendente. El punto más alto es llamado *raíz* o *nodo*, siguiendo por los nodos internos y culminando en nodos *hoja*, estos últimos ubicados en la base del algoritmo. Todos los nodos estando conectados por líneas llamadas *ramas* como se muestra en el siguiente diagrama:

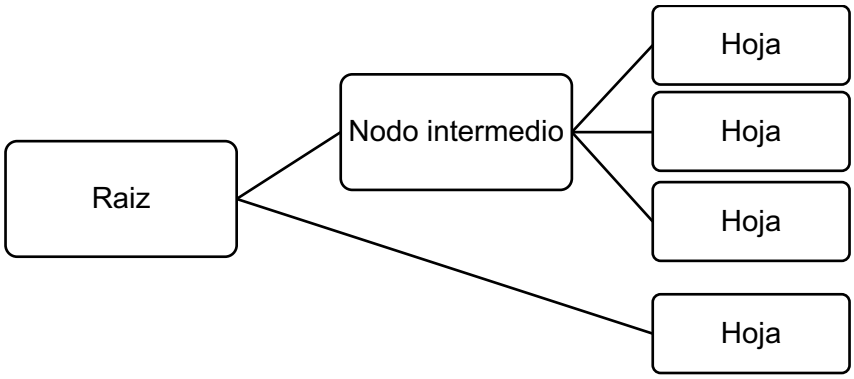

***Imagen 2**. Diagrama de flujo en Árbol de decisión*

La entrada de la información da inicio en el nodo raíz donde se aplican condicionales buscando las características que nos serán de utilidad para clasificar los datos entrantes. Dependiendo de las necesidades que se busquen cumplir con el árbol de decisión, se descartará o se aplicará un segundo procesamiento a los datos rechazados. Los nodos

intermedios tienen la tarea de clasificar a otros y darles paso a los nodos hojas que reflejan el resultado final del procesamiento de este algoritmo. La esencia de este algoritmo es tomar un elemento complejo y fragmentarlo para conocerlo más a detalle, pudiéndose comparar con la frase de “divide y vencerás”, como se puede ver reflejado en *Imagen 2*.

El algoritmo de árbol de decisión presenta las siguientes ventajas:

- **Alta flexibilidad:** su flexibilidad es en la estructuración que se le puede dar a este algoritmo y que inconvenientes como una gran similitud entre dos variables es resuelta por el algoritmo simplemente seleccionando una de las dos variables para el análisis evitando de esta forma cuellos de botella.
- **Fácil interpretación:** por la naturaleza y simplicidad en la estructura del algoritmo se puede ver rápidamente los parámetros o métricas usadas para la clasificación de las variables.
- **Poca o nula preparación de los datos de entrada:** permite la entrada de valores decimales y enteros, también descripciones físicas tales como color, material, textura, entre otros, sin olvidar que también puede usar valores omitidos, algo que no es capaz de aceptar el algoritmo de Naive Bayes (IBM, ¿Qué es un árbol de decisión?, 2021).

Por el contrario, tiene algunas desventajas, las cuales son las siguientes:

- **Estimadores de alta varianza:** una pequeña variación en los datos produce cambios drásticos en todo el árbol de decisión, un método para aminorar el impacto de este problema es promediar las estimaciones, pero esta solución es muy limitada.
- **Propenso a sobreajustes:** tratando de que las entradas sean procesadas correctamente se crean nodo los cuales no aportan a la clasificación ralentizando el procesamiento de estos, una forma de corregir esta problemática es descartando nodos que no aporten nada o en otras palabras “podar el árbol de decisión”.

3.2.2 K-NN

También conocido como “algoritmo de k vecinos más cercanos” consiste en clasificador de aprendizaje supervisado que mide la proximidad entre un centro previamente establecido y los valores más cercanos. Los resultados se obtienen en base al porcentaje de los valores más cercanos al centro y son almacenados en la memoria, por ende, se requiere de un

único proceso de entrenamiento. Por ejemplo, si se tienen dos valores a analizar con cinco registros cada uno, se medirá en porcentajes de 50%. En caso de aumentar a cuatro valores con 5 registros entonces se reduce a 25%. Se recomienda que el análisis se haga usando números impares, esto con la finalidad de buscar entre un conjunto un valor que sobresalga. La forma de analizar los datos es mediante la determinación de distancia en las métricas, para alcanzar a identificar las distancias se usan métodos como: distancia euclidiana, Distancia de Manhattan, Distancia Minkowski y Distancia de Hamming (IBM, ¿Qué es Knn?, 2021)

El algoritmo KNN también presenta algunas ventajas como:

- **Fácil implementación:** tiene un alto nivel de precisión y simplicidad, son uno de los primeros algoritmos que aprende un científico de datos.
- **Fácil de adaptar:** el algoritmo en su fase de entrenamiento acepta nuevos parámetros y ajusta lo nuevo con lo que ya se encontraba con anterioridad.

Las desventajas del algoritmo son:

- **No escala adecuadamente:** se origina con respecto a lo anterior mencionado, se trata de un algoritmo de memoria, al incorporar más valores al análisis requiere de mayor memoria y potencia de procesamiento de cómputo, aumentando los costos y tiempo de implementación de mejoras, aunque hay estructuras como Ball-Tree que trata de reducir el problema no lo hace por completo.
- **Propenso a sobreajuste:** a menor cantidad de datos ingresados el ajuste de nuevos parámetros será más fácil de procesar, por el contrario, agregar mayor cantidad de parámetros al algoritmo hace que se creen métricas que atentan a la velocidad de procesamiento de los datos o creando métricas nada útiles para el procesamiento de los datos.
- **Problema de dimensionalidad:** los datos ingresados no deben tener un cierto número de almacenamiento en el archivo o también conocido como fenómeno de pico.

3.2.3 Naive Bayes

Se trata de uno de los tipos de algoritmos de clasificación aplicado al aprendizaje automático que incluye una fase de entrenamiento basado en la técnica de clasificación basado en el teorema de Bayes. Este teorema busca encontrar la probabilidad de que ocurra un suceso dado un segundo suceso que afecte al primero o que por el contrario no esté relacionada usando el método de máxima verosimilitud que mide la probabilidad de que ocurra o se cumpla la hipótesis planteada o que el parámetro estimador que hemos planteado sea el correcto. Debido a su simplicidad resulta en una manera fácil de construir modelos que demuestren algún comportamiento ya que se puede aplicar este algoritmo a conjuntos de datos pequeños, todo este algoritmo funciona mediante la siguiente ecuación:

$$P(A|R) = \frac{\mathrm{P(R|A)P(A)}}{\mathrm{P(R)}}$$

Ecuación 1. *Naive Bayes*

Donde P(A) es la probabilidad A, P(R|A) es la probabilidad de R con respecto A, P(R) es la probabilidad de R y, por último, P(A|R) es la probabilidad siguiente que se suceda A con respecto a R.

Un ejemplo de este algoritmo es planteado considerando una piña la cual tiene un promedio de 30 centímetros de alto con 15 centímetros de grosor. El color y textura de una piña se puede comparar con todo el conjunto de datos buscando similitudes entre las alturas, anchos, colores y texturas de los otros. Cuando los valores de las clases (características de la piña) no coinciden u ocurren con los datos de entrenamiento, entonces la estimación de similitud baja a cero. Esto es un inconveniente ya que, observando la ecuación, cualquier dato siguiente al ser multiplicado dará cero causando un error. Por ello, se puede aplicar un *pseudocontador* que tenga las mismas características pero que no sea del todo igual al parámetro original. Transcribiendo y reflejando las diferentes características del ejemplo de la piña resultará en la siguiente ecuación:

$$P = \frac{\mathrm{P(altura|piña)P(grosor|piña)P(color|piña)P(textura|piña)P(piña}}{\mathrm{P(piña)}}$$

Ecuación 2. *Reemplazo de valores para Naive Bayes*

Donde R es reemplazada por *piña* (el parámetro principal) y A equivale a las características (Roman, 2019)

3.3 Hidrología de Puebla

Dentro del territorio del estado de Puebla se tiene la presencia del segundo volcán más grande del país, el Popocatépetl o también llamado Don Goyo con una altura de 5,426 metros.

El estado de Puebla se ubica en la quinta posición en el censo poblacional realizado por el Instituto Nacional de Estadística y Geografía (INEGI), con una extensión territorial de 34,251 kilómetros cuadrados colindando al norte con Tlaxcala e Hidalgo, al noreste con Veracruz, al sur con Oaxaca, al suroeste con Guerrero y al oeste con dos estados Morelos y Estado de México. A pesar de que en el estado no se cuenta playas, cuenta con una excelente oportunidad en turismo en varios ámbitos como el turismo religioso, gastronómico, ecoturismo, entre otros. Esto se debe a sus 9 pueblos mágicos y 3 lugares que fueron declarados por la Unesco como Patrimonio de la Humanidad, sin olvidar que fue protagonista en algunos acontecimientos históricos para el país (UNESCO, Pueblos Mágicos de México, 2020) . El estado de Puebla tiene una hidrología interesante como se ve representado en la *Imagen 3*.

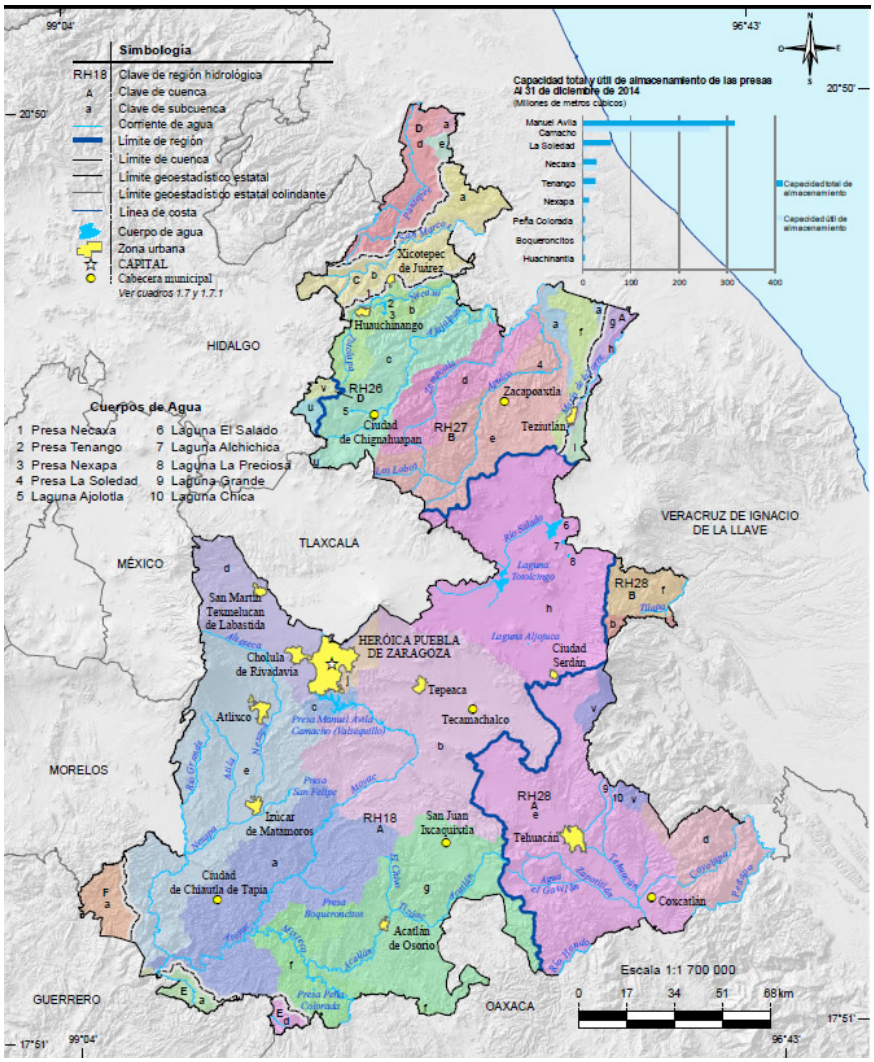

***Imagen 3**. Mapa de la Red Hidrográfica Digital de Puebla, Escala 1:250 000 (Obtenido de: INEGI - CONAGUA, 2007)*

El territorio del estado de puebla está dividido en tres regiones las cuales son: Región Balsas, Región Tuxpan-Nautla y Región Papaloapan, las cuales fueron ordenadas de mayor a menor según su tamaño. Recordando que el territorio poblano tiene una superficie de 34 252km^2 se puede representar en porcentaje el terreno que abarca cada una de las regiones y considerando el orden anteriormente mencionado se ordenan por porcentajes de la siguiente manera: 59.15%, 23.76% y 16.88%, el porcentaje restante pertenece a la región de Pánuco la cual es demasiado pequeña para ser considerada y tiene poco impacto en comparación de las otras regiones (p. 13). Una mejor forma de visualizar las regiones hidrológicas del estado de puebla es en la *Imagen 4.* Examinando más a profundidad la hidrología de Puebla es considerando los cuerpos de agua presentes en este estado siendo un total de: 22 ríos, 7 presas y 7 lagunas distribuidas a lo largo del estado, algunos cuerpos

de agua siendo desembocaduras de otros o sirviendo de conexión, los ríos son: Atoyac, Nexapa, Pantepec, San Marcos, Necaxa, Ajajalpa, Apulco, Salado, Zempoala, Tilapa, Mixteco, Acatlán, Zapoteco, Coyolapa, Alseseca, Atila, La Huertilla, Axuxuca, El Chivo, Tizáaz, Texcapa y Hondo. Las presas son: Manuel Ávila Camacho, Necaxa, Tenango, Nexapa, La Soledad, Boqueroncito y San Felipe. Las lagunas son: El Salado, Totolcingo, Grande, Chica, Ajolotla, Alchichica y La Preciosa, como se observa en la *Imagen 4* (González Villareal, 1986).

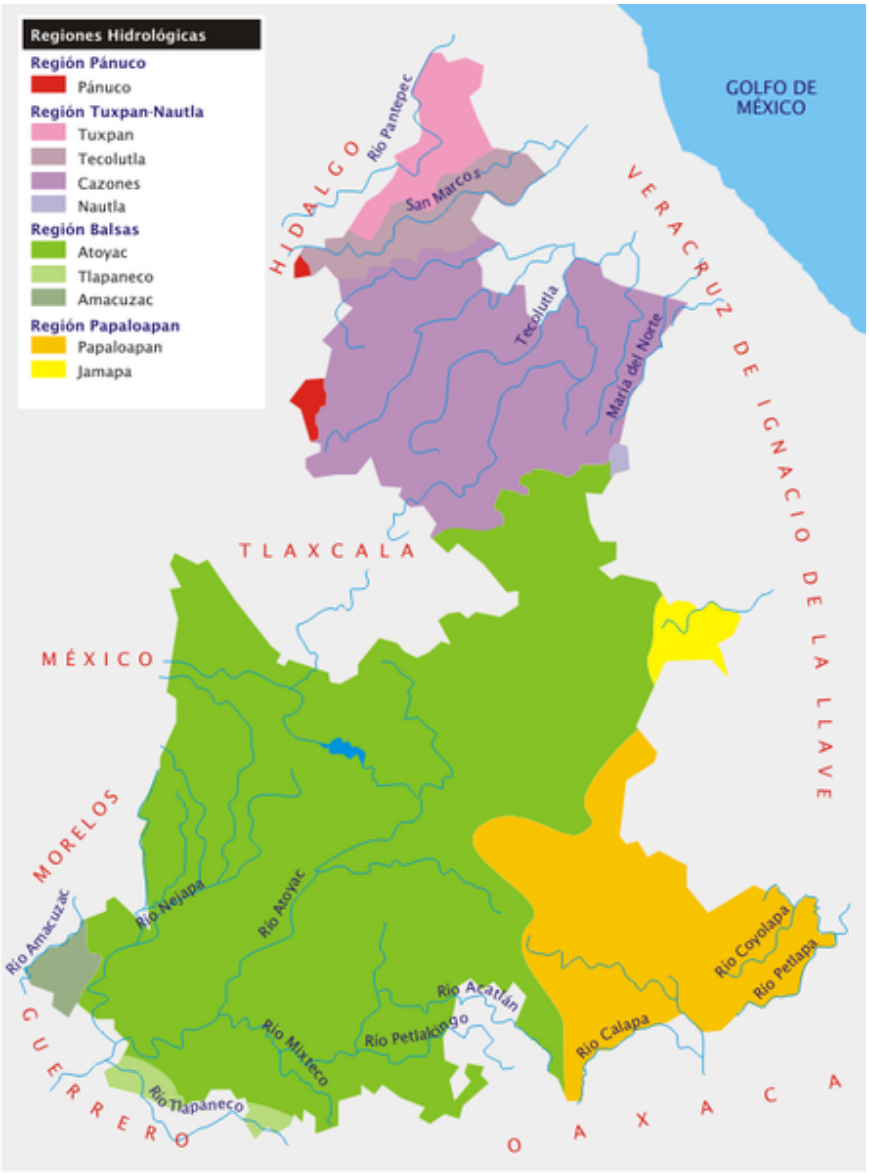

Imagen 4. *Hidrología del estado de Puebla por zonas.(Obtenida de: INEGI 2006* (Tulcingo, 2018 - 2021) *)*

3.3.1 Tipos de cuerpos de agua

Conocer el tipo de cuerpo de agua proporciona información sobre cómo se pueden comportar los contaminantes. La cantidad de zonas que logran abarcar dicho contaminante, situaciones hipotéticas donde algún contaminante se encontró con otro puede ocasionar una contaminación cruzada.

Tipo de cuerpo de agua	Características	Imagen
Lótico	También llamadas aguas loticas o corrientes, es un tipo de masa de agua que se mueven siempre en una misma dirección como ríos, manantiales, riachuelos y arroyos. (demográfico, 2020)	***Imagen 5.*** *Montañas (Obtenida de Pixabay, bajo licencia CC)*
Humedal	Son cuerpos de agua que retienen o almacenan agua, si el agua se encuentra en un terreno con los sedimentos adecuados estos se pueden filtrar directamente a los acuíferos. La diferencia entre un cuerpo de agua lentico y un humedal es en la poca cantidad de agua que se presenta en los humedales a comparación de un cuerpo de agua lentico. (Biodiversidad, 2023)	***Imagen 6.*** *Pantano (Obtenida de: Pixabay, bajo licencia CC)*
Subterráneo	Como su nombre lo indica, que se encuentra por debajo de la superficie terrestre, ubicadas bajo tierra, grietas, huecos, bajo arena, rocas o sedimentos. Estos espacios donde se encuentra este cuerpo de agua son llamados acuíferos. Es llamado como “el agua que no vemos”. (Hermann & Prunes, 2022)	***Imagen 7****. Cueva (Obtenida de: Pixabay, bajo licencia CC)*
Lentico	Cuerpos de agua de interior, quietas o estancadas tales como los lagos, lagunas, charcas, humedales o pantanos. (demográfico, 2020)	***Imagen 8.*** *Lago (Obtenida de: Pixabay, bajo licencia CC)*
Humedal (Lentico / lotico)	Es una combinación entre un cuerpo lentico o lotico, pero con la poca cantidad de agua en él, por ejemplo, poca profundidad, pero una gran extensión o poca profundidad, pero con un flujo de agua en él. (demográfico, 2020)	***Imagen 9.*** *Páramo Paisaje (Obtenida de: Pixabay, bajo licencia CC)*

*Tabla 2****.*** *Tipos de cuerpos de agua y su descripción*

3.3.2 Contaminantes más frecuentes

Teniendo como base los registros de la CONAGUA en un periodo de tiempo de 8 años del 2012 al 2021 y aplicando un pre-procesamiento para el conocimiento de los datos, se encuentran los más frecuentes. Considerando la cantidad de registros que contiene cada uno de los contaminantes y para identificarlos se usaron programas especializados de minería de datos. Dichos programas muestran elementos más frecuentes de una forma rápida, practica e inmediata. Para una mayor claridad de los contaminantes se incluye información sobre ellos, facilitando la comprensión de estos. Los tipos de contaminantes para tener un aproximado del contenido del agua y además dando entrada para el posterior análisis de la información. A continuación, se muestran los contaminantes más frecuentes (***¡Error! No se encuentra el origen de la referencia.***).

Nombre del contaminante	Descripción	Daño al cuerpo humano	Tipo de contaminante	Clave
Coliformes Fecales	Son grupo de organismos coliflores, ocasionalmente originados por materia fecal, pueden ser originados por agua potable con ingredientes añadidos, desechos industriales, material vegetal y suelo en descomposición	Disentería: inflamación de los intestinos, se propaga por alimentos contaminados o agua sucia, produciendo diarrea grave que incluye mucosa o sangre en las heces	Bacteriana	COLI_FEC
Sólidos disueltos totales	De las siglas SDT o TDS es una métrica para medir la concentración de minerales, sale, cloruros, metales y muchos otros contaminantes disueltos en el agua.	Dependiendo del mineral excedente que contenga el agua es un daño u otro	Mineral	SDT
Nitrógeno de Nitritos	El nitrito es una forma resultante de la oxidación biológica del nitrógeno amoniacal.	El nitrito reaccionan con la hemoglobina reduciendo la capacidad para transportar oxigeno de este último.	Bacteriana / mineral	N_NO2
Nitrógeno Amoniacal	Tipo de nitrógeno en forma de amoniaco que también es llamado nitrógeno amoniacal, su comportamiento es dependiente del pH, a pH ácido permanece disuelto en el agua mientras con pH alcalino se transforma en gas de amoniaco transportándose por medio del aire	Quemaduras en la piel, ojos, boca y pulmones.	Químico	N_NH3
Nitrógeno de Nitratos	La forma más oxidada del nitrógeno siendo muy poco tóxica, teniendo el mínimo impacto en el ecosistema y la salud. Habiendo dos posibilidades, la primera que sea asimilado por las plantas o evaporando el nitrógeno.	En cantidades excesivas produce la enfermedad de “los bebes/niños/pañal azules” siendo un problema metabólico con dificultades para metabolizar el aminoácido triptófano	Químico	N_NO3

		dando paso a una coloración azulada.		
Fosforo Total	Es la suma de todas las formas de fósforo existentes como: ortofosfato o fosfatos, fosfatos condensados y fosforo orgánico.	En altas concentraciones y exponiéndose en cortos intervalos tiene consecuencias en el hígado y los riñones.	Orgánico	P_TOT
Nitrógeno Orgánico	Se encuentra en las moléculas orgánicas como las proteínas, no tiene algún efecto característico es esencial para procesos microbianos.	Daños a corto plazo como quemaduras, espasmos, dilatación de los tejidos en garganta y vías respiratorias superiores.	Atmosférico	N_ORG
Nitrógeno Kjeldahl	Procedimiento para medir los niveles de nitrógeno orgánico en muestras orgánicas e inorgánicas. Conteniendo Oxido de mercurio, selenio o una mezcla de sulfato de cobre y selenio, y dióxido de titanio	Produce síntomas que si no se atienden de inmediato producen daño irreversible, sulfato de cobre causa daños en el hígado y riñón, el dióxido de titanio irrita las vías respiratorias. Al contener estas sustancias nocivas para el cuerpo humano es fácil inferir que causará mucho daño al cuerpo humano	Químico	N_TOTK
Arsénico Total	Forma de medir los diversos tipos de arsénico que hay, los cuales son: orgánico, inorgánico y gas. El arsénico es un mineral encontrado en los sedimentos de los cuerpos de agua.	En enormes cantidades puede ser cancerígeno y causar lesiones cutáneas	Orgánico	AS_TOT
Plomo Total	El plomo es un metal altamente tóxico, este es una métrica para medir los distintos tipos de plomo por ejemplo: sales del plomo y compuestos orgánicos del plomo.	El plomo al tener contacto con la sangre ocasiona daño cerebral, riñones y causar interrupción en la producción de glóbulos rojos.	Metálico	PB_TOT

Tabla 3. Contaminantes más frecuentes

3.3.3 Contaminantes menos frecuentes

Los contaminantes más frecuentes son aquellos provocados por fuentes naturales causados por eventos meteorológicos, orgánicos u otros motivos, químicos y que son provocados por el humano, ya sea intencionalmente como desechos industriales o error humano, accidentes que no tuvieron la intención de contaminar. Sin embargo, en la *Tabla 4* se presentan los contaminantes con menor presencia en el estado de Puebla.

Nombre del contaminante	Descripción	Daño al cuerpo humano	Tipo de contaminante	Clave
Antimonio Total	Elemento químico semimetálico que tiene cuatro categorías, representadas en colores las cuales son: blanco azulado, negro y amarillo, usada en aleaciones metálicas, ofrecer resistencia a los materiales contra el fuego, pinturas, esmaltes, elementos cerámicos, vulcanización del caucho y fuegos artificiales. Puede encontrarse en el agua por causas naturales e industriales.	Exposición a altos niveles y tiempo prolongado causa erupciones con picazón y enrojecimiento en la dermis, garganta y pulmones, tos, respiración con un silbido o falta de aire, dolor de cabeza, mero, náuseas, vómitos, dolor abdominal y falta de sueño. Puede matar fácilmente a fauna de un tamaño pequeño.	Mineral	SB_TOT
Bario Soluble	Metal que es encontrado en el medioambiente, normalmente encontrado con otros elementos químicos, sus compuestos son usados en la industria como lubricantes, elaboración de pinturas, bloques de construcción, gomas, azulejos y vidrio.	En altas cantidades puede causar parálisis o incluso la muerte, en pequeñas cantidades resulta en dificultad para respirar, presión sanguínea alta, inflamación en el cerebro, hígado, riñones y corazón.	Química	BA_SOL
Dureza Calcio	Método que busca la concentración de iones de calcio y magnesio en el agua	Sin consecuencias negativas para la salud.	Mineral	DUR_Ca
Dureza Magnesio	Suma en la concentración de magnesio en el agua	Sin consecuencias negativas para la salud	Mineral.	DUR_Mg

Tabla 4. *Contaminantes menos frecuentes*

3.4 Métodos de evaluación para medir la contaminación del agua.

El concepto de calidad del agua hace referencia al proceso de evaluación que se somete a dicho recurso con la finalidad de obtener una descripción de su composición radiológica, física, química y biológica del agua, la razón de este proceso es para obtener mejores resultados en las diversas áreas donde es empleada el agua, reduciendo los daños a los ecosistemas o a los propios seres humanos

(EUROINNOVA, 2022). La tabla 5 muestra algunos métodos existentes para medir la calidad del agua.

Nombre método	Descripción	¿Cuándo se utiliza?	Unidad de medidas
Cualitativos (LENNTECH, 2019)	Este método puede ser comprobado con los sentidos como el olfato, gusto y vista	Cuando el agua tiene una apariencia o se tiene sospechas de contaminación se puede aplicar este método en primea instancia	No tiene unidades de medida ya que es por medio de los sentidos pero se busca coloraciones inusuales, malos olores o un mal sabor
Materia suspendida o Turbiedad (Carbotecnia, 2021)	Se trata de la suma de elementos solidos disueltos y suspendidos en el agua, el procedimiento consiste en extraer una muestra, filtrarla, secarla y pesar el resultado, se aconseja dividir por tipo de residuo.	Obstrucciones en los ductos por donde pasa el agua o con el aumento en la densidad del agua	Se mide por mg/L o como lo estima la norma NOM-127: 0 -300 Excelente 300 – 600 Nivel bueno 600 – 900 Nivel aceptable 900 – 1200 Nivel no recomendable 1200 a más Inaceptable
Niveles de pH (Board, Folleto informativo pH, 2023)	Consiste en medir la presencia de iones de hidrogeno o sales que presenta el agua, las sales pueden variar desde magnesio, calcio o sodio, valores altos no representan daños a la salud pero puede ocasionar fallas en nuestros organismos	Cuando el agua afecta rápidamente a la biodiversidad la cual tiene contacto o el agua se encuentra en zonas ricas en sales minerales	Se miden por niveles y se miden de la siguiente forma: 4 – 9 mayoría de ambientes naturales 7.5 – 8.4 agua de mar 6.5 – 8.5 agua dulce
Niveles de patógenos e indicadores microbiológicos (Enriquez, 2011)	La presencia de bacterias, parásitos, virus y hongos puede ser resultado del cambio medio ambiental o debido a cusas humanas	Usualmente cuando se presentan enfermedades gastrointestinales o consumo de agua contaminada	La medida es UFC o “Unidad Formadora de Colonias” es un conteo de la cantidad de microorganismos vivos en un líquido, el agua destinada para beber debe tener cero de Ecoli y Coliformes fecales
Medición de Oxígeno Disuelto (Board, Folleto Informativo Oxígeno Disuelto, 2023)	La cantidad de oxígeno que tiene el agua a determinada temperatura, tiene dos nombres este método, concentración de OD y porcentaje de saturación	Cuando la biodiversidad de un cuerpo de agua se reduce o incremento en muertes de adultos y jóvenes	La unidad de medida es mg/L es decir los miligramos por litro

Conductividad del agua (optek, 2023)	Es la facilidad que tiene la energía eléctrica de fluir por el agua, presenta una visión clara de los iones o sales que contiene, ya que el agua es naturalmente resistente a la conducción de energía	Cuando el agua tiene una medida sospechosamente alta en sales	Se mide mediante Siemens, ejemplo con muestra a 25°: 0.05 µS/cm Agua de alta pureza 500 µS/cm Agua de grifo 1,76 mS/cm Ácido acético 10 353 mS/cm NaOH 10% Donde el símbolo µS simboliza micro Siemens y mS simboliza mili Siemens

Tabla 5. Métodos para medida de calidad del agua

La información presentada por la CONAGUA para medir la calidad del agua está basada en la medida de la presencia de oxígeno disuelto, demanda de oxígeno y demanda biológica de oxígeno. Esto se puede visualizar en el mismo sitio web donde fueron obtenidos los datos y considerando la norma NMX-AA-028-SCFI-2021 que determina el oxígeno disuelto en aguas naturales, residuales y residuales tratadas (CONAGUA, 2024).

4. Marco metodológico.

En este capítulo se presentan los pasos metodológicos a seguir para identificar patrones de la contaminación en el agua, siendo en esencia el proceso de análisis de datos basado en el modelo CRISP-DM.

4.1 Conociendo los datos

El primer paso para la comprensión de los datos es generar descripciones mediante tablas para diferentes conjuntos de datos, lapsos de tiempo, tipos de datos, entre otros. De igual forma, es importante comprender el origen de los datos que se usaron para este trabajo. Los datos fueron recopilados por la Comisión Nacional del Agua (CONAGUA) que abarcan un periodo de tiempo 9 años, es decir, desde el 2012 al 2021. La *Anexo Tabla 1* muestra un censo de los sitios donde se hizo la toma de muestras para un posterior llenado de los datos, dividiéndola en dos principales columnas que son: Nombre de sitio K y los números de casos de cada uno de los sitios. La mayoría de los registros provienen de plantas de tratamiento de agua ubicadas muy cerca de cuerpos de agua, la mayoría ubicados en el rio Atoyac. Los datos se estructuraron con los nombres de los sitios en la primera columna, se agruparon los sitios por secciones, dividiéndolos por una diagonal igual que el número de casos que fueron registrados.

Las herramientas de minería de datos aportan información sobre los datos que son ingresados para su procesamiento. Si esta información es organizada en una tabla descriptiva tendría una estructura similar a la de la *Tabla 6 que e*xpone 5 características de los contaminantes. La fila de *nombre de sitio* y *clave de sitio* se refieren a los sitios donde fueron obtenidos los datos, mostrando una diferencia de un valor a otro. Esto debido a que existen algunos sitios que son presas de aguas las cuales tienen dos puntos o compuertas que son considerados como “punto A” y “punto B” dentro del mismo sitio tratándose de dos registros diferentes. La clasificación de los cuerpos de agua permite conocer la cantidad de agua y saber las condiciones en las cuales los contaminantes se pudieran desarrollar. Se tomaron datos de 9 años y se almacenaron 2,824 registros dentro de *clave de monitoreo* con 252 tipos de contaminantes.

Parámetros	Descripción	Número de contaminantes
Clave sitio	Hay 113 claves de sitios registrados	252 contaminantes registrados
Clave monitoreo	En cada ocasión que se encontraba la presencia de algún elemento se registraba en esta columna, existen 2824	
Nombre del sitio	Hay 114 sitios registrados	
Tipo de cuerpo de agua	Hay 5 tipos de cuerpos de agua donde se obtuvieron 2824 registros	
Fecha	Parámetro que muestra cuando fue detectado algún elemento en el agua, va del 2012 al 2021, es decir, 9 años	

Tabla 6. *Parámetros usados para el registro de los contaminantes*

En la *Anexo Tabla 2* se visualizan todos los contaminantes que fueron registrados, dicha tabla contiene más información como la clave con la que fue registrado el contaminante, la unidad de medida para cada uno, el tipo de contaminante, número de registros nulos o número de registros vacíos, registros llenos o registros donde si hubo una unidad de medida para ese contaminante, la columna de porcentaje de perdida se realizó dividiendo la columna de registros vacíos entre los registros con unidades registradas. Sin embargo, a partir del contaminante en la posición 42 o el contaminante de *cianuros totales* los registros vacíos superaron aquellos que estaban llenos. A partir de ese punto comenzaron a dividirse entre el número de registros totales los cuales son 2,824 registros. Las últimas dos columnas son usadas para los rangos mínimos y máximos registrados en cada uno de los contaminantes, sin aplicar ningún tipo de filtrado en simbología incongruente debido a que esta tabla fue hecha en Excel.

En Tabla A3 del Anexo se muestran todos los contaminantes analizados, la primera columna corresponde con la clave con la que fue registrado dicho contaminante, la segunda se trata del contaminante, seguido por la unidad de medida con la que se miden dichos contaminantes y el tipo de contaminante. La columna del número de registros nulos indica el número de registros por contaminante que se encuentra vacío. La columna a continuación es la contra parte tratándose de la cantidad de registros que tiene el contaminante. El porcentaje de perdida se consiguió a través de la equivalencia de las pérdidas con respecto a los registros que estaban llenos y para finalizar las últimas dos columnas corresponden a los valores de la menor y mayor cifra registrada respectivamente.

Insertando los datos dentro de una aplicación de minería de datos y yendo al apartado de estadísticas se enlistan los diferentes parámetros con sus distintos valores. Por ello, con nuestros datos descubrimos los que encabezan la lista y son aquellos que presentan la menor cantidad de pérdidas. Los datos se pueden organizar en una tabla, donde la primera de ellas será clasificada por año sólo seleccionando los cuatro contaminantes. Con menor número de pérdidas, el programa realiza esto tomando los registros obtenidos para esos años y configurando los contaminantes promedio. Los resultados se muestran en la *Tabla 7*.

	Coliformes fecales COLI_FEC	**Solidos Disueltos Totales SDT**	**Nitrógeno de Nitritos N_NO2**	**Nitrógeno Amocioanal N_NH3**
2012	1294.53	643.50	0.04	13.41
2013	3749.18	668.14	0.15	1.36
2014	15953.70	581.17	0.18	2.67
2015	17105.60	598.58	0.08	4.86
2016	12900.17	618.61	0.17	5.93
2017	16530.91	621.12	0.21	10.07
2018	14073.71	652.01	0.22	8.76
2019	82147.16	653.20	0.19	9.41
2020	13857.23	686.68	0.29	3.93
2021	238967.56	420.65	0.06	3.94

Tabla 7. *Contaminantes más frecuentes organizados por año*

Los contaminantes presentan un aumento en los parámetros cada cierto tiempo, donde los años con los registros más grandes son 2017, 2019 y 2021. El siguiente paso para realizar es analizar estos mismos contaminantes, buscando algún comportamiento en los meses registrados, recordando que se muestran los datos promedio de todos los meses en los años que abarca el repositorio de datos. Como complemento a un análisis exploratorio se encuentra la siguiente tabla:

	Coliformes Fecales COLI_FEC	Solidos Disueltos Totales SDT	Nitrógeno de Nitritos N_NO2	Nitrógeno Amoniacal N_NH3
Enero	11663.40	741.86	0.16	4.22
Febrero	13382.03	716.50	0.15	12.27
Marzo	13281.72	727.94	0.09	6.34
Abril	48919.48	733.65	0.17	7.47
Mayo	14763.76	650.27	0.16	6.71
Junio	14300.48	585.33	0.16	4.12
Julio	14307.71	568.44	0.19	4.08
Agosto	40472.99	550.33	0.22	5.07
Septiembre	13091.63	495.88	0.24	2.70
Octubre	11494.52	531.56	0.13	2.57
Noviembre	3257.73	621.84	0.07	7.90
Diciembre	5991.25	695.61	0.11	15.29

Tabla 8. *Contaminantes más frecuentes registrados por mes*

Nuevamente se presentan dos comportamientos y con el aumento de los registros de los contaminantes que ocurren durante los meses de *abril* y *agosto* ocurriendo algo inusual con respecto a los años y es que en hay dos aumentos súbitos en los registros más o menos iguales en comparación con el disparo en los datos registrados en el 2021. El contenido de la *Tabla 8* tiene el mismo principio que la tabla anterior. Se organizaron todos los registros considerados en distintos meses, posteriormente se hizo un recuento siempre agrupándolos según el contaminante mostrado en cada columna para obtener los valores.

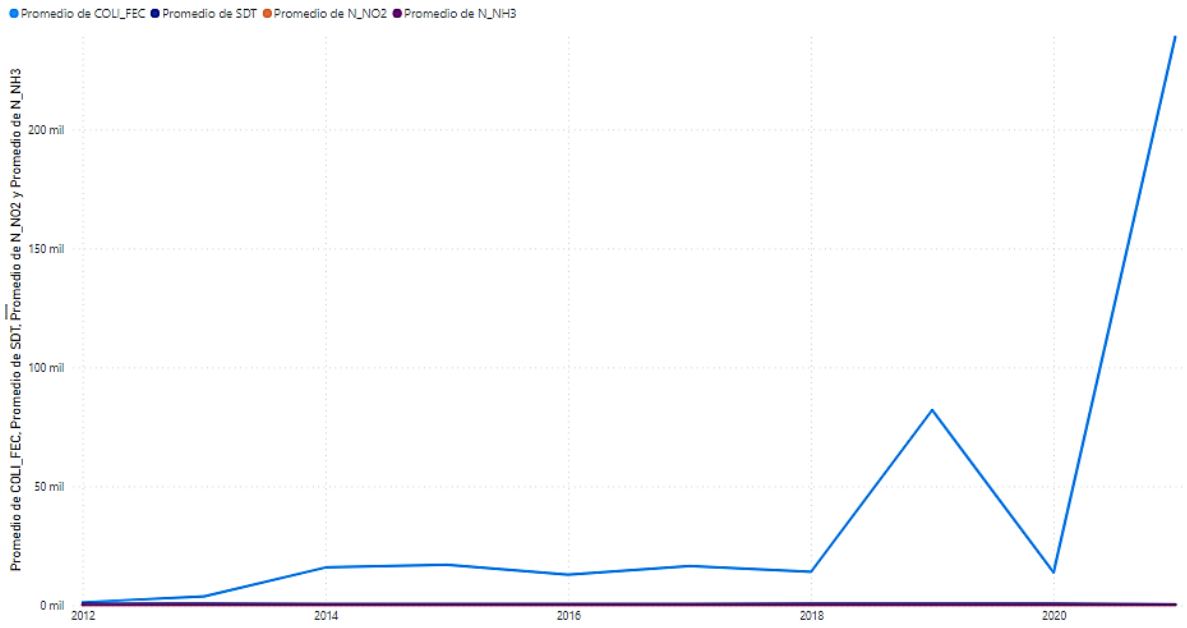

Imagen 10. *Contaminantes con mayor número de registros por año*

Al convertir la *Tabla 7* a una gráfica se observa de la *Imagen 10* , evidenciando un súbito aumento en los parámetros obtenidos de un año a otro comenzando en el 2018 al 2020 para ser despedido en el año 2021 en donde los registros en promedio fueron los más altos para el contaminante de Coliformes fecales.

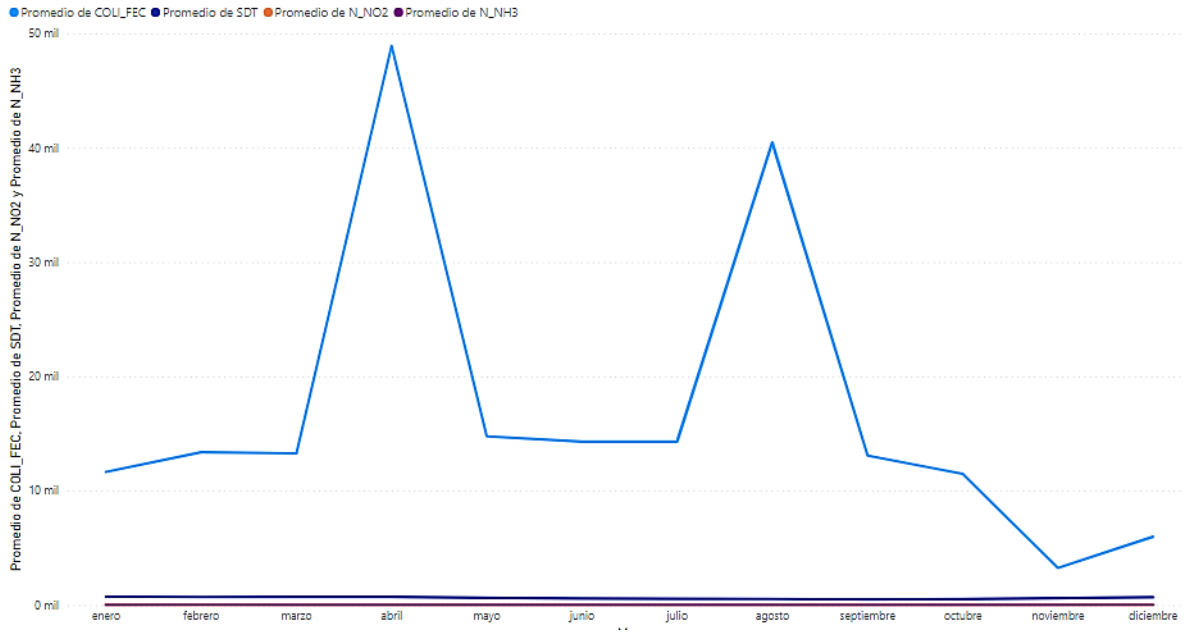

Imagen 11. *Contaminantes con menor número de perdida por meses*

Para esta gráfica *Imagen 11* tomando en cuenta la *Tabla 8* se percibe un comportamiento ligeramente diferente con respecto a la anterior *Imagen 10*. Al comenzar el año se nota un ligero aumento en los registros de Coliformes fecales para dos grandes crestas en los meses de abril y agosto y un descenso en los meses de octubre y noviembre. El segundo contaminante con mayor presencia son los *sólidos disueltos totales*.

Las medidas de los contaminantes pueden ser diferentes lo cual puede crear confusión entre los registros, sin embargo, esto es sólo un análisis exploratorio tomando en cuenta los cuatro contaminantes con el mayor número de pérdidas.

4.2 Limpieza de los datos

Efectuar una limpieza de información es obligatoria en el área de análisis de datos y para tener un mejor resultado a la hora de realizar el procesamiento y de esta forma generar *dashboards* más limpios y fáciles de comprender. La predicción de información sirve como apoyo a la toma de decisiones y termina con resultados aceptables o satisfactorios. Para ello se dividieron los datos en distintos conjuntos con diferentes características, de forma que al interpretarlos se identifican los escenarios que ocasionan la contaminación en el agua.

4.2.1 Homogenizar datos

La organización de los datos para que sean homogéneos agiliza las tareas de procesamiento y análisis. Hacerlo de forma manual no es una tarea recomendable por la alta tasa de error que genera y es mejor utilizar un programa especializado. Por esa razón se hizo un subproceso dentro de RapidMiner el cual se ve representado en la *Imagen 12*.

Imagen 12. *Pre-procesamiento de los datos con operadores en RapidMiner*

Este pre-procesamiento se ubica al inicio de cada experimento como una forma de limpieza y de reducción de errores. Este proceso se realiza de izquierda a derecha, donde los primeros operadores "*Replace*" reemplazan la simbología de "*mayor que*" y "*menor que*"

que son necesarios para reducir los datos ya que el programa no llega a interpretar estas situaciones cualitativas. En otras palabras, un registro que se presente como “<5” puede ser una infinidad de números en un rango del 0 al 5 incluyendo números decimales. Esta ambigüedad no es práctica cuando se comparan registros con un número entero o un número decimal que si puede ser cuantificable.

El tercer operador distingue entre los registros numéricos a la derecha y la información del lado izquierdo que incluyen clave de registro, sitio donde fue extraído, clave del sitio, tipo de cuerpo de agua y algún otro parámetro que se utilice para analizar información. Es importante no descartar este operador en los experimentos realizados en este trabajo ya que se buscan patrones en los contaminantes y usar la fecha es vital para encontrarlos durante el paso del tiempo.

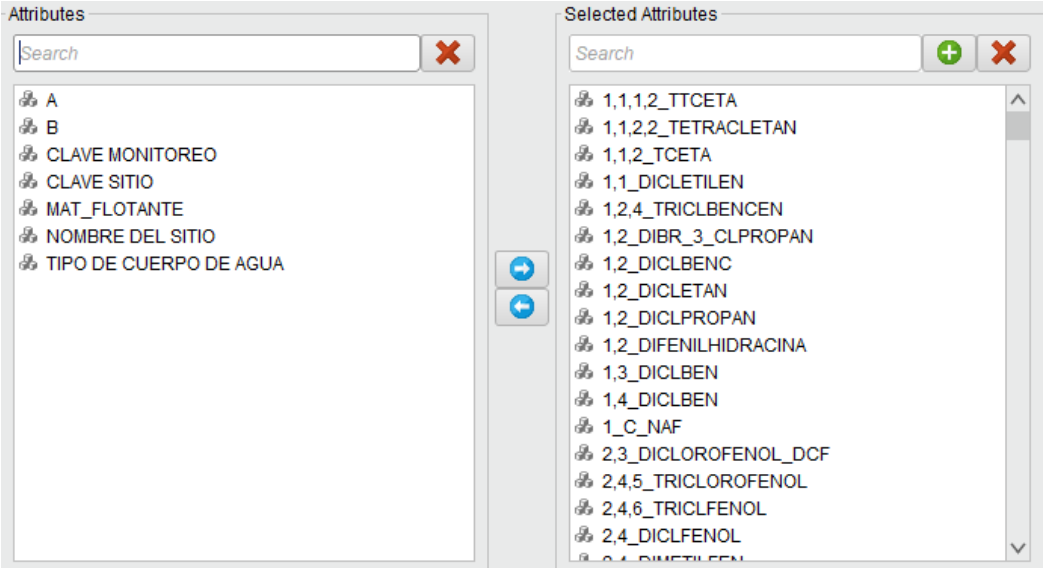

Imagen 13. *Selección de todos los elementos que sean contaminantes*

Los dos últimos filtros son exclusivos para la selección de los contaminantes los cuales serán explicados en el siguiente punto.

4.3 Selección de parámetros a descartar

La selección consta de optar por el parámetro que después nos aportará valor en el proceso de minería y creación de *dashboards*. Los parámetros que se descartan son aquellos que estén completamente vacíos. Gracias a las estadísticas de RapidMiner nos permite conocer

a más profundidad los parámetros y cambiar las posiciones de los parámetros para posicionar los registros vacíos en puestos superiores según el número de pérdidas.

Name	Type	Missing
SB_TOT	Nominal	2823
TEMP_AGUA_30	Nominal	2823
TEMP_AGUA_31	Nominal	2823
TEMP_AGUA_32	Nominal	2823
TEMP_AGUA_33	Nominal	2823
TEMP_AGUA_34	Nominal	2823
BA_SOL	Nominal	2823

***Imagen 14**. Vista estadística de contaminantes en RapidMiner*

Es importante recordar que los datos cuentan con 2,824 registros para cada parámetro. Así que los registros más cercanos a este número en cuestión de pérdidas se encontraran vacíos, ir al *Anexo Tabla 2*. Sin embargo, hay algunos parámetros que RapidMiner descartó por estar completamente vacíos, siendo los siguientes: OH, H_HELMINTOS, MAT_FLOTANTE, SULFUROS_SOL, BA_SOL, SB_SOL, DUR_Ca y DUR_Mg.

4.4 Casos Experimentales

Para identificar los contaminantes vacíos, tipos de contaminantes, contaminantes más comunes y contaminantes menos comunes se plantean diferentes experimentos que están basados en la clasificación de contaminantes hechos en (Naseem, y otros, 2021).

4.5 Modelo de procesamiento

El modelo de procesamiento consiste en el flujo que siguieron los datos buscando cumplir siempre el objetivo de encontrar patrones de comportamiento en los diferentes contaminantes presentes en el agua del estado de Puebla. Se hizo una explicación del modelo, pero la información de los diversos operadores usados en RapidMiner proviene directamente de (RapidMiner, 2021).

4.5.1 Pre-procesamiento

En esta fase se usan tres operadores que sirven para la limpieza y preparación de los datos para los diferentes experimentos a los que son sometidos. Todos los experimentos usan el mismo tipo de modelo, sin embargo, cambian a la hora de seleccionar los datos.

El primer operador *Retrive* hace una llamada a los datos que fueron cargados previamente, el segundo operador es llamado *subprocess* que contiene a su vez otros operadores. Los operadores que contiene son los mostrados en la *Imagen 12.* El tercer operador de selección de atributos consiste en una verificación de los elementos contaminantes en el filtro anterior. A pesar de estar “visualmente” terminado todo lo que conforma el proceso de pre-procesamiento la realidad es que se debe agregar un *multiply* que repite el resultado del proceso anteriormente mencionado para que pueda ser usado para los filtros que lo proceden.

4.5.2 Procesamiento

El procesamiento de los datos después de la limpieza nos otorga diferentes resultados los cuales son de gran provecho. En el primer y segundo bloque se elimina la simbología atípica. Dicha simbología son los signos de *mayor que* y *menor que*. Descartar algunos parámetros (columnas de Excel) que no son de utilidad para el *parse number* y el ultimo filtro es la selección del parámetro usado para la medida de calidad del agua mediante el porcentaje de oxígeno.

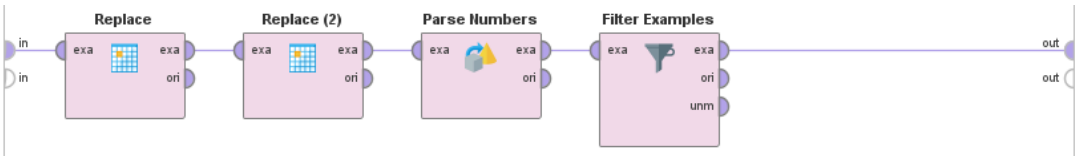

Imagen 15. *Limpieza de datos*

El parámetro más fácil de explicar es *Correlation Matrix* gracias a que se le asigna pesos de cada uno de los elementos para posteriormente aplicar una técnica estadística llamada relación buscando algún tipo de "parentesco" o "relación" entre cada uno de los elementos para finalizar con una tabla que es llamada "matriz de correlación".

La última tarea es la más extensa y genera diversos gráficos del modelo de procesamiento de contaminantes, tales como, gráficas de *árbol* de decisión, vectores de rendimiento y la validación. Para que estas graficas sean conseguidas se debe seguir el proceso que comienza con la selección de roles, que en esencia se trata de priorizar o posicionar en una posición más alta a algún elemento, y el elemento al que se refiere es la saturación de oxígeno disuelto con el código de *OD_%*. La razón de esta jerarquización es debido a que a mayor cantidad de oxígeno disuelto en el agua se tiene mejor calidad. El operador *Discretize* consiste en agrupar los valores en diferentes estados para limitar el número de datos y de esta forma "fragmentarlos" para crear conjuntos de valores más manejables, en este se agruparon de la siguiente manera:

Nombre de clase	Límite de las clases	Agua contaminada
POBRE	60	SÍ
ACEPTABLE	79	SÍ
ADECUADO	89	NO
EXCELENTE	100	NO

Tabla 9. *Clases para los datos en operador de Discretización*

Los números de la tabla anterior son representados por porcentajes según la presencia que tenga cada elemento, es decir, habiendo un 60% de presencia de algún elemento será clasificado con la nota de "POBRE", esto será de gran utilidad para el árbol de decisiones. El operador *Normalize* es usado para escalar valores para que se ajusten a un rango específico y se aplicó el método de *Z_transformation* que consiste en restar la media de todos los datos y luego los divide por la desviación estándar finalizando en una distribución

de los datos con una media de cero y una varianza de uno por lo que no se alejan demasiado del centro preparando aún más el proceso para el árbol de decisión que es el último bloque.

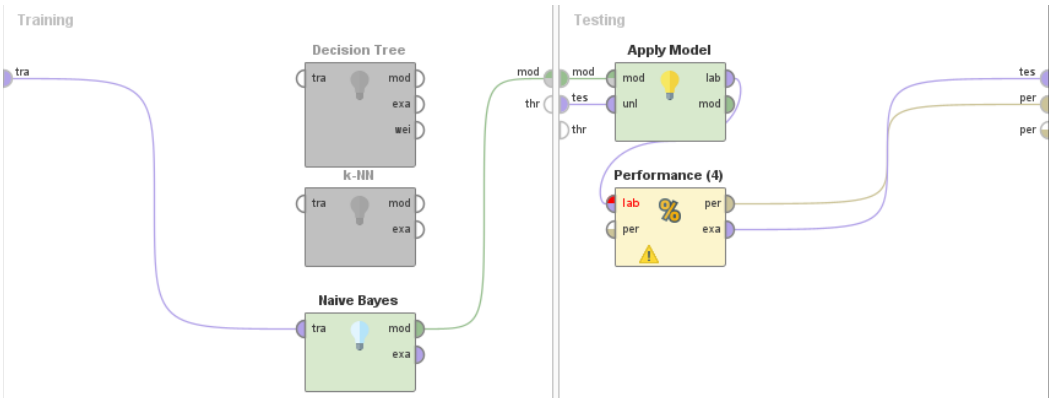

Imagen 16. *Vista interior en operador de Validation en RapidMiner*

En un árbol de decisión es importante ejecutar un entrenamiento para perfeccionar el proceso de clasificación de los datos, se ejecuta el entrenamiento al llegar al 90% de eficacia se corta y se repite el proceso de entrenamiento, guardando los mejores resultados para usarlos en el siguiente ciclo, una vez culminado el entrenamiento se aplica el modelo con la tabla de pesos evaluando el desempeño y enviando los resultados al operador.

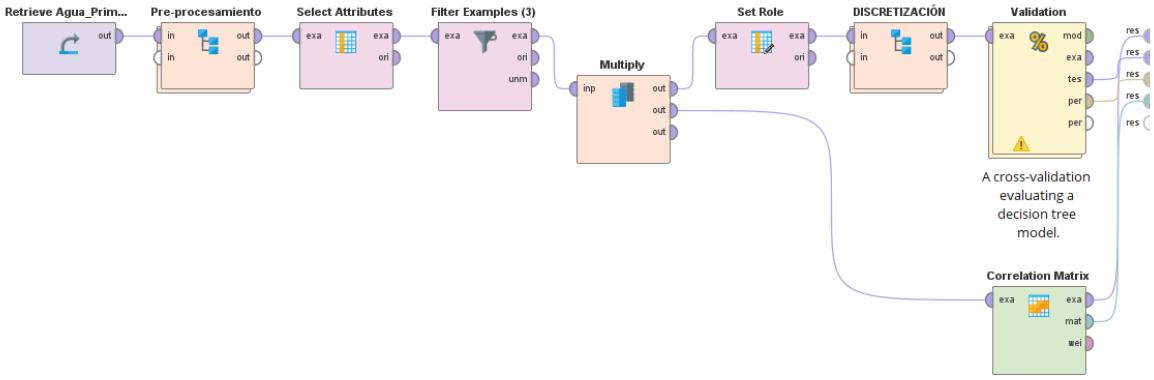

Imagen 17. *Modelo de procesamiento de datos*

4.6 Resultados

Los experimentos contenían una gran cantidad de contaminantes los cuales son analizados. Un elemento importante que se debe presentar previo a la explicación de los experimentos hechos son los resultados generales de todos experimentos para tener que

al ver de forma individual cada uno de estos se comprenda de la mejor forma posible, los resultados generales fueron puestos en una tabla como la siguiente:

Tipo de parámetro	Número de registros con porcentaje de pérdida de 99 a 100	Número de registros con porcentaje menor a 90	Contaminantes que se conservaron
Físicos	55	5	TEMP_AMB, TEMP_AGUA, TURBIEDAD, CONDUC_CAMPO
Químicos	13	8	pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT, OD_%
Biológicos	0	2	COLI_FEC, COLI_TOT, E_COLI
Aniones	11	1	N_NO3
Cationes	10	0	
Metales pesados	15	6	AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT

Tabla 10. *Número de registros que tienen un porcentaje de pérdida mayor al 90%*

Los parámetros biológicos no entran en la anterior tabla debido a que esta sólo contiene dos contaminantes las cuales al sumar su porcentaje de pérdida da como resultado 2.80% que es poco a comparación a los rangos de perdida que van de un 99% a un 100%. Una vez que se explica el escenario general de los experimentos se realiza la explicación de los resultados de los experimentos individualmente teniendo siempre en cuenta que los resultados se mostraran en tres elementos importantes que son: grafica de asignación de pesos, matriz de correlación y árbol de decisión.

Casos	Contaminantes	F-measure Tree Decision	F-measure Knn	F-measure Naive Bayes
Físicos	TEMP_AMB, TEMP_AGUA, TURBIEDAD, CONDUC_CAMPO	70.71%	67.68%	54.43%
Químicos	pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT	78.54%	73.84%	75.23%
Biológicos	COLI_FEC, COLI_TOT, E_COLI	56.98%	45.77%	63.78%
Metales pesados	AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT	60.52%	51.67%	17.14%
Físicos – Químicos	TEMP_AMB, TEMP_AGUA, TURBIEDAD, CONDUC_CAMPO, pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT	75.26%	75.24%	76%
Físicos – Metales pesados	TEMP_AMB, TEMP_AGUA, TURBIEDAD, CONDUC_CAMPO, AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT	67.14%	68.76%	20.46%
Químicos – Metales pesados	pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT, AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT	77.82%	74.62%	76.70%
Biológicos – Metales pesados	COLI_FEC, COLI_TOT, E_COLI, AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT	61.15%	57.67%	50.19%
Biológicos - Químicos	COLI_FEC, COLI_TOT, E_COLI, pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT	78.71%	69.63%	73.89%
Combinación los cuatro primeros casos	TEMP_AMB, TEMP_AGUA, TURBIEDAD, CONDUC_CAMPO, pH_CAMPO, DUR_TOT, DBO_SOL,DBO_TOT, DQO_SOL, DQO_TOT COLI_FEC, COLI_TOT, E_COLI, AS_TOT, CR_TOT, PB_TOT, CD_TOT, NI_TOT, HG_TOT	78.19%	72.40%	75.54%

Tabla 11. *Porcentaje de acierto para los experimentos y sus contaminantes.*

4.6.1 Experimento parámetros físicos

Las relaciones encontradas entre distintos parámetros son en la conductividad eléctrica y el porcentaje de oxígeno disuelto. Se trata de una relación negativa por lo que si uno baja el otro baja también de forma simultánea. Para la segunda relación se vuelve a presentar el porcentaje de oxígeno disuelto con la temperatura ambiente y la temperatura del agua, parámetros similares y temperaturas ya que la temperatura ambiente y del agua tienen relación.

Attributes	CONDUC_CAMPO	OD_%	TURBIEDAD	TEMP_AMB	TEMP_AGUA
CONDUC_CAMPO	1	-0.476	-0.058	0.012	0.134
OD_%	-0.476	1	0.037	0.282	0.258
TURBIEDAD	-0.058	0.037	1	0.026	0.074
TEMP_AMB	0.012	0.282	0.026	1	0.737
TEMP_AGUA	0.134	0.258	0.074	0.737	1

***Imagen 14**. Matriz de correlación en parámetros físicos*

4.6.2 Experimento parámetros químicos

En este experimento se presentan tres grupos importantes los cuales son: grupo de demanda de oxígeno, grupo del *porcentaje de oxígeno disuelto* y el tercero es el *oxígeno disuelto*. Para el primer grupo, las demandas de oxígeno se dividen en bioquímico y química donde se hace una distinción entre total y soluble. La relación más alta se presenta entre *demanda química soluble* con la *demanda bioquímica soluble* con 0.906, siguiendo con la *demanda química total* con *demanda bioquímica soluble* y *demanda química soluble* con valores de 0.765 y 0.760 respectivamente. La última relación de este grupo son las más bajas registradas y se trata de la demanda bioquímica total con la *demanda bioquímica soluble*, la *demanda química soluble* con la *demanda bioquímica total* con la *demanda bioquímica total.* La menor de ellas es la *demanda química total* con la *demanda bioquímica total*, con registros de 0.677, 0.633 y 0.557. El segundo gran grupo es entre el *porcentaje de oxígeno disuelto* y las demandas tanto *bioquímicas como químicas totales* y *solubles*, donde la relación es negativa, esto quiere decir que si una de ellas baja afecta al resto. El tercer y último grupo es el *oxígeno disuelto* con las demandas, se repite la situación como

el anterior grupo de relaciones negativas. Hay una más relación entre *el oxígeno disuelto* y el *porcentaje de oxígeno disuelto*, pero no tiene tanta variedad como los tres anteriores mencionados anteriormente.

Attributes	DBO_SOL	DBO_TOT	DQO_SOL	DQO_TOT	pH_CAMPO	OD_%	OD_mg/L	DUR_TOT
DBO_SOL	1	0.677	0.906	0.765	-0.057	-0.438	-0.418	0.045
DBO_TOT	0.677	1	0.633	0.557	-0.027	-0.332	-0.321	0.021
DQO_SOL	0.906	0.633	1	0.760	-0.061	-0.483	-0.459	0.051
DQO_TOT	0.765	0.557	0.760	1	-0.056	-0.373	-0.360	0.066
pH_CAMPO	-0.057	-0.027	-0.061	-0.056	1	0.140	0.141	0.132
OD_%	-0.438	-0.332	-0.483	-0.373	0.140	1	0.933	-0.060
OD_mg/L	-0.418	-0.321	-0.459	-0.360	0.141	0.933	1	-0.062
DUR_TOT	0.045	0.021	0.051	0.066	0.132	-0.060	-0.062	1

Imagen 15. *Matriz de correlación de parámetros químicos*

4.6.3 Experimento parámetros biológicos

La matriz de relación de los parámetros biológicos es la menos variada en cuanto a la cantidad de parámetros. La primera relación es entre parámetros similares tratándose al mismo tiempo de la relación más alta de la matriz entre los *coliformes fecales* y los *coliformes fecales totales*. La mayor relación es la de *Escherichia coli* y *Coliformes fecales* de casi una unidad, *Escherichia coli* tiene una segunda relación con los *coliformes fecales* pero es la menor de toda la matriz.

Attributes	COLI_FEC	COLI_T...	E_COLI	OD_%
COLI_FEC	1	0.743	0.992	-0.045
COLI_TOT	0.743	1	0.715	-0.059
E_COLI	0.992	0.715	1	-0.032
OD_%	-0.045	-0.059	-0.032	1

Imagen 16. *Matriz de correlación de parámetros biológicos*

4.6.4 Experimento parámetros metales pesados

El registro de *Cadmio* total tiene relación con otros dos parámetros, los cuales son el *níquel total* y *el plomo total* siendo este último la relación más alta registrada en esta matriz. El *níquel* tiene una segunda relación con el cromo total y la tercera con el *plomo total*. No hay que olvidar que existen otras dos relaciones que son tan sutiles y que no deben ser omitidas y es *cadmio total* con el *cromo total* además del *cadmio total* y *arsénico total*.

Attributes	OD_%	AS_TOT	CD_TOT	CR_TOT	HG_TOT	NI_TOT	PB_TOT
OD_%	1	0.108	0.051	-0.016	0.019	-0.032	-0.034
AS_TOT	0.108	1	0.156	0.067	-0.004	0.074	0.062
CD_TOT	0.051	0.156	1	0.169	0.035	0.279	0.393
CR_TOT	-0.016	0.067	0.169	1	-0.001	0.374	0.075
HG_TOT	0.019	-0.004	0.035	-0.001	1	-0.009	-0.005
NI_TOT	-0.032	0.074	0.279	0.374	-0.009	1	0.315
PB_TOT	-0.034	0.062	0.393	0.075	-0.005	0.315	1

Imagen 17. *Matriz de correlación de parámetros metales pesados*

4.6.5 Experimento parámetros físicos-químicos

La combinación de estas dos clases de parámetros permitió crear nuevas relaciones. La primera es la conductividad del campo tiene relación positiva con los parámetros de las demandas se oxígeno tanto químicas como bioquímicas solubles o totales. Otra relación de la conductividad es con la dureza total del agua, las relaciones negativas son con el parámetro de *porcentaje de oxígeno*. Volviendo con la *dureza total* muestra delación la turbiedad y ambos parámetros *de temperatura ambiental* y *temperatura del agua*, al igual que el *pH*.

Attributes	DBO_SOL	DBO_TOT	DQO_SOL	DQO_TOT	CONDUC_CAMPO	pH_CA...	OD_%	TURBIEDAD	DUR_TOT	TEMP_AMB	TEMP_AGUA
DBO_SOL	1	0.674	0.906	0.765	0.499	-0.054	-0.437	0.014	0.046	-0.137	-0.028
DBO_TOT	0.674	1	0.632	0.554	0.338	-0.024	-0.331	0.019	0.020	-0.107	-0.049
DQO_SOL	0.906	0.632	1	0.761	0.499	-0.060	-0.484	0.051	0.051	-0.159	-0.047
DQO_TOT	0.765	0.554	0.761	1	0.412	-0.053	-0.374	0.091	0.059	-0.118	-0.045
CONDUC_CAMPO	0.499	0.338	0.499	0.412	1	0.183	-0.486	-0.061	0.508	0.014	0.137
pH_CAMPO	-0.054	-0.024	-0.060	-0.053	0.183	1	0.137	-0.097	0.133	0.273	0.305
OD_%	-0.437	-0.331	-0.484	-0.374	-0.486	0.137	1	0.037	-0.062	0.288	0.261
TURBIEDAD	0.014	0.019	0.051	0.091	-0.061	-0.097	0.037	1	0.201	0.028	0.080
DUR_TOT	0.046	0.020	0.051	0.059	0.508	0.133	-0.062	0.201	1	0.263	0.303
TEMP_AMB	-0.137	-0.107	-0.159	-0.118	0.014	0.273	0.288	0.028	0.263	1	0.735
TEMP_AGUA	-0.028	-0.049	-0.047	-0.045	0.137	0.305	0.261	0.080	0.303	0.735	1

Imagen 18. *Matriz de correlación de parámetros físicos-químicos*

4.6.6 Experimento parámetros físicos-metales pesados

El parámetro de turbiedad tiene relación con varios parámetros que son los siguientes: cadmio total, arsénico total, cromo total y níquel, esto quiere decir que la presencia de estos metales en el agua produce turbiedad en el agua, posiblemente por el tamaño de las partículas de los parámetros.

Attributes	CONDUC_CAMPO	OD_%	TURBIEDAD	AS_TOT	CD_TOT	CR_TOT	HG_TOT	NI_TOT	PB_TOT	TEMP_AMB	TEMP_AGUA
CONDUC_CAMPO	1	-0.470	-0.061	0.121	-0.034	-0.010	-0.019	-0.035	0.017	0.025	0.147
OD_%	-0.470	1	0.041	0.109	0.057	-0.026	0.019	-0.033	-0.030	0.271	0.244
TURBIEDAD	-0.061	0.041	1	0.232	0.299	0.351	0.002	0.416	0.146	0.025	0.077
AS_TOT	0.121	0.109	0.232	1	0.158	0.068	-0.004	0.077	0.063	0.192	0.216
CD_TOT	-0.034	0.057	0.299	0.158	1	0.126	0.036	0.236	0.400	0.004	0.014
CR_TOT	-0.010	-0.026	0.351	0.068	0.126	1	-0.003	0.337	0.061	-0.008	-0.026
HG_TOT	-0.019	0.019	0.002	-0.004	0.036	-0.003	1	-0.010	-0.005	-0.001	-0.023
NI_TOT	-0.035	-0.033	0.416	0.077	0.236	0.337	-0.010	1	0.296	-0.011	-0.017
PB_TOT	0.017	-0.030	0.146	0.063	0.400	0.061	-0.005	0.296	1	0.014	-0.000
TEMP_AMB	0.025	0.271	0.025	0.192	0.004	-0.008	-0.001	-0.011	0.014	1	0.736
TEMP_AGUA	0.147	0.244	0.077	0.216	0.014	-0.026	-0.023	-0.017	-0.000	0.736	1

Imagen 19. *Matriz de correlación de parámetros físicos-metales pesados*

4.6.7 Experimento parámetros químico-metales pesados

Para este experimento sólo hay una nueva relación y es la *dureza total* con *el arsénico total*, esto da entender que el *arsénico* vuelve la muestra del agua algo acida y posiblemente con la presencia del parámetro de *pH* ocurra una relación.

Attributes	DBO_SOL	DBO_TOT	DQO_SOL	DQO_TOT	pH_CAMPO	OD_%	OD_mg/L	AS_TOT	CD_TOT	CR_TOT	HG_TOT	NI_TOT	PB_TOT	DUR_TOT
DBO_SOL	1	0.676	0.906	0.767	-0.055	-0.440	-0.428	-0.073	-0.043	0.031	-0.009	-0.010	-0.007	0.044
DBO_TOT	0.676	1	0.635	0.549	-0.025	-0.337	-0.327	-0.060	-0.028	0.043	-0.008	0.014	0.006	0.020
DQO_SOL	0.906	0.635	1	0.765	-0.061	-0.487	-0.470	-0.069	-0.042	0.055	-0.000	0.028	0.009	0.050
DQO_TOT	0.767	0.549	0.765	1	-0.054	-0.380	-0.368	-0.045	-0.006	0.090	-0.003	0.063	0.019	0.066
pH_CAMPO	-0.055	-0.025	-0.061	-0.054	1	0.134	0.128	0.043	-0.048	0.010	-0.021	-0.065	-0.049	0.134
OD_%	-0.440	-0.337	-0.487	-0.380	0.134	1	0.959	0.110	0.051	-0.027	0.020	-0.040	-0.038	-0.058
OD_mg/L	-0.428	-0.327	-0.470	-0.368	0.128	0.959	1	0.101	0.028	-0.018	0.028	-0.046	-0.066	-0.063
AS_TOT	-0.073	-0.060	-0.069	-0.045	0.043	0.110	0.101	1	0.138	0.072	-0.004	0.075	0.050	0.414
CD_TOT	-0.043	-0.028	-0.042	-0.006	-0.048	0.051	0.028	0.138	1	0.174	0.037	0.265	0.328	0.063
CR_TOT	0.031	0.043	0.055	0.090	0.010	-0.027	-0.018	0.072	0.174	1	-0.002	0.358	0.084	0.066
HG_TOT	-0.009	-0.008	-0.000	-0.003	-0.021	0.020	0.028	-0.004	0.037	-0.002	1	-0.010	-0.005	-0.003
NI_TOT	-0.010	0.014	0.028	0.063	-0.065	-0.040	-0.046	0.075	0.265	0.358	-0.010	1	0.316	0.090
PB_TOT	-0.007	0.006	0.009	0.019	-0.049	-0.038	-0.066	0.050	0.328	0.084	-0.005	0.316	1	0.033
DUR_TOT	0.044	0.020	0.050	0.066	0.134	-0.058	-0.063	0.414	0.063	0.066	-0.003	0.090	0.033	1

Imagen 20. *Matriz de correlación de parámetros químico-metales pesados*

4.6.8 Experimento parámetros biológicos-metales pesados

Exactamente las mismas relaciones que los experimentos individuales, es decir, *biológicos* y *metales pesados*, pero al juntarlos no ocurre ningún tipo de relación nueva en otros parámetros de ambos tipos de contaminantes.

Attributes	COLI_FEC	COLI_TOT	E_COLI	OD_%	AS_TOT	CD_TOT	CR_TOT	HG_TOT	NI_TOT	PB_TOT
COLI_FEC	1	0.743	0.992	-0.045	-0.017	-0.028	0.127	-0.008	0.030	-0.024
COLI_TOT	0.743	1	0.715	-0.060	-0.016	-0.028	0.073	-0.007	0.008	-0.020
E_COLI	0.992	0.715	1	-0.031	-0.012	-0.026	0.127	-0.006	0.020	-0.026
OD_%	-0.045	-0.060	-0.031	1	0.108	0.053	-0.022	0.019	-0.030	-0.031
AS_TOT	-0.017	-0.016	-0.012	0.108	1	0.151	0.071	-0.004	0.077	0.064
CD_TOT	-0.028	-0.028	-0.026	0.053	0.151	1	0.174	0.036	0.283	0.382
CR_TOT	0.127	0.073	0.127	-0.022	0.071	0.174	1	-0.001	0.377	0.080
HG_TOT	-0.008	-0.007	-0.006	0.019	-0.004	0.036	-0.001	1	-0.009	-0.005
NI_TOT	0.030	0.008	0.020	-0.030	0.077	0.283	0.377	-0.009	1	0.325
PB_TOT	-0.024	-0.020	-0.026	-0.031	0.064	0.382	0.080	-0.005	0.325	1

Imagen 21. *Matriz de correlación de parámetros biológicos-metales pesados*

4.6.9 Experimento parámetros biológicos-químicos

Al igual que el experimento anterior (*Experimento parámetros biológicos-metales pesados*) no se presenta una nueva relación entre los parámetros si los comparamos individualmente.

Attributes	COLI_FEC	COLI_TOT	E_COLI	DBO_SOL	DBO_TOT	DQO_SOL	DQO_TOT	pH_CAMPO	OD_%	DUR_TOT
COLI_FEC	1	0.797	0.992	0.075	0.059	0.047	0.054	-0.002	-0.043	-0.000
COLI_TOT	0.797	1	0.765	0.066	0.050	0.044	0.047	0.001	-0.051	0.006
E_COLI	0.992	0.765	1	0.069	0.054	0.038	0.047	0.001	-0.030	0.003
DBO_SOL	0.075	0.066	0.069	1	0.677	0.906	0.765	-0.057	-0.438	0.045
DBO_TOT	0.059	0.050	0.054	0.677	1	0.633	0.557	-0.027	-0.332	0.021
DQO_SOL	0.047	0.044	0.038	0.906	0.633	1	0.760	-0.061	-0.483	0.051
DQO_TOT	0.054	0.047	0.047	0.765	0.557	0.760	1	-0.056	-0.373	0.066
pH_CAMPO	-0.002	0.001	0.001	-0.057	-0.027	-0.061	-0.056	1	0.140	0.132
OD_%	-0.043	-0.051	-0.030	-0.438	-0.332	-0.483	-0.373	0.140	1	-0.061
DUR_TOT	-0.000	0.006	0.003	0.045	0.021	0.051	0.066	0.132	-0.061	1

Imagen 22. *Matriz de correlación de parámetros biológicos-químicos*

4.6.10 Experimento de todos los parámetros juntos

Simplemente todos los contaminantes juntos sin, no existe alguna relación entre los distintos tipos de contaminantes que se presentan y para ser más exactos no hay apenas relación entre contaminantes del mismo tipo, las relaciones más cercanas que se pueden llegar a percibir de los contaminantes es entre contaminantes similares, un ejemplo de ello es el contaminante COLI_FEC que la relación más cercana que llegó a presentar es consigo mismo.

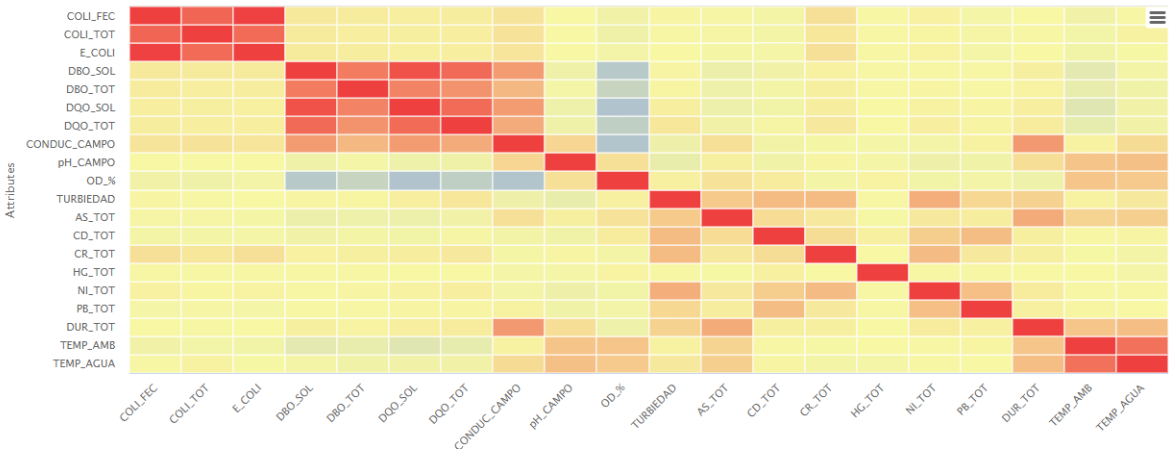

Imagen 23. *Matriz de correlación de parámetros físico-biológicos-químico-metales pesados*

5. Conclusiones y trabajo a futuro

5.1 Conclusión

En este trabajo se clasificaron los contaminantes de agua según su tipo. Se enlistaron los diferentes puntos de muestreo y se realizaron algunos experimentos para los contaminantes buscando cuales son los que estaban más presentes. Se usó el método planteado por la CONAGUA el cual consiste en emplear el parámetro de oxígeno, tanto en porcentaje como la cantidad, que está presente en el agua como medida de su calidad. Además, se consideró la saturación y demanda de oxígeno aplicando distintos algoritmos para el análisis de datos. Se obtuvieron los porcentajes de aciertos o éxito expuestos en la *Tabla 11* y resultó que considerando el algoritmo de árboles de decisión se obtiene un 78,71% de precisión con los parámetros Biológicos - Químicos, el algoritmo KNN un 74.62% con los parámetros Químicos – Metales pesados y el algoritmo de Naive Bayes un 75.54% con la combinación de todos los experimentos propuestos. Se puede concluir que el que tuvo mayor relevancia es el algoritmo de árbol de decisión, los porcentajes obtenidos son aceptables pese a las pérdidas de datos y a algunas incongruencias en el repositorio original de datos. Cabe mencionar que las relaciones planteadas en los resultados fueron satisfactorias debido a que se presentaron relaciones que resultan satisfactorias y puntos clave que impulsan la continuidad de estudio de este trabajo de tesis.

5.2 Trabajo a futuro

Como trabajos a futuro se propone la implementación en varios sectores que podrían ser de mejora para la continuación de este trabajo. El primer trabajo consiste en buscar otro repositorio de datos como complemento o reemplazo del usado en este trabajo. Esto para indagar sobre las fuentes de información donde fueron obtenidos de forma que los datos sean más recientes y exactos.

La segunda mejora consiste en llevar un mejor registro de los datos considerando tres tipos de ambientes: el primero son los ambientes controlados como las plantas de tratamiento de agua, el segundo son los ambientes no controlados como ríos o espacios donde la presencia humana es nula o escasa, el tercero es la combinación de los dos anteriores estarán ubicados en localidades rurales. Lo recomendable es que haya puntos de registros en los tres ambientes mencionados ya que habrá más variedad en los registros.

El último trabajo a futuro consiste en aplicar otras técnicas de minería de datos y generar más combinaciones de datos para lograr una precisión más alta en los resultados. Las técnicas propuestas que se pueden aplicar son: análisis de regresión, detección de valores atípicos y clasificación. No obstante, la aplicación de inteligencia artificial al análisis de datos facilitaría la tarea de minería de datos como redes neuronales o procesamiento de memoria a largo plazo por el lapso de tiempo en que los datos son registrados.

6. Bibliografía

Aslam, B., Maqsoom, A., Hassan Cheema, A., Ullah, F., Alharbi, A., & Muhammad, I. (2022). *Water Quality Management Using Hybrid.* Pakistan: Digital Object Identifier.

Biodiversidad, C. N. (09 de 02 de 2023). *Biodiversidad mexicana*. Obtenido de https://www.biodiversidad.gob.mx/monitoreo/simoh-mx/cuerpos-de-agua

Board, S. W. (2023). Folleto Informativo Oxígeno Disuelto. *Water Board*, 5. Obtenido de https://www.waterboards.ca.gov/water_issues/programs/swamp/docs/cwt/guidance/3110sp.pdf

Board, S. W. (07 de 09 de 2023). Folleto informativo pH. *Water Boards*, 5. Obtenido de https://www.google.com/search?q=medidas+ph+agua&sca_esv=600294646&rlz=1C1CHZN_esMX1037MX1037&sxsrf=ACQVn0_GjHmCaOAvblLj99SF9JWWu0EvRw%3A1705882902285&ei=FrWtZZeKEZiekPIPg_mIkAk&ved=0ahUKEwiX7ayD3e-DAxUYD0QIHYM8ApIQ4dUDCBA&uact=5&oq=medidas+ph+agua&gs_lp=E

Carbotecnia. (13 de 10 de 2021). *Significado de los sólidos disueltos totales en el agua*. Obtenido de https://www.carbotecnia.info/aprendizaje/quimica-del-agua/solidos-disueltos-totales-tds/#:~:text=Los%20s%C3%B3lidos%20disueltos%20totales%20(SDT%2C%20o%20TDS%20por%20sus%20siglas,se%20reporta%20en%20mg%2FL.

CONAGUA. (28 de 02 de 2024). *Calidad del agua en México*. Obtenido de https://www.gob.mx/conagua/articulos/calidad-del-agua

demográfico, M. p. (28 de 02 de 2020). *Categorías y tipos de masas de agua*. Obtenido de https://www.miteco.gob.es/es/agua/temas/estado-y-calidad-de-las-aguas/aguas-superficiales/categorias_y_tipos_de_masas_de_agua.html#:~:text=Aguas%20l%C3%B3ticas%20o%20corrientes%3A%20masas,%2C%20riachuelos%2C%20arroyos%2C%20ramblas.

Enriquez, E. O. (2011). Indicadores fecales y parógenos en agua descargada al Río Bravo. *SciELO*, 9. Obtenido de https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-57792011000400449

EUROINNOVA. (28 de 01 de 2022). *como se mide la calidad del agua*. Obtenido de https://www.euroinnova.mx/como-se-mide-la-calidad-del-agua

Fuertes, J. G. (10 de 09 de 2018). *Los usos del agua*. España: iAgua.

González Villareal, F. (1986). *Plan Nacional Hidráulico 1975.* Obtenido de Academia.edu: https://www.academia.edu/33265934/Plan_Nacional_Hidr%C3%A1ulico_1975

Haixia, L., Jianhong, C., & Xiangwei, B. (22 de 02 de 2021). Feature Extraction of Marine Water Pollution Based on Data Mining. *Special Issue Mathematical Modeling and Computational Methods in Science and Engineering III*, pág. 15.

Hermann, E., & Prunes, E. (2022). ¿Qué es el agua subterránea y por qué es tan importante? *Descubre WWF*, 1.

IBM. (02 de 03 de 2021). *¿Qué es Knn?* Obtenido de https://www.ibm.com/mx-es/topics/knn

IBM. (02 de 03 de 2021). *¿Qué es un árbol de decisión?* Obtenido de https://www.ibm.com/mx-es/topics/decision-trees

IBM. (02 de 03 de 2021). *Guía de CRISP-DM.* Obtenido de https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

Jatnika, H. (21 de 06 de 2021). Analysis of Data Mining in the Group of Water. *Journal of Physics: Conference Series*, pág. 8.

LENNTECH. (25 de 07 de 2019). *Evaluación de la calidad el agua*. Obtenido de https://www.lenntech.es/la-evaluacion-de-la-calidad-agua-faq-calidad-agua#:~:text=La%20calidad%20del%20agua%20se,de%20iones%20de%20hidr%C3%B3geno%20presentes.

Mosle Hmoud, A.-A., & Fawaz Waselallah, A. (2021). Modelling and Prediction of Water Quality by Using. *Sustainability*, 18.

Naseem, A., Muhammad Izzuddin, S. I., Mardiana Idayu, A., Khalid, U., Mohamad Shaiful, M. Y., Mohd Talha, A., . . . Yazan Khalaf, A. A. (28 de 02 de 2021). Water - MDPI. *Academic Open Access Publishing*, pág. 34. Obtenido de Modification of the Water Quality Index (WQI) Process for Simple Calculation Using the Multi-Criteria Decisio-Making (MCDM) Method: A Review: https://www.mdpi.com/2073-4441/13/7/905

optek. (03 de 04 de 2023). *Mediciones de conductividad*. Obtenido de https://www.optek.com/es/Mediciones-de-conductividad.asp

Raeburn, A. (22 de 02 de 2024). *¿Cuál es la diferencia entre exactitud y precisión?* Obtenido de asana: https://asana.com/es/resources/accuracy-vs-precision

RapidMiner. (24 de 03 de 2021). *RapidMiner Documentation.* Obtenido de https://docs.rapidminer.com/9.9/studio/operators/utility/multiply.html

Roman, V. (25 de 04 de 2019). *Algoritmos de Naive Bayes: Fundamentos e implementación*. Obtenido de Medium: https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

Sameer, A., & Ashok K., K. (19 de 7 de 2021). Pattern recognition of water quality variance in Yamuna River (India) using hierarchical agglomerative cluster and principal component analyses. *Environmental Monitoring and Assessment*, pág. 18. Obtenido de Springer Link: https://link.springer.com/article/10.1007/s10661-021-09318-1

Theyazn H. H, A., Mohammed, A.-Y., Hasan, A., & Mashael, M. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics and Biomechanics*, 12.

Tulcingo, A. M. (2018 - 2021). *Toma de protesta.* Tulcingo: Plan Municipal de Desarrollo.

UNESCO. (01 de 12 de 2020). *Pueblos Mágicos de México*. Obtenido de https://www.gob.mx/sectur/articulos/pueblos-magicos-206528

7. Anexos

Nombre de sitio		Rango de casos
Ampliación unión antorchista Atoyac la carmelita p1 (26) Después de la confluencia con el río san francisco Echeverría antes de la planta de tratamiento Atoyac sur La autopista p1 (24) Puebla san francisco (a) aguas arriba Puente carretera san martin-tlaxcala p1(15) Rio nexapa 20 Tlapalac antes de confluencia con rio Atoyac Bacardi y compañía, s.a. de C.V. (aguas abajo) Bacardi y compañía, s.a. de C.V. (aguas arriba) Cervecería Cuauhtémoc Moctezuma, s.a. de C.V. (aguas abajo) Después de bachoco Nexapa entrada a puctla Nexapa puente don roque Nexapa-las fajanas p1 (33) Puctla	Puente san francisco totimehuacan Rio Acatlán Rio Atoyac (después de la confluencia del rio mixteco) Rio mixteco (antes de confluir al rio Atoyac) Rio Necaxa 4 Rio Necaxa 5 Rio nexapa 1 Rio nexapa 10 Rio nexapa 11 Rio nexapa 14 Rio nexapa 16 Rio nexapa 17 Rio nexapa 19 Rio nexapa 21 Rio nexapa 5 Rio nexapa 6 Rio nexapa 8 Rio nexapa 9 Rio san Lorenzo 1 Sabormex, s.a. de C.V. (antes cafés de Veracruz, s.a. de C.V.) (aguas abajo) Lázaro Cárdenas	43 – 41
Rio nexapa 0 Rio nexapa 12 Sabormex, s.a. de C.V. (antes cafés de Veracruz, s.a. de C.V.) (aguas arriba) Agua fría Nexapa los molinos p1 (31) Rio Necaxa 1 Rio nexapa 7 Villa Ávila Camacho Rio cazones 5 Rio nexapa 4 Rio nexapa 15 Paso jardín Rio nexapa 13 Barranca honda Santa maría Texmelucan Puente marquez Arroyo metlapanapa Arroyo prieto Ra-1 rio Atoyac, antes tlahuapan (estación testigo) Rio rabanillo	Rio xochiac aguas abajo del colector el Carmen confluencia con rio xochiac Santa maría coronango Arroyo zapatero Puebla alseseca (a) aguas arriba Antes de la unión del rio tlanalapan p1(12) Presa Necaxa 2 Presa Manuel Ávila Camacho 6 Presa Tenango Sistema de represas y corredores biológicos de la cuenca hidrográfica del rio Necaxa 5 Sistema de represas y corredores biológicos de la cuenca hidrográfica del rio Necaxa 7 Sistema de represas y corredores biológicos de la cuenca hidrográfica del rio Necaxa 8 Manuel Ávila Camacho (obra de toma) p1(32) Presa los carros obra de toma Presa Necaxa 1 Presa Manuel Ávila Camacho 1 Presa Manuel Ávila Camacho 5 Presa la soledad	41 – 13

Nombre del sitio		Rango de casos
Moratilla 4 Mpio. Cuayucatepec Municipal s.j. tepeteopan Municipal san José Miahuatlán Xoxtla 1 La poblana Mpio. Zinacatepec Ooap de Tehuacán aeropuerto 3 Pozo ap localidad de ahuehuetzingo Blanca Lilia lendle rugarcia Claudia Sánchez regules Delia Escalante de Sánchez José Rosendo fco. Cortes hdz. La loma Laguna Santiago ovando San Antonio chichicuautla, spr de ri Tepoyanes 4 Ahuatepec puente Balcones del sur soapap Balcones del sur soapap 2 Cervecería Cuauhtémoc Moctezuma, s.a. de C.V. (aguas arriba)	Colonia unión antorchista Rio nexapa 22 San Baltazar tetela pozo centro San Baltazar tetela pozo de la " y" griega Sociedad dos barrancas de acteopan Morelos Sociedad el rosario spr de ri Arroyo metlapanapa 1 Arroyo metlapanapa 2 Arroyo metlapanapa 3 Barranca agua escondida Cuerpo de agua superficial, en apoyo a procedimiento fiscalía del estado y unidad jurídica 1 Cuerpo de agua superficial, en apoyo a procedimiento fiscalía del estado y unidad jurídica 2 Manantial el túnel Rio agua fría, aguas arriba del derrame Rio agua fría, en el sitio del derrame Rio nexapa (puente de puctla, aguas arriba de la descarga de la ptar de izucar de matamoros)	11 – 1

Anexo Tabla 1. Nombre de sitios divididos en rangos

CLAVE	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
COLI_FEC	Coliformes Fecales	NMP/100 mL	Biológico (bacteriano)	12	2812	0.43%	1	992
SDT	Sólidos Disueltos Totales	mg/L	Mineral, metálica y salina	19	2805	0.68%	37.12	28692
N_NO2	Nitrógeno de Nitritos	mg/L	Biológico	23	2801	0.82%	0.021	3.138
TEMP_AMB	Temperatura ambiente	°C	Temperatura	23	2801	0.82%	2.1	38.2
N_NH3	Nitrógeno Amoniacal	mg/L	Bioquímico	31	2793	1.11%	0.02	154.59
N_NO3	Nitrógeno de Nitratos	mg/L	Biológico	34	2790	1.22%	0.02	43.325
P_TOT	Fósforo Total	mg/L	Química	36	2788	1.29%	0.004	39.095
N_ORG	Nitrógeno Orgánico	mg/L	Biológico	37	2787	1.33%	0.001	164.648
N_TOTK	Nitrógeno Kjeldahl	mg/L	Química	44	2780	1.58%	0.023	<5.267
AS_TOT	Arsénico Total	mg/L	Natural	46	2778	1.66%	0.01	0.1
PB_TOT	Plomo Total	mg/L	Metálica	55	2769	1.99%	0.005	0.6
N_TOT	Nitrógeno Total (Cálculo)	mg/L	Química	59	2765	2.13%	0.122	<9.811
CD_TOT	Cadmio Total	mg/L	Química	59	2765	2.13%	0.003	0.16
CR_TOT	Cromo Total	mg/L	Química	65	2759	2.36%	0.005	0.7
HG_TOT	Mercurio Total	mg/L	Natural	69	2755	2.50%	0.0005	0.5
DUR_TOT	Dureza Total	mg CaCO3/L	Mineral	71	2753	2.58%	27.7	2780
COT	Carbono Orgánico Total	mg/L	Biológico	138	2686	5.14%	0.5	1502.052
ABS_UV	Absorción UV	U Abs/cm	Molecular	188	2636	7.13%	0.03	4.26
E_COLI	Escherichia coli	NMP/100 mL	Biológico (bacteriano)	201	2623	7.66%	3	2419600
TURBIEDAD	Turbiedad	UNT	Óptica	202	2622	7.70%	2.5	16780
NI_TOT	Níquel Total	mg/L	Química	203	2621	7.75%	0.001	0.48
pH_CAMPO	Potencial de Hidrógeno	UpH	Mineral	204	2620	7.79%	4.59	13.66
TEMP_AGUA	Temperatura agua	°C	Temperatura	204	2620	7.79%	5.5	48.2
DBO_TOT	Demanda Bioquímica de Oxígeno Total	mg/L	Bioquímico	206	2618	7.87%	2	9000
COLOR_VER	Color Verdadero	U Pt/Co	Molecular, Bioquímica y Química	208	2616	7.95%	5	2000
CONDUC_CAMPO	Conductividad eléctrica	µS/cm	Física	212	2612	8.12%	58	35865
DQO_TOT	Demanda Química de Oxígeno Total	mg/L	Química	226	2598	8.70%	10.1	99200
DBO_SOL	Demanda Bioquímica de Oxígeno Soluble	mg/L	Biológico (orgánico)	285	2539	11.22%	2	3400
DQO_SOL	Demanda Química de Oxígeno Soluble	mg/L	Química	301	2523	11.93%	10	48806.4
COT_SOL	Carbono Orgánico Soluble	mg/L	Biológico	370	2454	15.08%	0.514	782.9
ORTO_PO4	Fósforo Reactivo total (o-fosfatos)	mg/l	Química	378	2446	15.45%	0.007	50.624
TOX_V_15_UT	Toxicidad Vibrio fischeri 15min UT	UT	Bacteriano	404	2420	16.69%	1	73.05
TOX_FIS_D_48UT	Toxicidad Daphnia magna, 48 h	UT	Química	418	2406	17.37%	1.01	18.51
OD_mg/L	Oxígeno Disuelto	mg/L	Química	441	2383	18.51%	1	48

CLAVE	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
OD_%	Oxígeno Disuelto_%	% Saturación	Química	454	2370	19.16%	10	139.17
COLI_TOT	Coliformes Totales	NMP/100 mL	Biológico (bacteriano)	476	2348	20.27%	3	2419600
TOXI_FIS_V_5_UT	Toxicidad Vibrio fischeri 5min UT	UT	Bacteriano	539	2285	23.59%	1.01	89.526
SST	Sólidos Suspendidos Totales	mg/L	Material particulado	626	2198	28.48%	10	18900
SAAM	Sustancias Activas al Azul de Metileno	mg/L	Químico	698	2126	32.83%	0.05	58.1
CAUDAL	Caudal	L/s	Físico cuerpo de agua	705	2119	33.27%	0.001	1923000
CN_TOT	Cianuros Totales	mg/L	Química, Biológico y Metálico	830	1994	41.62%	0.001	0.15
TOX_V_30_UT	Toxicidad Vibrio fischeri, 30min	UT	Bacteriano	2200	624	77.90%	1.02	71.12
TRANSPARENCIA	Transparencia disco de Secchi	m	Óptica	2622	202	92.85%	0.05	4
CONDUC_CAMPO_SUP	Conductividad eléctrica superficie	µS/cm	Físico	2623	201	92.88%	92	3000
TEMP_AGUA_0.5	Temperatura Prof. 0.5 m.	°C	Temperatura	2623	201	92.88%	15.1	31.3
TOX_D_48_SUP_UT	Toxicidad Daphnia magna, Superficial, 48 h	UT	Química	2624	200	92.92%	1.07	1.64
TOX_FIS_SUP_15_UT	Toxicidad Vibrio fischeri Superficial, 15min	UT	Bacteriano	2624	200	92.92%	3.11	37.17
CLOROF_A	Clorofila A	mg/m3	Biológico	2633	191	93.24%	<0.5	393.2
TEMP_AGUA_SUP	Temperatura agua superficie	°C	Temperatura	2633	191	93.24%	15.4	31.3
FE_TOT	Fierro Total	mg/L	Químico y Metálico	2634	190	93.27%	0.02	178.6
TEMP_AGUA_1	Temperatura Prof. 1 m.	°C	Temperatura	2638	186	93.41%	15	30.4
PO4_TOT	Fosfatos Totales (a partir de P total)	mg/L	Bioquímico	2639	185	93.45%	0.0153	13.21
FLUORUROS_TOT	Fluoruros Totales	mg/L	Químico, Bioquímico	2639	185	93.45%	0.2	4.56
MN_TOT	Manganeso Total	mg/L	Metálico	2639	185	93.45%	0.0015	20.05
ALC_TOT	Alcalinidad Total	mg CaCO3/L	Ácido	2641	183	93.52%	57	1650
CO3	Carbonatos	mg CaCO3/L	Ácido Carbónico	2641	183	93.52%	0	240
OH	Hidroxilos	mg CaCO3/L	Biológico	2641	183	93.52%	0	0
K_TOT	Potasio Total	mg/L	Mineral	2645	179	93.66%	1.22	31.8
NA_TOT	Sodio Total	mg/L	Química	2646	178	93.70%	14	457.8
SO4_TOT	Sulfatos Totales	mg/L	Sales	2646	178	93.70%	15.1	792.1
ALC_FEN	Alcalinidad a la Fenoftaleína	mg CaCO3/L	Ácido	2647	177	93.73%	0	135
HCO3	Bicarbonatos	mg CaCO3/L	Salinidad	2647	177	93.73%	41	1650
CLORUROS_TOT	Cloruros Totales	mg/L	Metálico, Salinidad y Bioquímico	2648	176	93.77%	11.98	809.96
pH_CAMPO	Potencial de Hidrógeno superficie	UpH	Mineral	2649	175	93.80%	6.17	9.8
TOX_FIS_SUP_5_UT	Toxicidad Vibrio fischeri Superficial, 5min	UT	Bacteriano	2650	174	93.84%	3.19	60.98

CLAVE PARÁMETRO	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
POT_REDOX_CAMPO_SUP	Potencial REDOX superficie	mV	Química	2650	174	93.84%	-320	382
SIO2	Silica Total (SiO2)	mg/L	Metálico	2652	172	93.91%	0.1	140
OD_%_SUP	Oxígeno Disuelto superficie_%	% Saturación	Química	2653	171	93.94%	10	148.3
OD_mg/L_SUP	Oxígeno Disuelto superficie	mg/L	Química	2653	171	93.94%	1	10.86
CA_TOT	Calcio Total	mg/L	Mineral	2654	170	93.98%	8.4	287.8
TEMP_AGUA_2	Temperatura Prof. 2 m.	°C	Temperatura	2654	170	93.98%	14.9	30
CONDUC_CAMPO_MED	Conductividad eléctrica media	µS/cm	Física	2656	168	94.05%	93	1312
POT_REDOX_CAMPO	Potencial REDOX	mV	Química	2658	166	94.12%	-312.3	518
POT_REDOX_CAMPO_MED	Potencial REDOX media	mV	Química	2658	166	94.12%	-295	384
TOX_D_48_FON_UT	Toxicidad Daphnia magna, Fondo, 48 h	UT	Química	2660	164	94.19%	<1	<1
TOX_FIS_FON_15_UT	Toxicidad Vibrio fischeri Fondo, 15min	UT	Bacteriano	2661	163	94.23%	<1	1.62
PROFUNDIDAD	Profundidad	m	Longitudinal	2661	163	94.23%	0.25	29.5
TEMP_AGUA_3	Temperatura Prof. 3 m.	°C	Temperatura	2662	162	94.26%	14.9	29.8
ZN_SOL	Zinc Total	mg/L	Sales	2663	161	94.30%	0.01021	4.2
MG_TOT	Magnesio Total	mg/L	Mineral	2663	161	94.30%	5.5	133.7
BENCENO	Benceno	mg/L	Química	2664	160	94.33%	0.001	0.008
TETRACL_CARB	Tetracloruro de Carbono	mg/L	Química	2664	160	94.33%	0	<0.0001
TRICLOROFLUOROMETANO	Triclorofluorometano	mg/L	Química	2664	160	94.33%	0	<0.2
TRANS_1,2_DICLOROETILENO	1,2-Dicloroetano	mg/L	Química	2665	159	94.37%	0	<0.001
CL_METILENO	Cloruro de Metileno	mg/L	Química	2665	159	94.37%	0.004	0.04
METILTERBUTILETER_MTBE_	Metilterbutiléter (MTBE)	mg/L	Química	2665	159	94.37%	0	0.24
TETRACLOROETILENO_	Tetracloroeteno (Tetracloroetileno)	mg/L	Química	2665	159	94.37%	0.0012	0.24
TEMP_AGUA_MED	Temperatura agua media	°C	Temperatura	2665	159	94.37%	15	27.5
CONDUC_CAMPO_FON	Conductividad eléctrica fondo	µS/cm	Física	2666	158	94.41%	95	1316
1,1,1,2_TTCETA	1,1,1,2-Tetracloroetano	mg/L	Química	2666	158	94.41%	0	<0.202
1,1,2_TCETA	1,1,2-Tricloroetano	mg/L	Química	2666	158	94.41%	0	<0.001
1,1_DICLETILEN	1,1-Dicloroetileno	mg/L	Química	2666	158	94.41%	0	<0.00015
1,2,4_TRICLBENCEN	1,2,4-Triclorobenceno	mg/L	Química	2666	158	94.41%	0	<0.2
1,2_DIBR_3_CLPROPAN	1,2-Dibromo-3-Cloropropano (DBCP)	mg/L	Química	2666	158	94.41%	0	<0.2

CLAVE PARÁMETRO	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
1,2_DICLPROPAN	1,2-Dicloropropano	mg/L	Química	2666	158	94.41%	0	<0.2
2_CEVE	2-Cloroetilvinil Éter	mg/L	Química	2666	158	94.41%	0	<0.2
ACRILONITRILO	Acrilonitrilo	mg/L	Química	2666	158	94.41%	0	<0.2
BR_CLMET	Bromoclorometano	mg/L	Química	2666	158	94.41%	0	0.19
BROMOMETANO	Bromometano	mg/L	Química	2666	158	94.41%	0	<0.2
CL_BENC	Clorobenceno	mg/L	Química	2666	158	94.41%	0	<0.0002
CLOROMETANO	Clorometano	mg/L	Química	2666	158	94.41%	0	<0.001
ESTIRENO	Estireno	mg/L	Química	2666	158	94.41%	0	<0.2
TRANS_1,2_DICLOROETILENO	Trans-1,2-Dicloroetileno	mg/L	Química	2666	158	94.41%	0.00104	0.00163
TRICLOROETILENO	Tricloroetileno	mg/L	Química	2666	158	94.41%	0.001	0.05
POT_REDOX_CAMPO_FON	Potencial REDOX fondo	mV	Química	2666	158	94.41%	-308	388
BROMOF	Bromoformo	mg/L	Química	2666	158	94.41%	0.001	0.01
1,1,2,2_TETRACLETAN	1,1,2,2-Tetracloroetano	mg/L	Química	2667	157	94.44%	0	<0.2
CIS_1,2_DICLETEN	Cis-1,2-Dicloroetileno	mg/L	Química	2668	156	94.48%	0	<0.21
BR_DICLMET	Bromodiclorometano	mg/L	Química	2668	156	94.48%	0	<0.21
CLOROFORMO	Cloroformo	mg/L	Química	2668	156	94.48%	0.0011	0.005
DIBR_CLMET	Dibromoclorometano	mg/L	Química	2668	156	94.48%	0	<0.2
THMs	Trihalometanos Totales	mg/L	Química	2668	156	94.48%	0.002	<0.44
TEMP_AGUA_4	Temperatura Prof. 4 m.	°C	Temperatura	2668	156	94.48%	14.5	29.2
M,P_XILENO	m,p-Xilenos	mg/L	Química	2670	154	94.55%	<0.002	0.01
O_XILENO	o-Xileno	mg/L	Química	2670	154	94.55%	<0.001	0.003
XILEN_ISOMEROS	Mezcla de Isómeros de Xilenos	mg/L	Química	2670	154	94.55%	<0.003	0.1
CLORURO_DE_VINILO	Cloruro de Vinilo	mg/L	Química	2671	153	94.58%	0	<0.001
ETILBENCENO	Etilbenceno	mg/L	Química	2671	153	94.58%	0	<0.001
TOLUENO	Tolueno	mg/L	Química	2671	153	94.58%	0	<0.001
1,4_DICLBEN	1,4-Diclorobenceno	mg/L	Química	2671	153	94.58%	0	<0.001
ACROLEINA	Acroleina	mg/L	Química	2672	152	94.62%	0	<0.0002
1,2_DICLBENC	1,2-Diclorobenceno	mg/L	Química	2672	152	94.62%	0	<0.001
1,3_DICLBEN	1,3-Diclorobenceno	mg/L	Química	2672	152	94.62%	0	<0.001
DICLOROBENCENOS	Diclorobencenos (Mezcla de Isomeros)	mg/L	Química	2672	152	94.62%	0	<0.003

CLAVE PARÁMETRO	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
TEMP_AGUA_FON	Temperatura agua fondo	°C	Temperatura	2673	151	94.65%	13.21	24.4
TOX_FIS_FON_5_UT	Toxicidad Vibrio fischeri Fondo, 5min	UT	Bacteriano	2676	148	94.76%	<1	<1
pH_CAMPO_MED	Potencial de Hidrógeno media	UpH	Mineral	2677	147	94.79%	6.1	9.8
OD_%_MED	Oxígeno Disuelto media_%	% Saturación	Química	2679	145	94.87%	10	113.2
OD_mg/L_MED	Oxígeno Disuelto media	mg/L	Química	2679	145	94.87%	1	10
TEMP_AGUA_5	Temperatura Prof. 5 m.	°C	Temperatura	2681	143	94.94%	14.2	28.9
pH_CAMPO_FON	Potencial de Hidrógeno fondo	UpH	Mineral	2686	138	95.11%	6.2	9.9
OD_%_FON	Oxígeno Disuelto fondo_%	% Saturación	Química	2689	135	95.22%	10	100
OD_mg/L_FON	Oxígeno Disuelto fondo	mg/L	Química	2689	135	95.22%	1	10
TEMP_AGUA_6	Temperatura Prof. 6 m.	°C	Temperatura	2690	134	95.25%	15	28.2
TEMP_AGUA_7	Temperatura Prof. 7 m.	°C	Temperatura	2703	121	95.72%	14	27.6
TEMP_AGUA_8	Temperatura Prof. 8 m.	°C	Temperatura	2713	111	96.07%	13.9	27.5
TEMP_AGUA_9	Temperatura Prof. 9 m.	°C	Temperatura	2735	89	96.85%	13.8	27.4
TEMP_AGUA_10	Temperatura Prof. 10 m.	°C	Temperatura	2744	80	97.17%	13.8	27.3
TEMP_AGUA_11	Temperatura Prof. 11 m.	°C	Temperatura	2755	69	97.56%	13.52	27.3
TEMP_AGUA_12	Temperatura Prof. 12 m.	°C	Temperatura	2759	65	97.70%	13.52	27.3
TOX_FIS_SUP_30_UT	Toxicidad Vibrio fischeri Superficial, 30min	UT	Bacteriano	2765	59	97.91%	<1	<1
TEMP_AGUA_13	Temperatura Prof. 13 m.	°C	Temperatura	2773	51	98.19%	14.84	26.5
TOX_FIS_FON_30_UT	Toxicidad Vibrio fischeri Fondo, 30min	UT	Bacteriano	2777	47	98.34%	<1	<1
TEMP_AGUA_14	Temperatura Prof. 14 m.	°C	Temperatura	2780	44	98.44%	14.72	26.4
AL_TOT	Aluminio Total	mg/L	Química	2789	35	98.76%	0.015	15.93
CU_TOT	Cobre Total	mg/L	Metálica	2789	35	98.76%	0.002	0.07
TEMP_AGUA_15	Temperatura Prof. 15 m.	°C	Temperatura	2789	35	98.76%	14.71	26.1
AS_SOL	Arsénico Soluble	mg/L	Natural	2793	31	98.90%	0	0.01
CD_SOL	Cadmio Soluble	mg/L	Química	2793	31	98.90%	0	0.003
CR_SOL	Cromo Soluble	mg/L	Metálica	2793	31	98.90%	0.0051	0.04
FE_SOL	Fierro Soluble	mg/L	Metálica	2793	31	98.90%	0.02	0.4
MN_SOL	Manganeso Soluble	mg/L	Metálica	2793	31	98.90%	0.004	0.5
HG_SOL	Mercurio Soluble	mg/L	Natural	2793	31	98.90%	0.00063	0.0015
PB_SOL	Plomo Soluble	mg/L	Metálica	2793	31	98.90%	0.00713	0.01
ZN_SOL	Zinc Soluble	mg/L	Química	2793	31	98.90%	0.01	0.05
NI_SOL	Níquel Soluble	mg/L	Química	2793	31	98.90%	0.0041	0.2
SO4_SOL	Sulfatos Solubles	mg/L	Química	2794	30	98.94%	24.2	730
FLUORUROS_SOL	Fluoruros Solubles	mg/L	Natural	2795	29	98.97%	0.245	0.788
CA_SOL	Calcio Soluble	mg/L	Química y Biológica	2795	29	98.97%	8.8	285.4

CLAVE PARÁMETRO	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
MG_SOL	Magnesio Soluble	mg/L	Mineral	2795	29	98.97%	8.7	83.7
K_SOL	Potasio Soluble	mg/L	Mineral	2795	29	98.97%	2.6	23.05
NA_SOL	Sodio Soluble	mg/L	Sales	2795	29	98.97%	13.7	289
CLORUROS_SOL	Cloruros Solubles	mg/L	Natural	2796	28	99.01%	0.25	29.5
TEMP_AGUA_16	Temperatura Prof. 16 m.	°C	Temperatura	2797	27	99.04%	15.2	23.6
BA_TOT	Bario Total	mg/L	Metálica	2800	24	99.15%	0.01	0.28
AL_SOL	Aluminio Soluble	mg/L	Metálica	2800	24	99.15%	0.015	11.19
CU_SOL	Cobre Soluble	mg/L	Metálica	2800	24	99.15%	0.0015	0.03
TALIO	Talio Total	mg/L	Metálica	2801	23	99.19%	0	<0.005
TALIO_SOL	Talio Soluble	mg/L	Metálica	2801	23	99.19%	0	<0.005
TEMP_AGUA_17	Temperatura Prof. 17 m.	°C	Temperatura	2803	21	99.26%	15.28	23.4
TEMP_AGUA_18	Temperatura Prof. 18 m.	°C	Temperatura	2804	20	99.29%	15.26	21.48
TEMP_AGUA_19	Temperatura Prof. 19 m.	°C	Temperatura	2808	16	99.43%	15.4	21.46
TEMP_AGUA_20	Temperatura Prof. 20 m.	°C	Temperatura	2809	15	99.47%	15.4	21.44
TEMP_AGUA_21	Temperatura Prof. 21 m.	°C	Temperatura	2812	12	99.58%	15.3	21.4
TEMP_AGUA_22	Temperatura Prof. 22 m.	°C	Temperatura	2812	12	99.58%	15.3	21.37
SULFUROS_TOT	Sulfuros (H2S)	mg/L	Metálica y Mineral	2813	11	99.61%	0.1	14.4
TEMP_AGUA_23	Temperatura Prof. 23 m.	°C	Temperatura	2813	11	99.61%	15.3	21.36
TEMP_AGUA_24	Temperatura Prof. 24 m.	°C	Temperatura	2814	10	99.65%	15.3	21.32
TEMP_AGUA_25	Temperatura Prof. 25 m.	°C	Temperatura	2814	10	99.65%	15.3	21.3
MAT_FLOTANTE	MATERIA FLOTANTE	AUSENTE/PRESENTE	Material particulado	2816	8	99.72%	USENTE	PRESENTE
TEMP_AGUA_26	Temperatura Prof. 26 m.	°C	Temperatura	2817	7	99.75%	17.9	21.3
SULFUROS_SOL	Sulfuros Solubles	mg/L	Metálica y Mineral	2817	7	99.75%	0	<0.05
GRAS_ACEI	Grasas y Aceites	mg/L	Materias Grasas	2818	6	99.79%	0	5.5
SOLID_SED	Sólidos Sedimentables	mL/L	Material particulado	2818	6	99.79%	0	<0.1
TEMP_AGUA_27	Temperatura Prof. 27 m.	°C	Temperatura	2818	6	99.79%	17.8	21.21
TEMP_AGUA_28	Temperatura Prof. 28 m.	°C	Temperatura	2819	5	99.82%	17.6	21.2
1,2_DIFENILHIDRACINA	1,2-Difenilhidracina	mg/L	Química	2820	4	99.86%	0	<0.2
1_C_NAF	1-Cloronaftaleno	mg/L	Química	2820	4	99.86%	0	<0.2
2,3_DICLOROFENOL_DCF	2,3-Diclorofenol	mg/L	Química	2820	4	99.86%	0	<0.2
2,4,5_TRICLOROFENOL	2,4,5-Triclorofenol	mg/L	Química	2820	4	99.86%	0	<0.2
2,4,6_TRICLFENOL	2,4,6-Triclorofenol	mg/L	Química	2820	4	99.86%	0	<0.2

CLAVE	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
2,4_DICLFENOL	2,4-Diclorofenol	mg/L	Química	2820	4	99.86%	0	<0.2
2,4_DIMETILFEN	2,4-Dimetilfenol	mg/L	Química	2820	4	99.86%	0	<0.2
2,4_DINITOL	2,4-Dinitrotolueno	mg/L	Química	2820	4	99.86%	0	<0.001
2,4_DNF	2,4-Dinitrofenol	mg/L	Química	2820	4	99.86%	0	<0.2
2,6_DINTOL	2,6-Dinitrotolueno	mg/L	Química	2820	4	99.86%	0	<0.2
2_CF	2-Clorofenol	mg/L	Química	2820	4	99.86%	0	<0.2
2_CLNAFT	2-Cloronaftaleno	mg/L	Química	2820	4	99.86%	0	<0.2
2_NIFENOL	2-Nitrofenol	mg/L	Química	2820	4	99.86%	0	<0.2
4_BRFENFENETER	4-Bromofenil Fenil Éter	mg/L	Química	2820	4	99.86%	0	<0.2
4_NITROFENOL	4-Nitrofenol	mg/L	Química	2820	4	99.86%	0	<0.001
ACENAF	Acenafteno	mg/L	Química	2820	4	99.86%	0	<0.2
ACENAFTIL	Acenaftileno	mg/L	Química	2820	4	99.86%	0	<0.2
ANTRA	Antraceno	mg/L	Química	2820	4	99.86%	0	<0.2
BEN_A_ANT	Benzo(A)Antraceno	mg/L	Química	2820	4	99.86%	0	<0.2
BEN_A_PIR	Benzo(A)Pireno	mg/L	Química	2820	4	99.86%	0	<0.046
BEN_B_FANT	Benzo(B)Fluoranteno	mg/L	Química	2820	4	99.86%	0	<0.2
BEN_G,H,I_PER	Benzo(GHI)Perileno	mg/L	Química	2820	4	99.86%	0	<0.2
BEN_K_FANT	Benzo(K)Fluoranteno	mg/L	Química	2820	4	99.86%	0	<0.2
CRISENO	Criseno	mg/L	Química	2820	4	99.86%	0	<0.2
DIBEN_ANT	Dibenzo(AH)Antraceno	mg/L	Química	2820	4	99.86%	0	<0.2
FENANTRENO	Fenantreno	mg/L	Química	2820	4	99.86%	0	<0.2
FLUORANTENO	Fluoranteno	mg/L	Química	2820	4	99.86%	0	<0.2
FLUORENO	Fluoreno	mg/L	Química	2820	4	99.86%	0	<0.2
IND_PIR	Indeno (1,2,3-CD) Pireno	mg/L	Química	2820	4	99.86%	0	<0.00992
NAFTALENO	Naftaleno	mg/L	Bioquímico	2820	4	99.86%	0	<0.046
PIRENO	Pireno	mg/L	Química	2820	4	99.86%	0	<0.2
HAPS	Total de hidrocarburos poliaromáticos	mg/L	Química	2820	4	99.86%	0	<2.7
DUR_Mg	Dureza Magnesio	mg/L	Mineral	2823	1	99.96%	87.06	87.06
DUR_Ca	Dureza Calcio	mg/L	Mineral	2823	1	99.96%	71.93	71.93
B2_ETILHEXIL_FTALATO	Bis (2-Etilhexil) Ftalato	mg/L	Bioquímico	2820	4	99.86%	0.001	3.07
BENCIDINA	Bencidina	mg/L	Química	2820	4	99.86%	0	<0.2
BIS2_CLIOISOPET	Bis (2-Cloroisopropil) Éter	mg/L	Química	2820	4	99.86%	0	<0.2
DI_2_ETILEXIL_ADIPATO	Di (2 Etilexil) Adipato	mg/L	Química	2820	4	99.86%	0	<0.2
DIET_FTAL	Dietilftalato	mg/L	Química	2820	4	99.86%	0	0.0001

CLAVE	NOMBRE PARÁMETRO	UNIDAD DE MEDIDA	TIPO DE DATO	NÚMERO REGISTROS NULOS	REGISTROS LLENOS	PORCENTAJE PERDIDAS	RANGO MÍNIMO	RANGO MÁXIMO
DIMETIL_FTAL	Dimetilftalato	mg/L	Química	2820	4	99.86%	0	<0.2
DINITRO_O_CRESOL	Dinitro-o-Cresol	mg/L	Química	2820	4	99.86%	0	<0.2
FENOL	Fenol	mg/L	Química	2820	4	99.86%	0	<0.2
FENOLES_TOT	Fenoles Totales	mg/L	Química	2820	4	99.86%	0	<0.03
H_HELMINTOS	Huevos de Helminto	No. de huevos / 5 L	Bioquímico	2820	4	99.86%	0	0
HEXACL_BUT	Hexaclorobutadieno	mg/L	Química	2820	4	99.86%	0	<0.2
HEXACL_CICLOPEN	Hexaclorociclopentadieno	mg/L	Química	2820	4	99.86%	0	<0.2
HEXACL_ETANO	Hexacloroetano	mg/L	Química	2820	4	99.86%	0	<0.001
ISO_FOR	Isófona	mg/L	Química	2820	4	99.86%	0	<0.2
N_NITROS_DIMETILA	N-Nitrosodimetilamina	mg/L	Química	2820	4	99.86%	0	<0.2
N_NITROSODIFENILAMINA	N-Nitrosodi-N-Propilamina	mg/L	Química	2820	4	99.86%	0	<0.24
N_NITROSODIFENILAMINA	N-Nitrosodifenilamina	mg/L	Química	2820	4	99.86%	0	<0.2
NITROBENCENO	Nitrobenceno	mg/L	Química	2820	4	99.86%	0	<0.001
PCBEN	Pentaclorobenceno	mg/L	Química	2820	4	99.86%	0	<0.2
PENTACLOROFENOL	Pentaclorofenol	mg/L	Química	2820	4	99.86%	0	<0.2
SE_TOT	Selenio Total	mg/L	Mineral	2820	4	99.86%	0	<0.02
TEMP_AGUA_29	Temperatura Prof. 29 m.	°C	Temperatura	2821	3	99.89%	17.5	21.2
SB_TOT	Antimonio Total	mg/L	Metálico	2823	1	99.96%	0	0.003
TEMP_AGUA_30	Temperatura Prof. 30 m.	°C	Temperatura	2823	1	99.96%	0	17.39
TEMP_AGUA_31	Temperatura Prof. 31 m.	°C	Temperatura	2823	1	99.96%	0	17.3
TEMP_AGUA_32	Temperatura Prof. 32 m.	°C	Temperatura	2823	1	99.96%	0	17.2
TEMP_AGUA_33	Temperatura Prof. 33 m.	°C	Temperatura	2823	1	99.96%	0	17
TEMP_AGUA_34	Temperatura Prof. 34 m.	°C	Temperatura	2823	1	99.96%	0	17
BA_SOL	Bario Soluble	mg/L	Metálico	2823	1	99.96%	0	0.2
SB_SOL	Antimonio Soluble	mg/L	Metálico	2823	1	99.96%	0	<0.003

Anexo Tabla 2. *Todos los contaminantes*

Experimento	Parámetro	Número de perdidas	Porcentaje de perdidas
Parámetros Físicos	TEMP_AMB	0	0.00%
	TEMP_AGUA	0	0.00%
	TEMP_AGUA_SUP	2322	100.00%
	TEMP_AGUA_MED	2322	100.00%
	TEMP_AGUA_FON	2322	100.00%
	TEMP_AGUA_0.5	2322	100.00%
	TEMP_AGUA_1	2322	100.00%
	TEMP_AGUA_2	2322	100.00%
	TEMP_AGUA_3	2322	100.00%
	TEMP_AGUA_4	2322	100.00%
	TEMP_AGUA_5	2322	100.00%
	TEMP_AGUA_6	2322	100.00%
	TEMP_AGUA_7	2322	100.00%
	TEMP_AGUA_8	2322	100.00%
	TEMP_AGUA_9	2322	100.00%
	TEMP_AGUA_10	2322	100.00%
	TEMP_AGUA_11	2322	100.00%
	TEMP_AGUA_12	2322	100.00%
	TEMP_AGUA_13	2322	100.00%
	TEMP_AGUA_14	2322	100.00%
	TEMP_AGUA_15	2322	100.00%
	TEMP_AGUA_16	2322	100.00%
	TEMP_AGUA_17	2322	100.00%
	TEMP_AGUA_18	2322	100.00%
	TEMP_AGUA_19	2322	100.00%
	TEMP_AGUA_20	2322	100.00%
	TEMP_AGUA_21	2322	100.00%
	TEMP_AGUA_22	2322	100.00%
	TEMP_AGUA_23	2322	100.00%
	TEMP_AGUA_24	2322	100.00%
	TEMP_AGUA_25	2322	100.00%
	TEMP_AGUA_26	2322	100.00%
	TEMP_AGUA_27	2322	100.00%
	TEMP_AGUA_28	2322	100.00%
	TEMP_AGUA_29	2322	100.00%
	TEMP_AGUA_30	2322	100.00%

Experimento	Parámetro	Número de perdidas	Porcentaje de perdidas
Parámetros Físicos	TEMP_AGUA_31	2322	100.00%
	TEMP_AGUA_32	2322	100.00%
	TEMP_AGUA_33	2322	100.00%
	TEMP_AGUA_34	2322	100.00%
	TURBIEDAD	0	0.00%
	CONDUC_CAMPO	0	0.00%
	CONDUC_CAMPO_SUP	2322	100.00%
	CONDUC_CAMPO_MED	2322	100.00%
	CONDUC_CAMPO_FON	2322	100.00%
	CA_TOT	2322	100.00%
	CA_SOL	2322	100.00%
	DUR_Ca	2322	100.00%
	MG_TOT	2322	100.00%
	MG_SOL	2322	100.00%
	DUR_Mg	2322	100.00%
	NA_SOL	2322	100.00%
	NA_TOT	2322	100.00%
	K_TOT	2322	100.00%
	K_SOL	2322	100.00%
	CO3	2322	100.00%
	HCO3	2322	100.00%
	SO4_TOT	2317	99.78%
	SO4_SOL	2321	99.96%
Parámetros Químicos	pH_CAMPO	1	0.04%
	pH_CAMPO_SUP	2322	100.00%
	pH_CAMPO_MED	2322	100.00%
	pH_CAMPO_FON	2322	100.00%
	DUR_TOT	15	0.65%
	DUR_Ca	2322	100.00%
	DUR_Mg	2322	100.00%
	DBO_SOL	84	3.62%
	DBO_TOT	10	0.43%
	DQO_SOL	98	4.22%
	DQO_TOT	27	1.16%
	OD_%	0	0.00%
	OD_mg/L	0	0.00%

Experimento	Parámetro	Número de perdidas	Porcentaje de perdidas
Parámetros Químicos	OD_%_SUP	2322	100.00%
	OD_mg/L_SUP	2322	100.00%
	OD_%_MED	2322	100.00%
	OD_mg/L_MED	2322	100.00%
	OD_%_FON	2322	100.00%
	OD_mg/L_FON	2322	100.00%
	ALC_FEN	2322	100.00%
	ALC_TOT	2322	100.00%
Parámetros Biológicos	COLI_FEC	2	0.09%
	COLI_TOT	63	2.71%
	E_COLI	201	7.66%
Parámetros Aniones	CO3	2322	100.00%
	HCO3	2322	100.00%
	CL_METILENO	2319	99.87%
	CLORURO_DE_VINILO	2319	99.87%
	CLORUROS_TOT	2318	99.83%
	TETRACL_CARB	2322	100.00%
	CLORUROS_SOL	2322	100.00%
	FLUORUROS_SOL	2322	100.00%
	FLUORUROS_TOT	2322	100.00%
	N_NO3	0	0.00%
	SO4_TOT	2317	99.78%
	SO4_SOL	2321	99.96%
Parámetros Cationes	CA_TOT	2322	100.00%
	CA_SOL	2322	100.00%
	DUR_Ca	2322	100.00%
	MG_TOT	2322	100.00%
	MG_SOL	2322	100.00%
	DUR_Mg	2322	100.00%
	NA_SOL	2322	100.00%
	NA_TOT	2322	100.00%
	K_TOT	2322	100.00%
	K_SOL	2322	100.00%

Experimento	Parámetro	Número de perdidas	Porcentaje de perdidas
Parámetros de Metales Pesados	AS_SOL	2321	99.96%
	AS_TOT	13	0.56%
	AL_SOL	2321	99.96%
	AL_TOT	2317	99.78%
	CR_TOT	18	0.78%
	CR_SOL	2321	99.96%
	CU_TOT	2317	99.78%
	CU_SOL	2322	100.00%
	PB_TOT	15	0.65%
	PB_SOL	2321	99.96%
	MG_TOT	2322	100.00%
	MG_SOL	2322	100.00%
	DUR_Mg	2322	100.00%
	CD_TOT	17	0.73%
	CD_SOL	2321	99.96%
	NI_SOL	2321	99.96%
	NI_TOT	14	0.60%
	ZN_TOT	2318	99.83%
	ZN_SOL	2321	99.96%
	HG_SOL	2321	99.96%
	HG_TOT	43	1.85%

Anexo Tabla 3. *Experimentos con contaminantes y respectivos porcentajes de perdida*

Casos	Contaminantes seleccionados	Descripción	Clave de contaminante
1	Conductividad eléctrica, Nitrógeno de Nitratos, Oxígeno Disuelto, Temperatura Agua, Temperatura Ambiente, Turbiedad	Parámetros físicos es una variable numérica que describe la calidad que es conformada por elementos como color, olor, sabor, turbidez, entre otros	CONDUC_CAMPO, N_NO3, OD_%, TEMP_AGUA, TEMPO_AMB, TURBIEDAD
2	Demanda Bioquímica de Oxígeno Soluble, Demanda Bioquímica de Oxígeno Total, Demanda Química de Oxígeno Soluble, Demanda Química de Oxígeno Total, Dureza Total, Saturación de Oxígeno Disuelto, Oxígeno Disuelto, Potencia de Hidrógeno.	Parámetros químicos se analiza la distribución de las sustancias individuales que constituyen la estructura del agua	DBO_SOL, DBO_TOT, DQO_SOL, DQO_TOT, DUR_TOT, OD_%, pH_CAMPO
3	Coliformes Fecales, Coliformes Totales, Saturación de Oxígeno Disuelto	Parámetros biológicos encargada de identificar su capacidad de transportar distintas sustancias, disolver otras y mantener su estructura original, además de analizar los elementos orgánicos e inorgánicos	COLI_FEC, COLI_TOT, OD_%, E_COLI
4	Nitrógeno de Nitratos, Saturación de Oxígeno Disuelto	Parámetros Aniones analiza la energía neutral o conductividad en el agua	N_NO3, OD_%
5	Calcio Soluble, Calcio Total, Dureza Calcio, Dureza Magnesio, Potasio Soluble, Potasio Total, Magnesio Soluble, Magnesio Total, Sodio Soluble, Sodio Total, Saturación de Oxígeno Disuelto	Parámetros Cationes es un conjunto de sales y sedimentos que aumentan la conductividad en el agua	CA_SOL, CA_TOT, DUR_Ca, DUR_Mg, K_SOL, K_TOT, MG_SOL, MG_TOT, NA_SOL, NA_TOT, OD_%
6	Arsénico Total, Cadmio Total, Cromo Total, Mercurio Total, Níquel Total, Saturación de Oxígeno Disuelto, Plomo Total	Parámetros de Metales Pesados se analizan los elementos metálicos cuya densidad es por lo menos cinco veces mayor que la del agua	AS_TOT, CD_TOT, CR_TOT, HG_TOT, NI_TOT, OD_%, PB_TOT

Casos	Contaminantes seleccionados	Descripción	Clave de contaminante
7	Arsénico Total, Cadmio Total, Coliformes Fecales, Coliformes Totales, Conductividad Eléctrica, Cromo Total, Demanda Bioquímica de Oxígeno Soluble, Demanda Bioquímica de Oxígeno Total, Demanda Química de Oxígeno Soluble, Demanda Química Total, Dureza Total, Mercurio Total, Nitrógeno de Nitratos, Níquel Total, Saturación de Oxígeno Disuelto, Plomo Total, Potencia de Hidrógeno, Temperatura de Agua, Temperatura Ambiente, Turbiedad	Estos experimentos son una combinación de los anteriores parámetros analizados	AS_TOT, CD_TOT, COLI_FEC, COLI_TOT, CONDUC_CAMPO, CR_TOT, DBO_SOL, DBO_TOT, DQO_SOL, DQO_TOT, DUR_TOT, HG_TOT, N_NO3, NI_TOT, OD_%, PB_TOT, pH_CAMPO, TEMP_AGUA, TEMP_AMB, TURBIEDAD
8	Demanda Bioquímica de Oxígeno Soluble, Demanda Bioquímica de Oxígeno Total, Demanda Química de Oxígeno Soluble, Demanda Química de Oxígeno Total, Saturación de Oxígeno Disuelto, Oxígeno Disuelto, Potencial de Hidrógeno		DBO_SOL, DBO_TOT, DQO_SOL, DQO_TOT, DUR_TOT, N_NO3, OD_%, pH_CAMPO
9	Coliformes Fecales, Coliformes Totales, Conductividad Eléctrica, Demanda Bioquímica de Oxígeno Soluble, Demanda Bioquímica de Oxígeno Total, Demanda Química de Oxígeno Soluble, Demanda Química de Oxígeno Total, Dureza Total, Saturación de Oxígeno Disuelto, Oxígeno Disuelto, Temperatura del agua, Temperatura Ambiente		COLI_FEC, COLI_TOT, CONDUC_CAMPO, DBO_SOL, DBO_TOT, DQO_SOL, DQO_TOT, DUR_TOT, OD_%, TEMP_AGUA, TEMP_AMB

Anexo Tabla 4. *Experimentos planteados para estimar la calidad del agua*