Proyecto Integrador I

2020-1

Christian Flor - A00355624, César Canales - A00345026, Carlos Restrepo - A00355028,
Daniel Fernández - A00354694, Felipe Sanchez - A00355727

Gas impact analyzer on crops in Valle del Cauca

(GIAOC-Valle del Cauca)

Fase 1. Identificación del problema.

Descripción del contexto problemático (causas y síntomas):

El Instituto de Hidrología, Meteorología y Estudios Ambientales IDEAM es una entidad del gobierno Colombiano dependiente del Ministerio de Ambiente y Desarrollo Sostenible. Se encarga del manejo de la información científica, hidrológica, meteorológica y todo lo relacionado con el medio ambiente en Colombia.

Mediante el portal web de Datos Abiertos del Ministerio de Tecnologías de la Información y las Comunicaciones www.datos.gov.co, el IDEAM proporcionó datos de Calidad de aire (Inmisión) y variables climatológicas, reportadas por las Autoridades Ambientales al SISAIRE (Subsistema de Información de la Calidad del Aire) durante los años 2011 - 2017.

El portal web tiene como función principal publicar de manera unificada, todos los datos producidos por las entidades públicas de Colombia, en formato abierto, con el fin de que éstos puedan ser usados de forma libre y sin restricciones por cualquier persona para desarrollar aplicaciones o servicios de valor agregado, hacer análisis e investigación, ejercer labores de control o para cualquier tipo de actividad comercial o no comercial.

Por otro lado, en el país la contaminación atmosférica se ha constituido en uno de los principales problemas ambientales; el deterioro de la calidad del aire ha propiciado que se incrementen los efectos negativos sobre la salud humana y el medio ambiente.

Identificación de los síntomas y necesidades de la situación problemática

- Se requiere plantear una manera de visualizar los datos más relevantes presentados en la base de datos a través de los distintos años que se presentan.
- Se requiere hacer un análisis de los datos para así obtener la información de mayor interés para el usuario.

Definición del problema

Colombia ha realizado esfuerzos para lograr la ejecución y cumplimiento de las estrategias dirigidas a prevenir y controlar la contaminación del aire (Politica de prevencion y control 2010).Por lo cual el principal objetivo es lograr informar a la población Colombiana que tan contaminado está el aire.

Estudios han mostrado que las plantas necesitan de ciertos compuestos químicos para su crecimiento y síntesis de alimentos. Muchos de estos compuestos se encuentran en el aire y son llevados a la tierra mediante precipitación. A pesar de que la mayoría de los gases son aprovechados por las plantas para cumplir con sus funciones autótrofas, en una cantidad muy elevada estos pueden inhibir su crecimiento. Una cantidad muy baja de ellos puede ser insuficiente. Este es un inconveniente que muchos agricultores pueden estar sufriendo a causa de la falta de información disponible acerca del tema o de la poca disponibilidad de ella, haciendo que pierdan grandes sumas de dinero tratando de hacer

prosperar un cultivo en una zona no apta para ello. Varios de estos temas han sido tratados en artículos de revistas reconocidas cómo la *National Geographic.*

Fase 2. Recopilación de la información necesaria.

Para tener claridad total en los conceptos involucrados, se realiza una búsqueda de las definiciones de los términos relacionados con el problema planteado.

Opinión profesional

Además de las búsquedas web se le preguntó a un profesor del curso de Programación en Red de la Universidad ICESI acerca de qué recomendaciones nos podía dar para el tratamiento de grandes volúmenes de datos. También se tuvo la oportunidad de conversar con un profesor de Inferencia Estadística de la misma universidad, lo que ayudó a clarificar muchas de las ideas que se tenían acerca del uso de métodos estadísticos para predecir patrones en los datos. Se consideró hacer uso de series de tiempo para predecir patrones pero se encontró un problema: Para que estas predicciones sean buenas la población debería mantenerse constante, al igual que la cantidad de industrias, el nivel de producción y otros factores que pueden ser de gran impacto en las variables de respuesta a analizar(crecimiento de cultivos).

Análisis estadístico

Se llegó a la conclusión de que dada la naturaleza de los datos la estadística inferencial no sería de gran ayuda al tratar de hacer predicciones significativas y precisas a largo plazo. Por lo tanto se optó por una opción igual de interesante qué se basa en conseguir la información que sea necesaria sobre el crecimiento de cultivos en alguna región del país, analizar la relación entre algunos factores presentes en la base de datos del IDEAM y la

variable de respuesta crecimiento y hacer uso de métodos estadísticos descriptivos para lograr explicar el comportamiento de la variable de respuesta ante ciertas concentraciones de los gases medidos.

Emisión de gases

Según un artículo de la revista *National Geographic,* tres de los tipos de gases de los que tenemos información en la base de datos(SO_2, NO y NO_2) hacen parte de la lluvia ácida. Mediante el proceso de precipitación está llega al suelo y actúa como un disolvente de los nutrientes en la tierra, haciendo para las plantas una tarea difícil el obtener agua y alimento. Otro artículo publicado por la misma revista habla sobre la importancia del ciclo del nitrógeno en las plantas. En específico dice que algunos de los microbios subterráneos tienen los medios bioquímicos para extraer nitrógeno del agua y convertir los nitritos en nitratos beneficiosos para las plantas. Pero cuanto más nitrógeno se les da, menos eficientes son para convertirlo, por lo tanto mucho del nitrógeno en tierra se puede quedar en forma de nitritos que son contraproducentes para las plantas.

Minería de datos:

"La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos." (Microsoft, 2019, párr. 1). En este sentido, se entiende que la minería de datos hace uso de diversas técnicas para encontrar patrones, predecir comportamientos, entre otros, en un conjunto de datos.

JSON

JSON (JavaScript Object Notation) es un formato ligero de intercambio de datos. Es fácil para los humanos leer y escribir. Es fácil para las máquinas analizar y generar. Se basa en

un subconjunto del lenguaje de programación JavaScript Standard ECMA-262 3rd Edition - December 1999. JSON es un formato de texto que es completamente independiente del lenguaje pero utiliza convenciones que son familiares para los programadores de la familia de lenguajes C, incluido C, C ++, C #, Java, JavaScript, Perl, Python y muchos otros. Estas propiedades hacen de JSON un lenguaje ideal para el intercambio de datos.

Referencias.

***GOV.CO*. (2019). DATOS DE CALIDAD DEL AIRE EN COLOMBIA 2011-2017. Recuperado de https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/DATOS-DE-CALIDAD-DEL-AIRE-EN-COLOMBIA-2011-2017/ysq6-ri4e**

***Independent-software*. (2016). GMAP.NET BEGINNERS TUTORIAL MAPS MARKERS, POLYGONS AND ROUTES. Recuperado de http://www.independent-software.com/gmap-net-beginners-tutorial-maps-markers-polygons-routes-updated-for-vs2015-and-gmap1-7.html**

***GOV.CO*. (2019). TÉRMINOS Y CONDICIONES DE USO PORTAL DE DATOS ABIERTOS. Recuperado de https://herramientas.datos.gov.co/es/terms-and-conditions-es**

JSON. (2013). Introducing JSON. *JSON organization*. Recuperado de https://www.json.org/json-en.html

***IDEAM*. (2016). Informe del Estado de la Calidad del Aire en Colombia 2016. Recuperado de http://www.ideam.gov.co/web/contaminacion-y-calidad-ambiental/informes-del-estado-de-la-calidad-del-aire-en-colombia**

***Ministerio de Ambiente, Vivienda y Desarrollo Territorial*. (2010).Política de Prevención y Control de la Contaminación del Aire. Colombia, Bogotá D.C. Recuperado de https://www.minambiente.gov.co/images/AsuntosambientalesySectorialyUrbana/pdf/Polit%C3%ACcas_de_la_Direcci%C3%B3n/Pol%C3%ADtica_de_Prevenci%C3%B3n_y_Control_de_la_Contaminaci%C3%B3n_del_Aire.pdf**

***World Health Organization Europe*. (2016). AirQ+: software tool for health risk assessment of air pollution. Recuperado de http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/activities/airq-software-tool-for-health-risk-assessment-of-air-pollution**

***United States Environmental Protection Agency*. (2015). Air Quality Dispersion Modeling - Preferred and Recommended Models. Recuperado de**

https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models

Addlink Software Científico. (2020). AERMOD View 9.8. Recuperado de https://www.addlink.es/productos/aermod-view

Cambridge Environmental Research Consultants, Environmental software and services. (s.f). ADMS-Urban. Recuperado de https://www.cerc.co.uk/environmental-software/ADMS-Urban-model.html

GOV.CO. (2019). Superficie Sembrada en Hectáreas con Cultivos Transitorios en el Valle del Cauca del año 2000 al 2018. Recuperado de https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Superficie-Sembrada-en-Hect-reas-con-Cultivos-Tran/vs5v-e66i

GOV.CO. (2019). Superficie Cosechada en Hectáreas con Cultivos Transitorios en el Departamento del Valle del Cauca del Año 2000 al 2018. Recuperado de https://www.datos.gov.co/Agricultura-y-Desarrollo-Rural/Superficie-Cosechada-en-Hect-reas-con-Cultivos-Tra/3d2z-wkgw

C. ZIMMER. (2019). Nitrogen as Horse, Earth, and Air. *National Geographic*. Recuperado de

https://www.nationalgeographic.com/science/phenomena/2008/03/21/nitrogen-as-horse-earth-and-air/

C. NUNEZ. (2019). WHAT IS ACID RAIN?. *National Geographic*. Recuperado de https://www.nationalgeographic.com/environment/global-warming/acid-rain/

Microsoft. (2019). Docs Microsoft : Conceptos de minería de datos. Recuperado de https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions

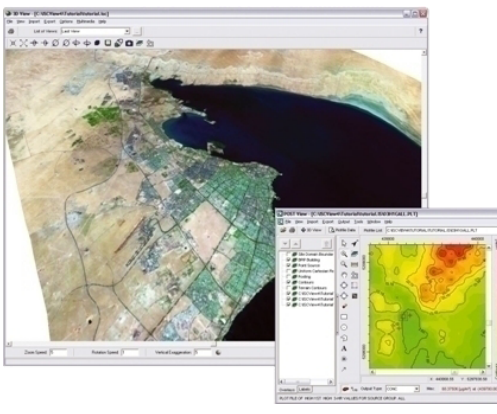
Marco teórico

De acuerdo a la base de datos sugerida por nuestro mentores llamada “DATOS DE CALIDAD DEL AIRE EN COLOMBIA 2011-2017” del IDEAM, esta base tiene 15.657.064 datos y 16 variables. Se realizó un análisis preliminar de las distintas variables en el que se exploran los distintos valores que pueden tomar y en el caso de ser una variable cuantitativa, se buscó cuál era el mínimo y el máximo valor que esta puede tomar. Esta investigación nos permite tener un mayor entendimiento de los datos, por lo consiguiente sabremos qué variables son relevantes y el intervalo de los posibles valores que pueden tomar las variables en esta gran base de datos.

Estado del arte.

AirQ+: software tool for health risk assessment of air pollution

La herramienta de software de la OMS / Europa AirQ + realiza cálculos que permiten cuantificar los efectos sobre la salud de la



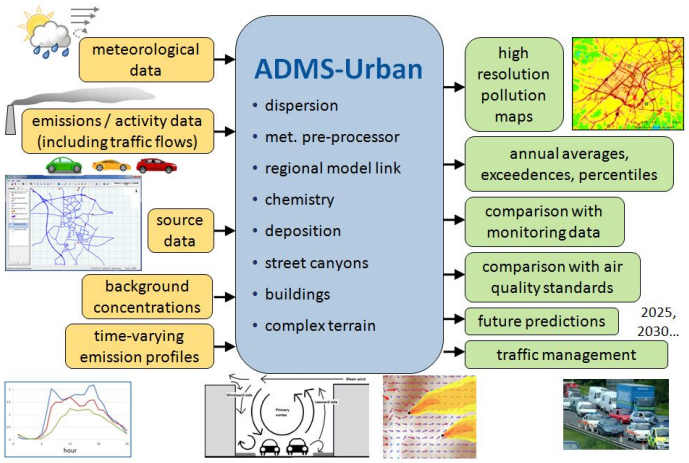
exposición a la contaminación del aire, incluidas las estimaciones de la reducción de la esperanza de vida.

Software EPA Aermod (modelamiento de calidad de aire)

Sirve para modelar emisiones y dispersión de contaminantes atmosféricos y su depositación. También es usado para para la estimación de impactos ambientales de todo tipo de fuentes de emisión atmosférica.

ADMS-Urban

Este programa permite modelizar la calidad del aire en diferentes entornos urbanos, diversos tipos de contaminantes en dispersión a causa de la meteorología y del tránsito, y la previsión de los focos de contaminación según el tipo de calle, su ubicación y su volumen de tránsito.



Fase 3. Búsqueda de soluciones creativas.

- **Ideas para el uso de la base de datos y desarrollo de la aplicación:**
    - Usar los datos de concentración de los diferentes gases en las distintas zonas de Colombia y ver si estos representan un factor decisivo para el crecimiento de las plantas, en especial de ciertos cultivos agrícolas y así desarrollar una aplicación que permita visualizar el crecimiento de dichos cultivos a través del tiempo.

- De acuerdo con la concentración de contaminantes, determinar las zonas de Colombia en donde se debe implementar un programa de reforestación con el objetivo de ayudar a contrarrestar la emisión de estos gases. De esta manera, desarrollar una aplicación que permite visualizar a través de un mapa, las zonas que requieren los programas mencionados.
- Desarrollar una aplicación que haciendo uso de los datos de aire en Colombia determine mediante el uso adicional de la herramienta de Google Maps, los niveles de contaminación en determinadas rutas entre dos puntos. Esto con el fin de informar a las personas sobre las implicaciones en la salud que podrían tener el hecho de usar esa ruta constantemente.

- Empleando los datos de concentración de los diferentes gases y la ubicación de fábricas y talleres industriales, determinar mediante el nivel de contaminación alrededor de estos últimos si están cumpliendo con la emisión máxima permitida por la ley y permitir visualizar esta información a través de una aplicación.

- **Ideas para el procesamiento de la información:**

  Las ideas se harán con base a Data Mining. Las técnicas que hay en el data mining y que se pueden usar para el desarrollo de la aplicación son las siguientes:

  - Mediante el uso de la técnica de clustering, se pueden agrupar los registros con una característica de interés para el análisis que compartan valores comunes. Esto

puede facilitar el análisis estadístico. Por lo cual, es pertinente definir los cultivos que son susceptibles a altas y bajas concentraciones de determinado gas.

- A través del uso de la técnica de Association Rule Learning encontrar combinaciones de factores (concentración de gas, humedad, temperatura, elevación, entre otros) determinantes en el crecimiento de los cultivos pero que aparentemente no tienen relación alguna.
- Usando la técnica de Rule Induction para análisis predictivos y aprovechando la gran cantidad de datos que están disponibles, definir reglas y comportamientos que presentan los cultivos bajo ciertas circunstancias.

Fase 4. Transición de las ideas a diseños preliminares.

Está fase se realiza a los dos objetivos generales presentados anteriormente. Primero se llevará a cabo la transición de ideas a diseños preliminares para el objetivo relacionado con el uso de la base de datos y por último se realizará al objetivo que hace referencia al procesamiento de datos.

1. Uso de la base de datos y desarrollo de la aplicación:

Primeramente, procedemos a descartar ideas inalcanzables.

Alternativa 2: Programa de reforestación

Esta alternativa se descarta ya que no se cuenta con información precisa acerca de cuándo se debe implementar un plan de reforestación dependiendo de la cantidad de contaminación que haya en un área.

Alternativa 3: Rutas a altos niveles de contaminación

Está alternativa puede ser de gran ayuda para las personas que viajan constantemente, pero las ideas de conocer las múltiples rutas que usan los usuarios y para cada ruta determinar su nivel de contaminación presentan un gran reto y mayor conocimiento sobre herramientas relacionadas con el uso de mapas. Por lo cual se descarta esta alternativa.

A continuación, se analizan las ideas prometedoras

Alternativa 1: Crecimiento de las plantas

Esta alternativa se va a implementar utilizando dos bases de datos, una que provee información acerca de la calidad de aire en Colombia a lo largo de los años y la otra que provee información acerca del crecimiento de diversos cultivos a través del tiempo, mediante una correlación se puede ver qué tanto afecta la contaminación al crecimiento de las plantas y mediante la aplicación reportar los hallazgos obtenidos.

Alternativa 4: Emisión de gases de fábricas y de talleres industriales

La idea con esta alternativa es tener un mapa de calor donde se puedan ver todas las fábricas y talleres industriales en Colombia junto con su nivel de contaminación, donde rojo significa muy contaminado y verde significa poco contaminado, y cuando se seleccione una se pueden ver detalles acerca de la emisión de gases de esa empresa en particular.

2. Procesamiento de la información

Procedemos igual que en el ítem anterior, comenzando por descartar las ideas inalcanzables.

Alternativa 2: Técnica de Association Rule Learning

Para realizar estas reglas de asociación existen diversos algoritmos, entre ellos los algoritmos Apriori y Eclat, los cuales son los más conocido. Sin embargo, el algoritmo

Apriori presenta una limitación y es su complejidad temporal, la cual es de tipo exponencial, específicamente 2^n. Aunque el algoritmo Eclat es un poco más eficiente en cuanto al tiempo, este sigue presentando una complejidad muy alta. Por todo esto y teniendo en cuenta que la base de datos con la que se trabaja presenta alrededor de 15 millones de datos, se descarta completamente está alternativa.

Alternativa 3: Técnica de rule induction

Con esta técnica se tenía pensado predecir comportamientos futuros en cuanto a la calidad de aire de acuerdo con los datos presentados en la base de datos. Sin embargo, se investigó con un profesional de estadística y este afirmó de que solo se podrían predecir comportamientos futuros a un plazo no mayor a un día. Dado que la idea que se tenía con esta técnica era conocer comportamientos a seis meses o más, se descarta esta alternativa.

Ahora analizaremos las alternativas que quedaron en consideración:

Alternativa 1: Técnica de clustering

Dado que la técnica de clustering encuentran características comunes en los datos y los agrupa de acuerdo a estos, está técnica se podría implementar para que encuentre qué propiedades en cuanto a la calidad de aire están afectado al crecimiento de las plantas. Esto se puede lograr ejecutando un algoritmo de clustering sobre las tres bases de datos que se tienen. Hay que resaltar, que algoritmos como el K-means (algoritmo propio de clustering) presentan complejidades temporales mucho menores a las de algoritmos de Association Rule Learning.

Fase 5. Evaluación y selección de la mejor solución.

Se deben definir los criterios que permitirán evaluar las alternativas de solución para poder elegir la solución que mejor satisfaga las necesidades del problema. Los criterios que se eligieron en este caso son los que se enumeran a continuación. Al lado de cada uno se ha establecido un valor numérico para establecer un peso.

1. **Uso de la base de datos y desarrollo de la aplicación:**

Criterio A: Facilidad para reunir la información

[3] Fácil, se tiene toda la información necesitada.

[2] Medio, se tiene una información base y la que falta es de fácil acceso.

[1] Difícil, no se tiene información previa y encontrar la información requiere de consultas avanzadas.

Criterio B: Impacto en la sociedad

[4] Muy importante: más del 80% de la población se ve afectada.

[3] Importante: entre un 50% y 80% de la población se ve afectada.

[2] Medianamente importante: entre un 20% y un 50% de la población se ve afectada.

[1] Poco importante: menos del 20% de la población se ve afectada.

Criterio C: Volumen de datos a trabajar

[3] Alto, se requiere trabajar con tres o más bases de datos.

[2] medio, se necesita trabajar como mucho con dos bases de datos.

[1] bajo, solo se requiere trabajar con una base de datos.

Evaluación:

Utilizando los criterios mencionados anteriormente, obtenemos la siguiente tabla al evaluar las alternativas

	Criterio A: Facilidad para reunir la información	Criterio B: Impacto en la sociedad	Criterio C: Volumen de datos a trabajar	Total
Alternativa 1: Crecimiento de las plantas	1	4	3	8
Alternativa 4: Emisión de gases de fábricas y de talleres industriales	2	1	3	6

Selección:

Con base en el análisis y evaluación de las alternativas nos disponemos a trabajar sobre el tema “Crecimiento de las plantas”. Esto más que todo porque es un aspecto de gran impacto en la sociedad, sobre todo en una como Colombia en el que una gran parte de la economía se mueve en el sector agrícola.

2. **Procesamiento de la información**

En vista de que de las tres alternativas que se tenían para el procesamiento de la información usando técnicas de minería de datos, dos fueron descartadas en la fase anterior, se toma como selección a la alternativa uno, que hace uso de la técnica de Clustering. Esto se debe principalmente a que con la técnica de clustering es posible encontrar propiedades de cómo la calidad de aire afecta al crecimiento de las plantas y esto es precisamente lo que se desea lograr con el proyecto.

FASE 6: PREPARACIÓN DE INFORMES Y ESPECIFICACIONES

6.1 Diagrama de clases.

6.2 Diagramas de secuencia.

6.3 Modelo Git.

6.4 Especificación de requerimientos funcionales.

6.5 Diccionario de la base de datos.

FASE 7: IMPLEMENTACIÓN DEL DISEÑO

La solución al problema está en el siguiente repositorio:

https://github.com/ChristianFlor/gas-impact-analyzer-in-crops

SÍNTESIS REFLEXIVA:

Con el desarrollo del método de la ingeniería se pudo llegar a la implementación de la solución que se encontró más eficiente y adaptada al problema presentado y a los recursos que se disponían. Se realizó una contextualización del problema y una primera aproximación a este en las primeras fases. A continuación de esto se usó la técnica de lluvia de ideas para buscar ideas

para el desarrollo de la aplicación y procesamiento de la información que resolvieran el problema tratado. Estas ideas se tuvieron que analizar y evaluar con criterios de calificación. Después de elegir las dos ideas (el desarrollo y el procesamiento de información) más factibles, se prosiguió a implementar estas mismas., Y finalmente a probarlas con pruebas unitarias. A través de este trabajo realizado, se percató que las herramienta como Soda permite una gran flexibilidad y facilidad de uso para consultar datos. LiveCharts nos brindó la posibilidad de graficar grandes volúmenes de datos con distintos tipos de gráficas. Así mismo, GIAC probó ser una manera muy intuitiva de resolver el problema que se planteaba. En efecto, GIAC es un programa intuitivo, "fácil" de usar, en el que se visualizan los datos de cómo afecta la calidad de aire en el crecimiento de las plantas a través de los años en el Valle del cauca.