Análisis de la Calidad del Aire
Segunda Entrega del Proyecto

Contenido

1. Branches del proyecto:
    - Develop: Se unifica las funcionalidades implementadas en las otras branches que se encargan de cada implementación.
    - Master: se utiliza para subir los resultados finales después de probar exitosamente las funcionalidades implementadas.
    - Release: se utiliza para buscar errores antes del lanzamiento del software.
    - MapSimulation: se crea con la finalidad de implementar la funcionalidad que tiene como objetivo visualizar la información de la concentración de las partículas contaminantes en los departamentos del país.
    - DataAnalysis: tiene como objetivo implementar el algoritmo de data mining con el cual se busca realizar una predicción con los datos de la base de datos y representarlos en una gráfica estadística.

2. Definición de reportes

    - Reporte 1. Simulación de niveles de contaminación partículas PM por medio de un mapa
    - Reporte 2. Diagrama de barras que representa las predicciones de los niveles de contaminación.

3. Listado de responsabilidades
    - Alejandro Martínez: ui
    - Johann Andrei Ocampo: documentación
    - José David Jiménez: model
    - Cristian Lasso: analítica de datos
    - Ana María Muñoz: configuración git

4. Data mining

Para analizar los datos decidimos usar **algoritmos de regresión**, que predicen una o más variables numéricas continuas, basándose en otros atributos del conjunto de datos. Para lo cual investigamos y encontramos en la página *docs.microsoft.com* información acerca de que algoritmos serían los adecuados para predecir un atributo continuo como: Pronosticar las ventas futuras, predecir los visitantes de un sitio a partir de tendencias históricas proporcionadas o generar una puntuación de riesgo a partir de datos demográficos.

Para ello recomendaba los 3 siguientes algoritmos:

1. Algoritmo de árboles de decisión.

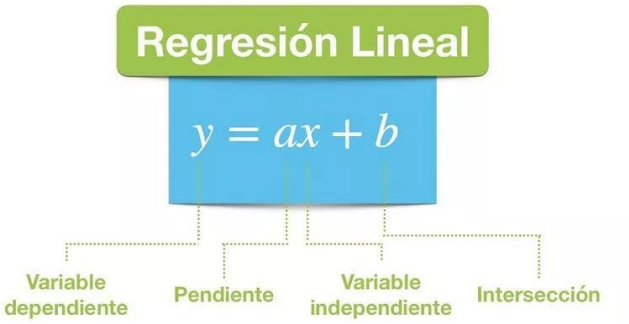
2. Algoritmo de serie temporal.
3. Algoritmo de regresión lineal.

Después de investigar acerca de cada uno tomamos la decisión de que el algoritmo de regresión lineal era el que se nos facilitaba más y sería igualmente efectivo.

Regresión Lineal Simple – Teoría

El modelo de Regresión Lineal es tan simple que muchos argumentan que no es digno de ser clasificado como Machine Learning. Este algoritmo es un método estadístico que nos permite resumir y estudiar las relaciones entre dos variables continuas cuantitativas. La Regresión Lineal es una técnica paramétrica utilizada para predecir variables continuas, dependientes, dado un conjunto de variables independientes.

Matemáticamente, la regresión usa una función lineal para aproximar o predecir la variable dependiente dada como:



En la cual Y es la variable dependiente, X la variable usada para la predicción, A es la pendiente y B la constante también conocida como intercepto (cuando X es igual a 0, Y será igual a B).

De esta forma el algoritmo se emplea para usar los datos ya conocidos y hallar el patrón de acuerdo al cual varía, para así crear una línea de predicción que puede decirnos cómo se comportan estos datos más allá de la información conocida. Paginas web: *docs.microsoft.com* -- *Regresión Lineal*

Alejandro Martínez - José D. Jiménez - Cristian Lasso - Johann A. Ocampo - Ana M. Muñoz

ESPECIFICACIÓN DE REQUERIMIENTOS

Massive Operator of Numerical Data Analysis

(M.O.N.D.A)

Contenido

1. Introducción.....5

1.2. Definiciones, Acrónimos y Abreviaturas.....5

2. Funciones del Producto.....6

3. Requisitos específicos.....6

1. Introducción

La Especificación de Requisitos Software se realiza con el propósito de que el lector pueda tener comprensión del programa que se desarrolla y así mismo de las funcionalidades planteadas para su interacción. De esta manera, se resalta que el programa cumple con el propósito general de analizar algunos de los parámetros que describen la calidad del aire en Colombia.

A partir de lo anterior, teniendo en cuenta la elección de unas variables que describen la cantidad de partículas contaminantes en el aire llamadas PM 10 y PM 2.5, se plantea como objetivos poder representar gráficamente la concentración de estas partículas en una región particular del país, específicamente en un departamento. El usuario tendrá la posibilidad de escoger el departamento del cual desea conocer la información y también se establecerá una predicción del comportamiento que podría presentar la contaminación en el área estudiada a partir del análisis de los datos obtenidos en la base de datos.

Finalmente, se destaca que el programa no permite que los datos sean modificados debido a que provienen de una base de datos ajena a sí. Los datos que serán estudiados y utilizados para la implementación de los reportes estadísticos y la predicción del posible comportamiento siguiente, provienen de mediciones reales realizadas por distintas entidades gubernamentales de Colombia en el año 2017.

1.2 Definiciones, Acrónimos y Abreviaturas

- Partículas PM10: Se pueden definir como aquellas partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 μm (1 micrómetro corresponde la milésima parte de 1 milímetro).
- Partículas PM2.5: es una mezcla que puede incluir sustancias químicas orgánicas, polvo, hollín y metales. Estas partículas pueden provenir de los automóviles, camiones, fábricas, quema de madera y otras actividades.
- Regresión lineal: consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente o respuesta se le identifica como Y y a la variable predictora o independiente como X. Una vez generado un modelo que se pueda considerar válido, es posible predecir el valor de la variable dependiente Y para nuevos valores de la variable predictora X.

2. Funciones del Producto

El objetivo del software es realizar análisis a partir de una base de datos sobre la contaminación en Colombia, para lograr predecir cuáles podrían ser los niveles de

concentración de partículas contaminantes en el aire en el país después del 2017. Para esto se hará uso de un algoritmo que permitirá analizar los datos y llevar regresión lineal y se mostrarán los resultados de los datos en 2 ventanas aparte; la primera tendrá un reporte de manera estadística y en la siguiente se mostrará un mapa el cual ubique los puntos de concentración de las partículas contaminantes en un área determinada de la región.

3. Requisitos específicos

1. El software debe ser capaz de cargar la base de datos desde la URL https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/DATOS-DE-CALIDAD-D,EL-AIRE-EN-COLOMBIA-2011-2017/ysq6-ri4e, al terminar esta accion, en pantalla debera aparecer un banner informativo con el mensaje “Base de datos cargada efectivamente”,
   en caso de que ocurra un error en la carga deberá aparecer un banner de error con el mensaje “Fallo en la carga de la base de datos, por favor vuelva a intentarlo”.
   El único tipo de dato válido será el String, como el dato tipo String no tiene ninguna restricción de escritura los escenarios posibles serán :
   - ➔ El programa accedera a la URL y podrá cargar la base de datos y saldrá el banner informativo antes expuesto.
   - ➔ El programa no podrá acceder a la URL y saldrá el banner de error antes expuesto.

   En caso de algún fallo, este no pasará de la aparición de una excepción.

2. El software contará con búsquedas, en el que podrá filtrar la información de la base de datos basado en las columnas que esta tiene y el valor a buscar en ella (Ej: Columna: Departamento; Valor: Valle del cauca), podrá someter la búsqueda a cuantos filtros desee y podrá ver los filtros que va haciendo en una lista. Al presionar el botón ”buscar”, en pantalla se mostrará una tabla en el que podrá ver todos los datos que según sus filtros cumplen con las condiciones. El unico tipo de dato que se recibira en esta funcion sera String, como este no tiene restricción de escritura los escenarios posibles serán:
   - ➔ El valor que el usuario digite no se encuentra en la columna donde quiere buscar y como consecuencia la tabla aparecerá vacías y saldrá un banner informativo donde dice “Valor no encontrado, por favor realice de nuevo su búsqueda” y se reiniciará la búsqueda.
   - ➔ Todos los valores que el usuario filtra se encontraran en sus respectivas columnas y la tabla mostrará todos los datos que cumplen con la condición.
3. Usará el algoritmo implementado para realizar las predicciones sobre el nivel de concentración de partículas contaminantes estudiadas en el aire de Colombia de la base de datos. Éste algoritmo solamente funcionará sobre la base de datos

especificada con anterioridad. Si funciona correctamente se mostrará un banner informativo donde indicará que todo salió exitosamente.

4. Después de que el algoritmo de predicción se ejecute correctamente, se abrirá una nueva ventana en donde se mostrará un reporte de texto junto a un diagrama de barras que representarán los datos de predicción.

5. Se abrirá una nueva ventana en donde se mostrará un mapa de Colombia donde se pueda visualizar el nivel de concentración de partículas contaminantes estudiadas en el aire de Colombia.
   .