FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

QUITO AIR QUALITY MODELING AND PREDICTION USING METEOROLOGICAL AND POLLUTION DATA

Trabajo de Titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en Sistemas de Computación de Informática

Profesor Guía

Ing. Mario Salvador González Rodríguez

Autor

Martín Nicolás Almeida Gachet

Año

2019

DECLARACIÓN DEL PROFESOR GUÍA

Declaro haber dirigido el trabajo, *Quito air quality modeling and prediction using meteorological and pollution data*, a través de reuniones periódicas con el estudiante Martín Nicolás Almeida Gachet, en el semestre 201920, orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación

_____

Mario Salvador González Rodríguez

Doctor en Ingeniería Informática y de Telecomunicación

CI: 0958376345

DECLARACIÓN DEL PROFESOR CORRECTOR

Declaro haber revisado este trabajo, *Quito air quality modeling and prediction using meteorological and pollution data*, del estudiante Martín Nicolás Almeida Gachet, en el semestre 201920, dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación

_____

Paulo Roberto Guerra Terán

Máster en Ciencias de la Computación

CI: 1002856050

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE

Declaro que este trabajo es original, de mi autoría, que se han citado las fuentes correspondientes y que en su ejecución se respetaron las disposiciones legales que protegen los derechos de autor vigentes.

_____

Martín Nicolás Almeida Gachet

CI: 1718947706

RESUMEN

En este documento se describe el proceso de diseño, implementación y resultados de diferentes funciones de Machine Learning para utilizarlos objetivamente con los datos meteorológicos del municipio de Quito. Datos que han sido recolectados desde el año 2004 hasta 2018 horariamente por la REMMAQ en nueve estaciones alrededor del distrito. Se enfoca principalmente en algoritmos de regresión lineal, análisis de series temporales y análisis de interpolación espacial.

Durante la definición de gráficos a ser implementados en la aplicación AirQ2 se realizó un proceso de análisis sobre la contribución de cada tipo de gráfico para obtener un mejor resultado al construir modelos de predicción sobre la contaminación del aire de Quito. Cada gráfico aquí descrito aporta con información específica y brinda una guía para conocer más sobre los diferentes contaminantes y estaciones de los datos recolectados, así como su contribución en la construcción de modelos de predicción.

ABSTRACT

This document describes the process of design, implementation and results of different Machine Learning functions to use them objectively with meteorological data from the municipality of Quito. Data that has been collected from 2004 to 2018 hourly by the REMMAQ in nine stations around the district. It focuses mainly on linear regression algorithms, time series analysis and spatial interpolation analysis.

During the definition of graphs to be implemented in the AirQ2 application, an analysis process was carried out on the contribution of each type of graph to obtain a better result when constructing prediction models on air pollution in Quito. Each graphic described here provides specific information and provides a guide to learn more about the different pollutants and stations of the data collected, as well as their contribution in the construction of prediction models.

ÍNDICE

1. INTRODUCCIÓN

La polución o contaminación ambiental es la presencia en el aire de componentes dañinos, cuya fuente son las actividades industriales y la necesidad que se deriva del desarrollo de la vida moderna. Es importante tener en cuenta que los recursos naturales, aunque son abundantes, no son inagotables. Por esto es preciso racionalizar y reducir su consumo.

Con la creciente preocupación por el medio ambiente han nacido asociaciones y organismos cuyo propósito es el estudio, conocimiento y protección de la naturaleza. A pesar de esto, no se ha detenido el proceso de agotamiento y malversación de los recursos naturales ni la contaminación del medio ambiente. Debido a los efectos de la actividad industrial y del aumento de la población mundial, el ser humano está conduciendo al planeta a una situación límite y de alto riesgo.

Se debe destacar que la calidad de aire en países de primer mundo ha incrementado en los últimos años gracias a los esfuerzos académicos y legislativos. No ocurre lo mismo en Latinoamérica, donde todavía se evidencia una baja calidad de aire, que se relaciona fuertemente a problemas de salud.

Estudios realizados en la ciudad de Quito durante la última década encontraron que después de dos a tres años los efectos de cada regulación legislativa se desgastan. Esto es producto del incremento del parque automotriz en la ciudad.

Utilizando tecnología para procesar grandes volúmenes de datos se busca analizar la relación entre diferentes contaminantes presentes en el aire de la ciudad de Quito y predecir el incremento de la contaminación. Así mismo, con la ayuda de herramientas de visualización se colocará de forma dinámica esta información en una aplicación web denominada AirQ2 (*Quito Air Quality*). Esto con el objetivo que usuarios finales puedan acceder a este tipo de análisis colocando filtros de interés como contaminante, fechas y zonas de la ciudad. Se puede obtener estos resultados gracias a técnicas estadísticas como regresión lineal, regresión lineal múltiple, *random forest*, PCA (*Principal Component Analisis*), análisis de

interpolación espacial por IDW (Interpolación con la Distancia Inversa Ponderada) y análisis de series temporales.

1.1. Alcance

Este proyecto propone la utilización de técnicas de Machine Learning para obtener predicciones sobre los datos meteorológicos del Municipio de Quito, enfocados principalmente al análisis temporal como input de los algoritmos de aprendizaje y la comparación con análisis tradicional de series de tiempo, así como análisis e interpolación espacial. Los datos que se utilizarán como materia prima son los archivos históricos de la calidad del aire del Municipio de Quito, que se encuentran disponibles en la página del mismo. Estos archivos se encuentran en formato Excel y contienen datos de los siguientes contaminantes: CO (monóxido de carbono), NO2 (dióxido de nitrógeno), O3 (ozono), PM2.5 (partículas menores a 2.5 micrómetros), PM10 (partículas menores a 10 micrómetros) y SO2 (dióxido de azufre). Adicionalmente se cuenta con datos sobre la dirección del viento, el nivel de humedad, radiación solar, velocidad del viento y nivel de precipitación de lluvia.

Es importante destacar que estos datos se encuentran divididos por sectores del Municipio de Quito, siendo estos los sectores donde se encuentran estaciones de la REMMAQ. Estas estaciones se encuentran en: Belisario, Carapungo, Centro, Cotocollao, El Camal, Guamaní, Jipijapa, Los Chillos, San Antonio y Tumaco. Por otro lado, también se cuenta con una frecuencia de recolección de datos de una hora, por lo que los datos disponibles han sido recolectados desde el primero de enero de 2004 todos los días cada hora hasta el 30 de junio de 2018.

Primero se deberá hacer un modelado y unión de los datos en formato Excel y transformarlos a un formato estándar que se guardará en un archivo csv por cada sector que tendrá la siguiente estructura (Tab. 1):

Tabla 1

Estructura de datos planteada

Campo	Nombre Completo
Station	Estación
Date_time	Fecha y hora

CO	Monóxido de carbono
DIR	Dirección del viento
HUM	Humedad
LLU	Lluvia
NO2	Dióxido de nitrógeno
O3	Ozono troposférico
PM2.5	Partículas menores a 2.5 micrómetros
PM10	Partículas menores a 10 micrómetros
PRE	Presión barométrica
RS	Radiación solar
SO2	Dióxido de azufre
TMP	Temperatura media
VEL	Velocidad del viento

Dichos archivos servirán como insumo para los algoritmos donde se trabajará con Machine Learning para el análisis de aprendizaje temporal y análisis de series de tiempo. Adicionalmente se implementará análisis e interpolación espacial, donde a partir de los datos de un contaminante en las estaciones se podrá obtener un aproximado de la concentración del contaminante en una coordenada determinada dentro del Municipio de Quito. Este proceso de limpieza de datos será realizado con el lenguaje de programación Python, utilizando principalmente las librerías Numpy y Pandas.

Estas soluciones de Machine Learning serán implementadas utilizando el lenguaje de programación R, que es ampliamente utilizado para computación estadística. Finalmente, los resultados de los métodos antes mencionados serán visualizados por medio de una aplicación web desarrollado en base a R Shiny, siendo este un paquete de R que brinda una interfaz de programación fácil de utilizar y robusto para crear aplicaciones web que soporten las librerías de R para el procesamiento de datos.

1.2. Justificación

Hoy en día existe una creciente preocupación ambiental. Las alcaldías y el gobierno deben tener una participación activa en este cambio de mentalidad para garantizar un planeta vivo a las futuras generaciones. Esta es la importancia de

predecir la concentración de contaminantes en el aire del Municipio de Quito, y esta predicción debe sustentarse en una sólida base tecnológica y científica.

Los habitantes del Municipio de Quito pueden no estar al tanto de los niveles de concentración de contaminantes a los que se exponen diariamente, y peor aún, cómo se proyectan estos niveles en el futuro al ritmo al que se mueve la metrópolis. Gracias a la construcción de estos modelos de análisis de series temporales, análisis espacial e interpolación espacial se podrá conocer los niveles de concentración de contaminantes y tomar medidas de prevención a nivel de ciudadanía y alcaldía.

Utilizando estos modelos basados en aprendizaje automático se va a predecir los niveles de concentración de diferentes contaminantes en el aire a partir de los datos recolectados por la REMMAQ. La propuesta consiste en un modelo de análisis de series temporales, análisis e interpolación espacial utilizando algoritmos de aprendizaje supervisado para pronosticar niveles de concentración de contaminación en diferentes puntos de Quito utilizando factores meteorológicos y los niveles de concentración de contaminación en las estaciones recolectoras de datos.

1.3. Metodología a utilizar

En el proyecto aquí propuesto se utilizará la metodología de investigación inductiva. Se tomará como punto de partida los datos recolectados de la calidad del aire de Quito para realizar predicciones utilizando diferentes algoritmos de Machine Learning.

Para realizar la gestión del proyecto se utilizará la metodología ágil SCRUM ya que el desarrollo de la solución se verá beneficiada con un método iterativo e incremental. Adicionalmente la constante comunicación entre las partes involucradas en la solución habilitará un ambiente propicio para realizar los cambios necesarios según las historias de usuarios recopiladas. Esto significa una reducción del tiempo utilizado en corregir potenciales errores que se hagan presentes a lo largo del desarrollo de la solución.

Para el desarrollo de la solución se utilizará la metodología CRISP-DM, cuyas siglas en ingles significan “proceso estándar de la industria para minería de datos”. Esta metodología divide la minería de datos en seis pasos:

Entendimiento del negocio

Se trata de entender de qué trata el negocio para el que se realiza la solución. Se debe tener en cuenta las posibles preguntas que puedan beneficiar al negocio y que la respuesta se encuentre en los datos que serán extraídos.

Entendimiento de los datos

Se debe entender el significado de los datos que se extraen. Puede que para el negocio unos datos sean más relevantes que otros, dependiendo de la información que brindan y la interpretación que se les pueda dar a estos. Se debe tener conocimiento de que preguntas pueden responder los datos que se obtienen en el proceso de minería de datos.

Preparación de los datos

Este es un proceso de limpieza de datos, frecuentemente se trata el procesamiento de datos no existentes y datos atípicos. Se propone tratar los datos no existentes con el cálculo del promedio de los datos más relacionados en el conjunto de datos. Los datos atípicos se propone re-calcularlos con el promedio como se busca hacer con los datos no existentes.

Modelado

En este paso se busca mantener un formato estándar de los datos según las necesidades del negocio. La estructura del conjunto de datos que se busca se describe en el alcance de este documento.

Evaluación

Se realizan pruebas de la solución con los datos procesados. Al tener una interpretación de los resultados obtenidos se puede determinar si el proceso realizado en los datos fue correcto o se obtiene retroalimentación para volver a pasos previos, dependiendo del caso.

Despliegue

Una vez que se tiene conformidad con los datos obtenidos y la interpretación de los resultados de las pruebas se procede al despliegue de la solución. Se obtiene retroalimentación del proceso realizado y volvemos al paso de entendimiento del negocio.

Es posible no realizar estos pasos en secuencia, se puede regresar a un paso anterior o adelantarse a un paso en caso de necesitarlo. Es importante mencionar que el proceso de minería de datos continúa después de que la solución haya sido desplegada. Es posible que siguiendo esta metodología después del despliegue de la solución se encuentren nuevas preguntas más centralizadas en el negocio. CRISP-DM es una metodología iterativa, por lo que cuando se procede a realizar el despliegue de la solución se recomienda volver al paso de entendimiento del negocio con la retroalimentación de las iteraciones previas.

1.4. Problema Por Resolver

Hoy en día existe una creciente preocupación ambiental. Las alcaldías y el gobierno deben tener una participación activa en este cambio de mentalidad para garantizar un planeta vivo a las futuras generaciones. Esta es la importancia de predecir la concentración de contaminantes en el aire del Municipio de Quito, y esta predicción debe sustentarse en una sólida base tecnológica y científica.

Los habitantes del Municipio de Quito pueden no estar al tanto de los niveles de concentración de contaminantes a los que se exponen diariamente, y peor aún, cómo se proyectan estos niveles en el futuro al ritmo al que se mueve la metrópolis. Gracias a la construcción de estos modelos de análisis de series temporales, análisis espacial e interpolación espacial se podrá conocer los niveles de concentración de contaminantes y tomar medidas de prevención a nivel de ciudadanía y alcaldía.

Utilizando estos modelos basados en aprendizaje automático se va a predecir los niveles de concentración de diferentes contaminantes en el aire a partir de los datos recolectados por la REMMAQ. La propuesta consiste en un modelo de análisis de series temporales, análisis e interpolación espacial utilizando algoritmos

de aprendizaje supervisado para pronosticar niveles de concentración de contaminación en diferentes puntos de Quito utilizando factores meteorológicos y los niveles de concentración de contaminación en las estaciones recolectoras de datos.

1.5. Objetivo General

Implementar un sistema web para la predicción de contaminación del aire de la ciudad de Quito para facilitar los procesos de análisis multidisciplinarios de la información procesada.

1.6. Objetivos Específicos.

- Desarrollar algoritmos de análisis de series temporales en base a librerías para machine learning incluidas en R.
- Desarrollar algoritmos de regresión lineal que puedan predecir el nivel de contaminación del aire en un punto determinado.
- Desarrollar algoritmos de interpolación espacial para calcular un valor aproximado de contaminación en los puntos entre las estaciones recolectoras de datos en la ciudad.

2. MARCO TEÓRICO

2.1. Lenguaje R

Es un ambiente de código abierto que provee herramientas para la computación estadística y la realización de gráficos. Tiene un enfoque de código abierto y está basado en el lenguaje S, brinda a los usuarios una amplia gama de herramientas para análisis estadístico. El proyecto R afirma que la ventaja de este lenguaje de programación es “la facilidad con la que se pueden producir gráficos con calidad de publicación” (*R Project*, 2018).

2.2. Shiny

Es un paquete de R que facilita la construcción de aplicaciones web interactivas utilizando R. Combina el poder computacional de R con la interactividad de la web moderna.

2.3. Machine Learning

Es un conjunto de técnicas usado para construir modelos complejos y algoritmos que son utilizados para realizar predicciones. Utiliza técnicas estadísticas para brindar a las computadoras la habilidad de “aprender” a partir de un conjunto de datos sin ser programados explícitamente para un propósito.

Su premisa básica utilizar análisis estadístico en sus algoritmos que reciben datos de entrada para predecir los datos de salida, mientras que estos datos de salida se convierten en datos actualizados de entrada.

Los procesos involucrados en Machine Learning son muy similares a los de minería de datos y modelado predictivo. Ambos requieren analizar los datos en busca de patrones y ajustar las acciones de los programas según los resultados.

Muchas personas se familiarizan con Machine Learning gracias al realizar compras por internet y observar publicidad relacionada a sus compras anteriores. Esto se da por motores de recomendación que utilizan Machine Learning para personalizar la entrega de publicidad online casi en tiempo real. Más allá del marketing personalizado, Machine Learning también se utiliza en detección de fraude, filtro de SPAM, detección de amenazas para la seguridad de redes, mantenimiento predictivo y la construcción de los titulares de noticias.

Generalmente los algoritmos de Machine Learning se categorizan en supervisados y no supervisados.

Los algoritmos supervisados requieren de un científico de datos o un analista de datos con habilidades en Machine Learning para proveer de entradas y salidas esperadas. Adicionalmente suministran retroalimentación sobre la precisión de las predicciones durante el entrenamiento del algoritmo. Son los científicos de datos quienes determinan que variables o características el modelo debería analizar y

usar para desarrollar predicciones. Una vez que el entrenamiento se completa, el algoritmo aplica lo que aprendió con datos nuevos.

Una famosa definición de la inteligencia artificial como el estudio de cómo hacer que los computadores hagan lo que, al momento, las personas son mejores (Rich y Knight, 1991, pp 150-152).

Según Russell, S y Norvig, P, la historia de la ingeligencia artificial se forza en concentrar una pequeña cantidad de personas y eventos e ignorar a otros que también fueron importantes (*Artificial Intelligence*, pp 10-11).

Por otro lado, los algoritmos de aprendizaje no supervisado no necesitan ser entrenados con datos de salida esperados. En cambio, estos abordan el problema utilizando métodos iterativos, que se conoce como “Deep Learning”, para revisar datos y llegar a conclusiones. Estos algoritmos, también llamados redes neuronales, se utilizan para tareas de procesamiento más complejo que los algoritmos de aprendizaje supervisado. Se suelen utilizar en reconocimiento de imagen, voz a texto y generación de lenguaje natural. Estas redes neuronales trabajan combinando millones de ejemplos de datos de entrenamiento e identificando automáticamente correlaciones sutiles entre varias variables. Una vez entrenado, el algoritmo puede utilizar su banco de asociaciones para interpretar nuevos datos. Estos algoritmos han sido factibles de ejecutar gracias a la era de Big Data, ya que requieren cantidades masivas de data de entrenamiento.

2.4. Regresión Lineal

Es una técnica estadística que se utiliza para estudiar la relación entre variables. Cuando se usa solamente dos variables (una dependiente y una independiente o predictora) se denomina regresión lineal simple, mientras que cuando se analiza más de dos variables (una dependiente y varias predictoras) se llama regresión lineal múltiple. En ambos casos se utiliza la regresión lineal para explorar y cuantificar la relación entre una variable y sus predictoras. Así mismo se desarrolla una ecuación lineal con fines predictivos.

Tusell, F afirma que, “Son frecuentes en la práctica situaciones en las que se cuenta con observaciones de diversas variables, y es razonable pensar en una

relación entre ellas. El poder determinar si existe esta relación —y, en su caso, una forma funcional para la misma— es de sumo interés. Por una parte, ello permitiría, conocidos los valores de algunas variables, efectuar predicciones sobre los valores previsibles de otra. Podríamos también responder con criterio estadístico a cuestiones acerca de la relación de una variable sobre otra.” (Análisis de Regresión. Introducción Teórica y Práctica basada en R, 2011, pp1).

2.5. Feature Selection

Es la selección de atributos de los datos que son más relevantes para el modelo predictivo en el que se está trabajando. Es un proceso que ayuda al crear un modelo predictivo preciso seleccionando características de los datos que brindan mayor precisión de predicción utilizando menos datos.

2.6. Random Forest

Es un algoritmo de aprendizaje supervisado. Crea varios árboles de decisión para realizar predicciones más precisas. Una de sus grandes ventajas es que Random Forest puede ser utilizado para problemas de regresión y clasificación.

2.7. PCA

PCA (*Principal Component Analysis*) es un método de reducción de dimensionalidad que suele usarse para reducir la dimensionalidad de grupos de datos muy grandes. Esto lo logra transformando un amplio grupo de variables en una más pequeña que todavía contenga la mayoría de información del grupo más amplio.

2.8. IDW

IDW (*Inverse Distance Weightning*) es un método determinista de interpolación espacial que estima un valor desconocido en una ubicación utilizando valores conocidos en puntos conocidos.

2.9. Time Series Forecasting (LSTM)

Las series temporales involucran datos recolectados secuencialmente en el tiempo. Las redes de *Long Short Term Memory* (LSTM) son una forma especial de redes neuronales recurrentes que son capaces de aprender dependencias a largo plazo.

3. ANÁLISIS DEL PROBLEMA

La Red Metropolitana de Monitoreo Atmosférico de Quito (REMMAQ) se puso en marcha en el año 2002. Su propósito según la Secretaría del Ambiente es “producir datos confiables sobre la concentración de contaminantes atmosféricos en el territorio del Distrito Metropolitano de Quito que sirvan como insumo para la planificación, formulación, ejecución y evaluación de políticas y acciones orientadas al mejoramiento de la calidad del aire y difundir esta información en condiciones comprensibles para el público en general” (2018).

Son nueve estaciones de recolección de datos las que conforman la REMMAQ. Cada estación tiene equipado una cantidad de sensores que se utilizan para obtener datos meteorológicos y la concentración de contaminantes en el aire. Los datos recopilados por cada estación son accesibles a través de un grupo de gráficos colgados en el sitio web de la secretaría del ambiente, en incluso pueden ser descargados.

La secretaría del ambiente ha seccionado estos datos históricos diferenciándolos por contaminante y variable meteorológica a la que corresponden los datos, y utilizando otra categoría para identificar la estación de la que provienen los datos. También estos datos se encuentran en archivos separados dependiendo del periodo de tiempo en el que se recopilaron los datos, siendo inexistente un control de limpieza de datos para estos archivos. Ya que la cantidad de datos existentes actualmente y la gran diversidad de formatos en los que se encuentran guardados, para realizar la agrupación de datos y brindarles un formato en común (limpieza de datos) se requiere de conocimientos en el uso de sistemas computacionales y lenguajes de programación como R y Python.

Los datos ambientales que proporciona la secretaría del ambiente pueden generar un gran impacto en campos tanto ambientales como jurídicos, vale la pena recalcar que en estos campos no se necesita tener un conocimiento técnico sobre los métodos de limpieza de datos ni aseguramiento de la calidad de los datos y menos aún de la generación de representaciones gráficas por medio de sistemas de computación. El diseño de procesos de análisis de datos necesariamente requiere de conocimientos sobre las características individuales de los diferentes contaminantes y las variables atmosféricas que se relacionan fuertemente al estudio del medio ambiente.

Actualmente no se tiene una percepción adecuada sobre que variables meteorológicas o que estaciones de recolección de datos se encuentran fuertemente correlacionadas, ni sobre que variables meteorológicas serían las más importantes para construir un modelo de predicción de contaminación basado en un contaminante que se desea predecir. Por esta razón se busca implementar las siguientes características dentro del sistema AirQ2:

- Visualización de relación entre estaciones por medio de clusters y dendogramas.
- Visualización de correlación entre variables meteorológicas por estación.
- Visualización de contaminación en diferentes puntos de la ciudad de Quito utilizando análisis de interpolación espacial por medio del método IDW y dispersado por medio de un heatmap.
- Visualización de la importancia de variables meteorológicas para la construcción de modelos de predicción de la variable seleccionada.
- Visualización de árboles de decisión creados por el algoritmo de Random Forest, utilizado para Feature Selection y predicción.
- Visualización de relación entre el dato de contaminación del día de la toma del dato versus el dato de contaminación del día anterior.
- Visualización de la importancia de componentes principales y la importancia de cada variable para la construcción de los componentes principales más relevantes según PCA.

- Visualización de predicción de contaminación según análisis de series temporales mediante algoritmo LSTM.

4. DISEÑO DE LA SOLUCIÓN

Según la metodología CRISP-DM el primer paso es el entendimiento de negocio. En este punto se debía entender el trabajo de la REMMAQ con la recolección de datos, el significado de los mismos, el propósito de este proceso y donde encontrar dichos datos. Los datos se colocan cada seis meses en la página web de la secretaría del ambiente en formato Excel. Estos contienen recolección de los datos contaminantes desde el año 2004 recolectados en forma horaria.

El siguiente paso es el entendimiento de los datos. Los datos dentro de los archivos Excel se encontraban divididos en estaciones, fecha y hora en la que el dato fue tomado, y la variable de contaminante a la que correspondía el dato. Los contaminantes que se encontraron dentro de este proceso de recolección se encuentran listados en Tab. 1.

Pasado este punto se procede a la preparación de los datos. En este proceso lo que se realizó fue una limpieza de los datos. Eliminando valores atípicos por NaN, y una vez hecho esto procesar los valores NaN por la media del contaminante en determinado punto. Cuando los datos obtenidos se encontraron completos se procedió a construir los diferentes archivos con los datos separados según la necesidad de la aplicación. Se propuso dos estructuras según la necesidad de la funcionalidad de la aplicación web a desarrollar. La primera estructura divide los datasets por estación (un archivo de datos por cada estación presente) y propone la estructura de Tab. 1. Con respecto a la segunda, se plantea una estructura donde se divide el dataset por variables, por lo que se obtiene un archivo de datos por cada variable de contaminante existente. La estructura propuesta se muestra en Tab. 2.

Tabla 2

Estructura de datos por contaminante

Campo	Nombre Completo
Date_time	Fecha y hora

BELISARIO	Valor en Belisario
CARAPUNGO	Valor en Carapungo
COTOCOLLAO	Valor en Cotocollao
EL_CAMAL	Valor en El Camal
LOS_CHILLOS	Valor en Los Chillos
TUMBACO	Valor en Tumbaco

Los siguientes pasos de modelado, evaluación y despliegue se describen a continuación.

Después de realizar en análisis del problema y las características deseadas se propone realizar la implementación de las funcionalidades sobre el sistema AirQ2 que se encuentra desarrollado en R utilizando el paquete Shiny. En esta aplicación se han hecho disponibles los conjuntos de datos que fueron provistos por la secretaría del ambiente. Previo a la carga de los datos en la aplicación AirQ2 se realizó un proceso de limpieza y modelado de datos para que sean accedidos por el sistema. Adicionalmente esta aplicación cuenta con un grupo de gráficos que facilitan el análisis de datos meteorológicos.

Las nuevas funcionalidades se encontrarán desplegadas sobre una nueva pestaña llamada “Predicción”. Dentro se encontrarán las opciones para acceder a las pantallas de:

- Clusters
- Correlation
- IDW
- Variable Importance
- Simple Tree
- Data Firm
- PCA
- Time Series

Cada una de estas pantallas contará con uno o dos gráficos descriptivos sobre los datos disponibles y el propósito de la pantalla como tal. Estos gráficos serán personalizables por los usuarios del sistema por medio de controles que podrán manipular para modificar el gráfico desplegado. Esto gracias a las herramientas

que brinda el paquete Shiny. Se debe destacar que los controles que se planea implementar son dinámicos, lo que significa que cambian dependiendo de la pantalla donde se encuentra el usuario. Esto ayuda para no sobrecargar la interfaz de usuario con controles inservibles ni saturar al servidor con información que no va a utilizar en la generación de los gráficos.

Para generar los gráficos de los datos se utilizarán librerías de R que facilitan esta tarea. Estas librerías son ggplot2 y plotly, que permiten la versátil generación de gráficos según el tipo de gráfico que se requiere. Existen gráficos que se generarán utilizando directamente los paquetes de factoextra y party. Específicamente estos gráficos serán utilizados en el análisis de series temporales y Random Forest.

Adicionalmente para el desarrollo del sistema se planea utilizar Scrum, con avances semanales y reuniones constantes con el profesor guía para la verificación de los avances realizados.

5. DESARROLLO E IMPLEMENTACIÓN DE LA SOLUCIÓN

La metodología de desarrollo utilizada en este proyecto fue Scrum. Para esto se definieron sprints de una semana de duración, donde se presentaron avances significativos respecto a la aplicación en desarrollo y el análisis respectivo de los resultados. La prioridad de los requerimientos que se debían desarrollar fueron acordados en cada reunión con el dueño del producto según sus necesidades.

Scrum es una metodología iterativa, por lo que en cada reunión del respectivo sprint se presentaron avances del sistema y se tomaba en cuenta la retroalimentación del dueño del producto para implementar mejoras. Adicionalmente estas reuniones semanales facilitaron para reducir requerimientos complejos en problemas pequeños y mostrar avances sobre estas pequeñas metas hasta resolver el requerimiento complejo.

Las funcionalidades implementadas en el sistema tienen como propósito el análisis y predicción de las variables de contaminación ambiental de forma visual. Esto quiere decir, por medio de gráficos mostrar las variables más importantes para la predicción de una variable seleccionada, mostrar mapas de calor complementados con interpolación espacial sobre el mapa de la ciudad de Quito, gráficos de

tendencia para análisis de series temporales, etc. Dichas funciones son dinámicas por medio de filtros colocados en un panel de control en la aplicación. Estos controles brindan al usuario la habilidad de filtrar los datos según estación, variable a analizar, tiempo, así como la posibilidad de seleccionar opciones para la visualización de los gráficos como decidir si remover o no valores atípicos, seleccionar la variable a analizar, etc. Estos controles son filtrados dependiendo del gráfico que se muestra en pantalla para brindar una interfaz de usuario más limpia, que contenga solo los controles que se van a utilizar en la pantalla actual.

En esencia el usuario ingresa al sistema a la pantalla deseada, modifica los filtros según sus requerimientos y la aplicación se encarga de cargar los datos a la función delegada para tratar el requerimiento de la pantalla. En esta función se construye el gráfico solicitado en base a los parámetros configurados por el usuario y lo muestra en pantalla, siendo posible para el usuario realizar el análisis deseado.

Las pestañas implementadas se encuentran debajo de la opción “Prediction” implementada en la aplicación para separar los gráficos utilizados en análisis de datos de los utilizados en predicción. Estas son: Clusters, Correlation, IDW, Variable Importance, Simple Tree, Data Firm, PCA, Time Series y Linear Relation. Esto se puede apreciar en Fig. 1, donde se muestra la interfaz de la aplicación.

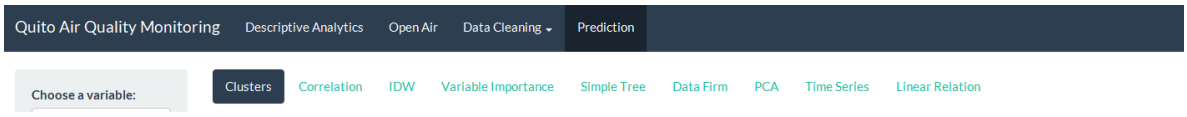

Figura 1. Menú implementado en AirQ2

En la pestaña de Clusters se implementaron dos gráficos, un dendograma y un gráfico de cluster, donde se puede apreciar la relación de las estaciones en base a una variable determinada en los filtros del panel izquierdo, número de clusters para realizar la agrupación, entre otros. Aquí se pude realizar un análisis de la relación entre estaciones basados en el contaminante seleccionado.

Cuando el usuario navega a la ventana de Correlation puede observar los filtros para seleccionar si realizar un análisis por contaminante o por estación y el contaminante o la estación para analizar. El gráfico mostrado en pantalla muestra

el nivel de correlación entre estaciones (en el caso de realizar un análisis por contaminante) o la correlación entre contaminantes (en el caso de realizar un análisis por estación).

Una vez seleccionada la pantalla de IDW se muestra un mapa del municipio de Quito. Sobre este mapa se puede ver marcas sobre cada estación de recopilación de datos, así como un mapa de calor superpuesto que muestra visualmente con colores los lugares de Quito que presentan una mayor concentración del contaminante que se está analizando.

En la pestaña de Variable Importance se encuentra solo un filtro que corresponde a la variable que se desea analizar. El gráfico mostrado en pantalla muestra visualmente las variables más importantes para construir un modelo de predicción del contaminante seleccionado. Se debe aclarar que este proceso se realiza por medio de random forest que permite realizar un análisis de Feature Selection.

Simple Tree es una pestaña donde se puede visualizar un árbol de decisión típico de random forest, donde se puede configurar el contaminante a analizar y la profundidad del árbol que se va a visualizar. De esta forma se tiene un control sobre la saturación visual en la interfaz gráfica y se tiene una idea de cómo funciona el algoritmo de random forest para la predicción de contaminantes.

La ventana Data Firm muestra un gráfico que revela la relación entre la medida de una variable con su lectura anterior. Aquí se tiene control sobre la estación que se desea analizar y el contaminante respectivo. Existen variables donde se puede apreciar una relación lineal entre la lectura actual comparada a la anterior lectura de ese contaminante.

PCA corresponde al análisis de componentes principales. En esta pestaña se visualizan dos histogramas. El izquierdo que corresponde al porcentaje de varianza explicada y el derecho que muestra la contribución de cada variable para el componente principal seleccionado. Los controles que se encuentran aquí son la estación que se desea analizar, el componente principal del que se quiere ver la contribución de las variables y el número de componentes principales que se quieres visualizar en el histograma.

Time Series es una pestaña donde se realiza análisis de series temporales. Aquí podemos visualizar un gráfico de la predicción del valor de un contaminante según la estación seleccionada en los filtros y el grupo de tiempo (diario, semanal o mensual). Adicionalmente se implementó un checkbox para que el usuario pueda decidir si hacer su análisis removiendo valores atípicos o manteniéndolos.

La relación lineal entre dos variables es especialmente importante para poder determinar si se puede utilizar el algoritmo de regresión lineal para la predicción de un contaminante en base a una variable o no. Eso se puede visualizar en la ventana de Linear Relation donde se implementaron los controles de estación, primera variable (dependiente), segunda variable (independiente) y la opción para remover valores atípicos en caso de así desearlo.

6. PRUEBAS Y RESULTADOS

En el sistema AirQ2 se implementó un total de nueve funcionalidades nuevas enfocadas a Machine Learning. Esto se realizó utilizando como base la aplicación AirQ2 desarrollada en el paquete de R llamado Shiny y utilizando los datos mostrados en su sección de analítica descriptiva. Dentro de esta aplicación se encuentran disponibles datos sobre la contaminación del aire de Quito recopilados por la REMMAQ desde el año 2004 hasta junio de 2018 con un periodo de una hora entre dato y dato.

Previo a la utilización de los datos colgados en la página web de la Secretaría del Ambiente se realizó un proceso de modelado y limpieza de datos, donde se obtuvo como resultado un archivo por cada estación con el siguiente formato:

Tabla 3

Estructura de datos por estación

Campo	Nombre Completo
Station	Estación
Date_time	Fecha y hora
CO	Monóxido de carbono
DIR	Dirección del viento
HUM	Humedad
LLU	Lluvia
NO2	Dióxido de nitrógeno

O3	Ozono troposférico
PM2.5	Partículas menores a 2.5 micrómetros
PM10	Partículas menores a 10 micrómetros
PRE	Presión barométrica
RS	Radiación solar
SO2	Dióxido de azufre
TMP	Temperatura media
VEL	Velocidad del viento

Una vez que se obtuvo los datos con un formato estándar se procedió a desarrollar la primera funcionalidad, llamada en la aplicación “Clusters”, donde el objetivo es observar la relación entre estaciones dependiendo de la variable que se quiere analizar, como se muestra en Fig. 2.

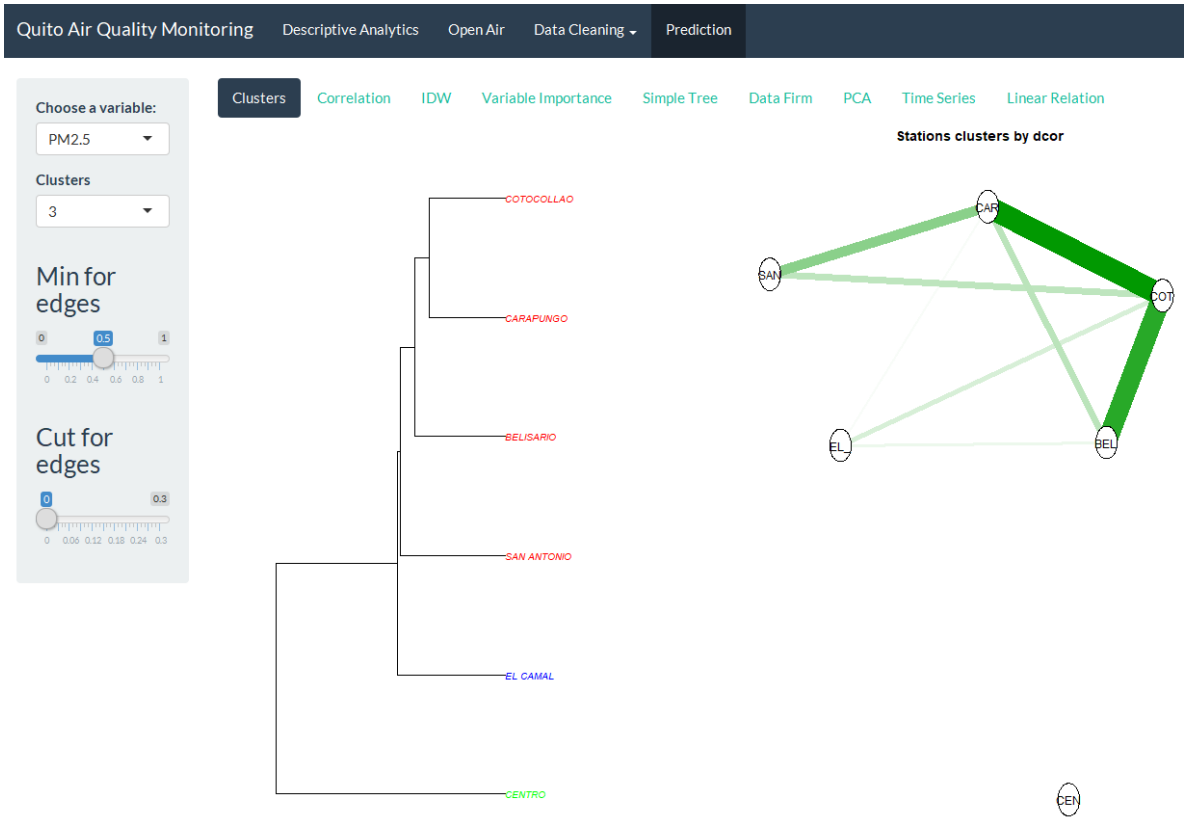

Figura 2. Relación entre estaciones según PM2.5

En el gráfico anterior se muestra la relación entre estaciones según el contaminante PM2.5 y muestra la agrupación entre estaciones según este contaminante. Así mismo en la figura verde se muestra la fuerza de la relación entre estaciones, mostrándose como un verde más intenso las conexiones más

fuertes. Se muestra también a la estación “Centro” aislada al resto, lo que concuerda con el dendograma de la izquierda donde no se la agrupa con ninguna estación.

La siguiente funcionalidad implementada muestra en un gráfico la correlación entre estaciones dependiendo del contaminante, o la correlación entre contaminantes dependiendo de la estación. Esto se puede apreciar en Fig. 3, donde se observa los filtros dinámicos para realizar este análisis a conveniencia del usuario.

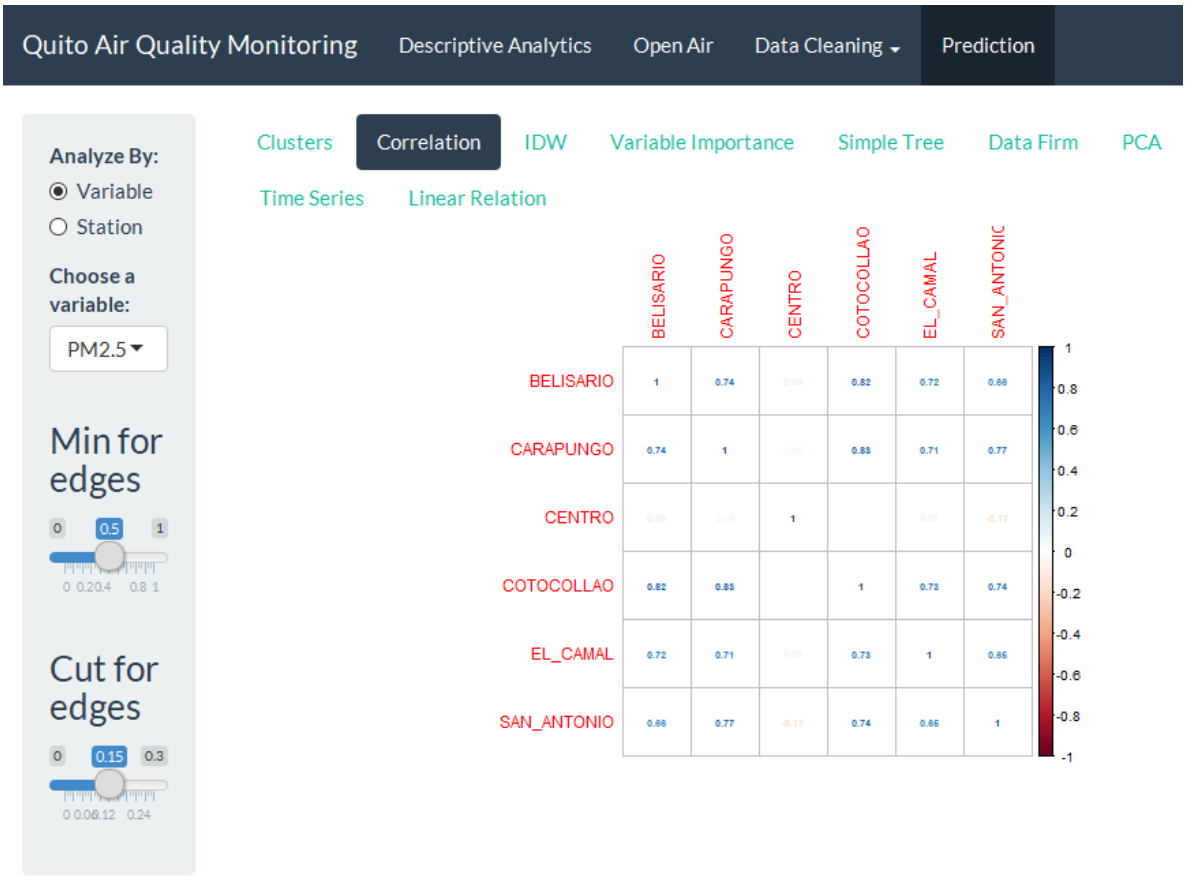

Figura 3. Correlación entre estaciones según PM2.5

Como se mencionó antes, se tiene una opción para ver la correlación entre contaminantes según la estación seleccionada. Eso es gracias a los controles dinámicos implementados en la ventana. En Fig. 4 se puede observar este cambio en el gráfico según los filtros seleccionados.

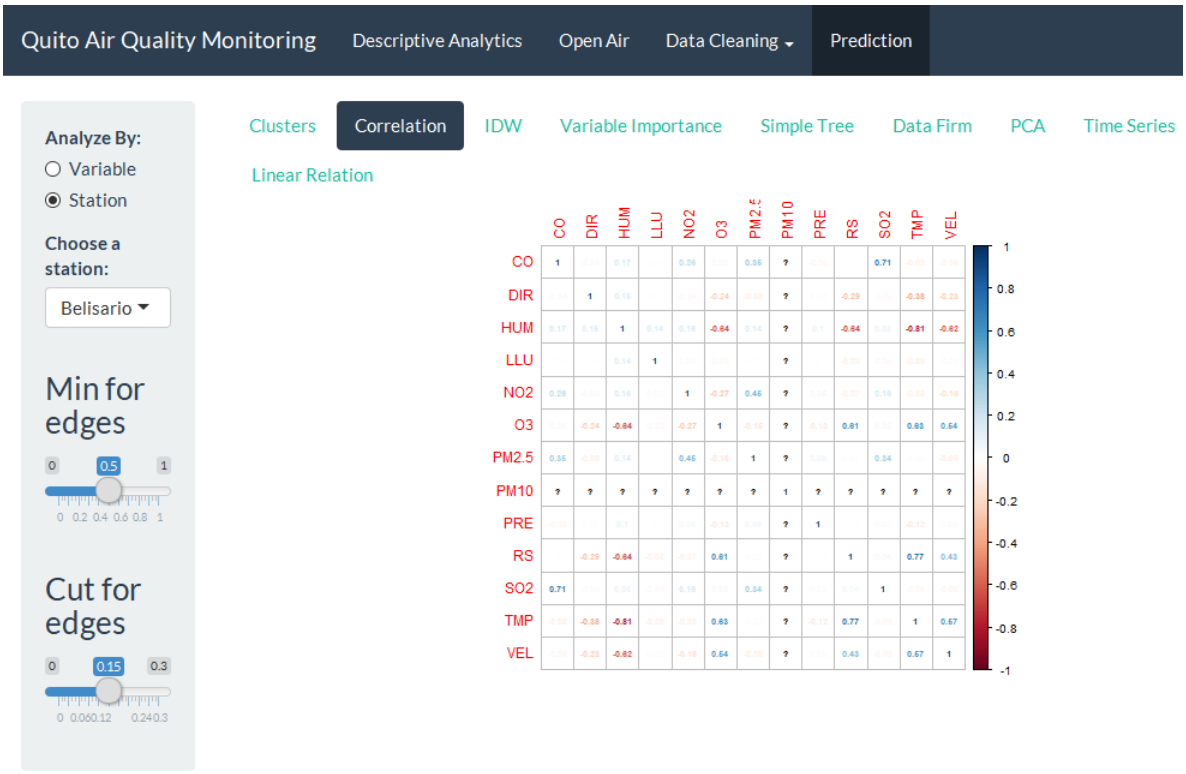

Figura 4. Correlación entre contaminantes según estación Belisario

Tanto en Fig. 3 como en Fig. 4 se mira como las estaciones o contaminantes con mayor correlación se les asigna un color azul más intenso junto con un valor más cercano al 1. Caso contrario, cuando la correlación es muy débil el color que corresponde es el rojo con un valor que se aproxima al -1. En el caso que en el recuadro se presente el símbolo “?” significa que ese contaminante no tiene valores medidos por la estación seleccionada.

En la ventana de IDW se realizó un análisis de interpolación espacial por el algoritmo de distancia inversa ponderada. En Fig. 5 se expone un mapa del municipio de Quito con un mapa de calor sobrepuesto sobre un área cubierta por las estaciones que contienen datos correspondientes a los filtros seleccionados. En el mapa de calor se muestra con color rojo las zonas con un mayor valor del contaminante seleccionado, mientras que el color azul muestra los valores más bajo.

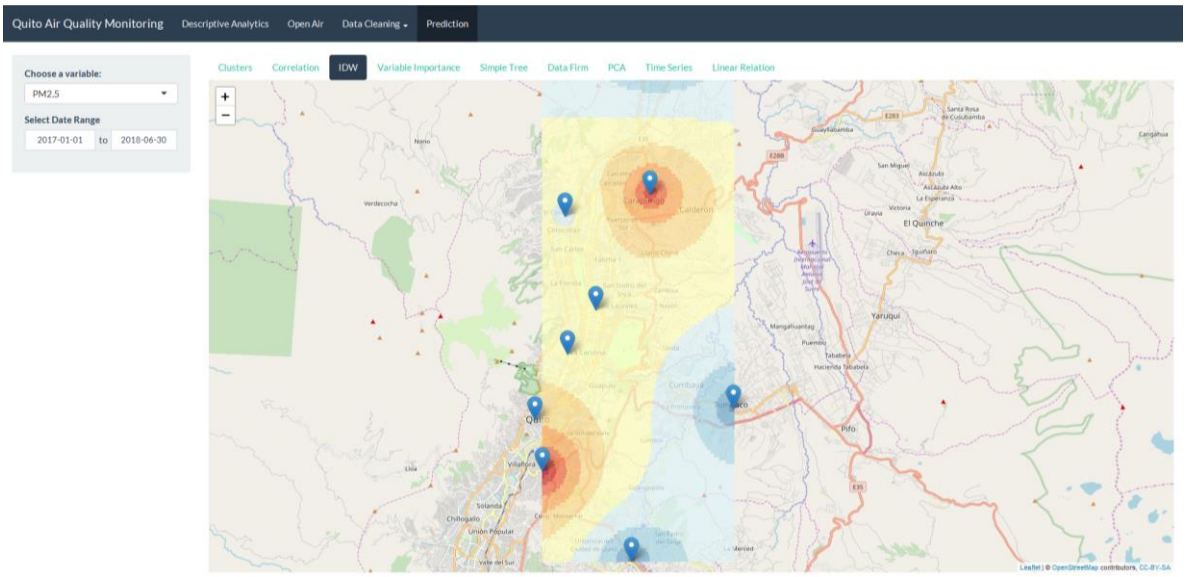

Figura 5. Mapa de calor sobre mapa de Quito basado en PM2.5

En análisis de interpolación espacial se agregó también filtros de fechas para mostrar la firma del contaminante sobre la ciudad en un determinado periodo.

Cuando se habla de predicción es importante conocer las variables más importantes para construir los modelos de predicción según el algoritmo que se vaya a utilizar. Es por esto que se agregó la pestaña de “Variable Importante” donde se observa (Fig. 6) un gráfico que expone a las variables más importantes para predecir un contaminante seleccionado según el algoritmo de Random Forest.

Figura 6. Importancia de variables de predicción para PM2.5

Lo que se muestra en Fig. 6 es el aporte de cada variable al modelo de predicción construido por Random Forest. Es importante de igual manera visualizar la coherencia entre estos datos graficados con la realidad, por lo que en la pestaña

de “Simple Tree” se obtiene una vista de un árbol típico construido por este algoritmo.

Como se muestra en Fig. 7, en esta ventana es posible explorar el árbol en base al contaminante seleccionado y la profundidad que se desea ver en pantalla.

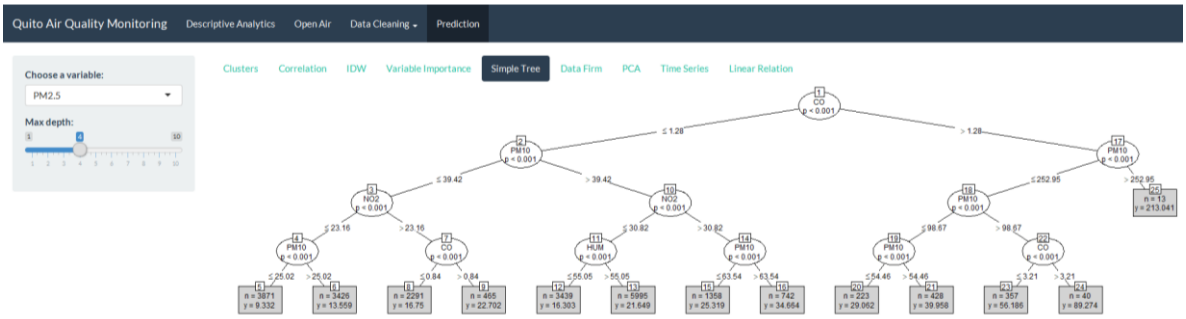

Figura 7. Árbol típico de Random Forest para PM2.5 con profundidad de 4

Random Forest construye varios de estos árboles de decisión y devuelve un resultado en base a la respuesta de la mayoría de árboles de decisión.

De igual forma es interesante observar la firma característica de un contaminante, que se analizó entre el valor medido contra el valor previo del contaminante, agrupado por datos diarios, semanales y mensuales. Adicionalmente se agregaron filtros de estación y contaminante para realizar el gráfico deseado, que se muestra en Fig. 8.

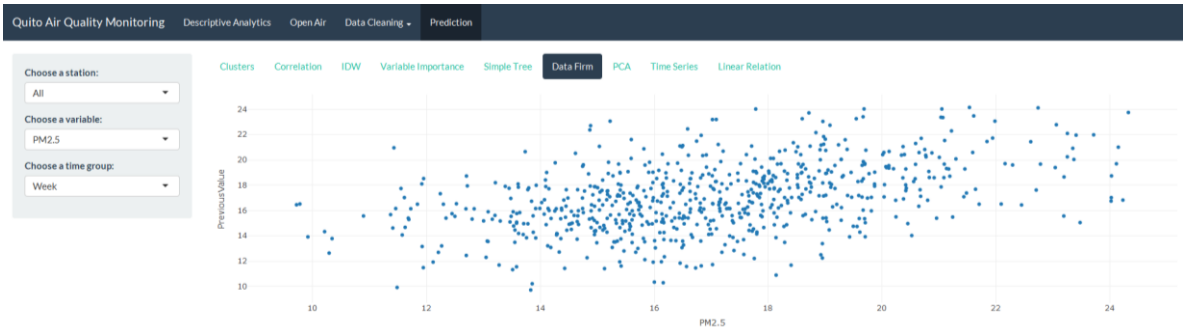

Figura 8. Firma semanal de PM2.5 en todas las estaciones

En Fig. 8 se observa la relación entre PM2.5 con el valor previo medido y se muestra una clara relación lineal. Estos hallazgos son de gran ayuda para la construcción de variables personalizadas al crear un modelo de predicción. Incluso

se puede utilizar como base para la elección de un algoritmo de Machine Learning como regresión lineal.

Otro método relevante para realizar un proceso de selección de variables es PCA (análisis de componentes principales). Este divide los datos en componentes principales y analiza la importancia de cada variable por componente principal. En Fig. 9 se puede ver la importancia de variables en los cinco primeros componentes especiales (sumatoria), así como el aporte de cada componente principal a los datos.

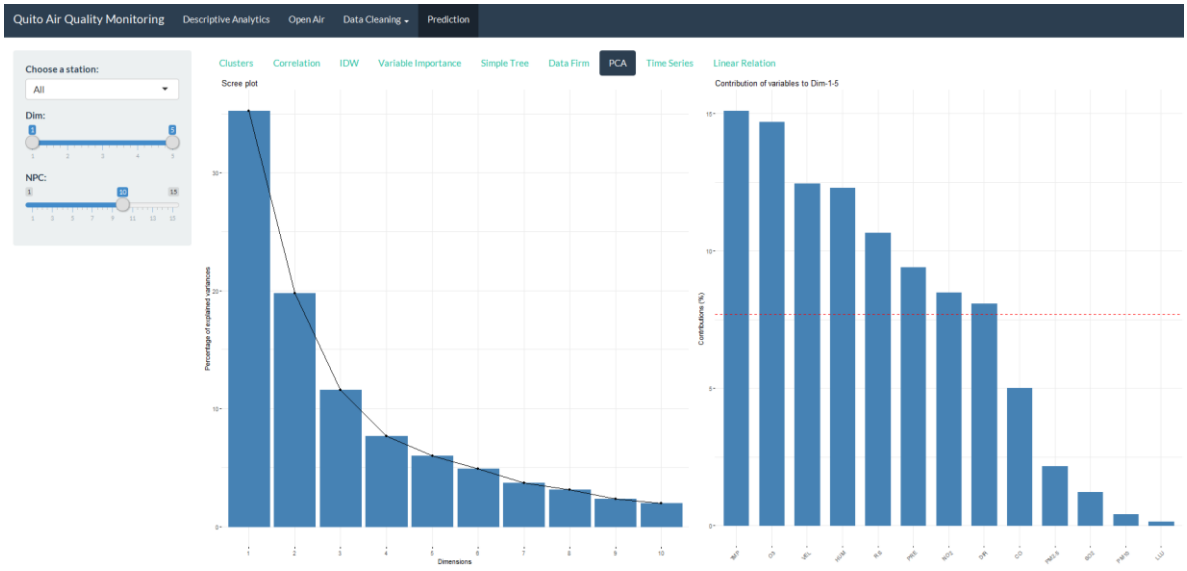

Figura 9. Análisis de componentes principales para todas las estaciones

Se puede apreciar dos gráficos, el izquierdo en Fig. 9 corresponde al aporte de cada componente principal, mientras el derecho muestra la importancia de cada variable al componente principal. La línea roja que se aprecia es la contribución media esperada por cada variable del modelo, sin embargo, aquellas que superan este valor claramente tienen una mayor importancia para construir modelos de predicción.

El análisis de series temporales corresponde a un área de Machine Learning donde se utilizan modelos de predicción para estimar valores futuros de un contaminante. Después de pre procesar los modelos de predicción y guardarlos en archivos encriptados (extensión h5) se logró utilizarlos en el análisis de contaminantes en

el tiempo, según su media diaria, semanal o mensual. Esto se puede apreciar en Fig. 10.

Figura 10. Análisis de serie temporal para PM2.5 con media semanal

Se puede apreciar que la figura de la línea de predicción (naranja) tiene la misma forma de los datos actuales (azul). Analizando las figuras de la media mensual (Fig. 12) y diaria (Fig. 11), se ve claramente que al utilizar datos más detallados (media diaria) el modelo se vuelve más exacto que con una media más generalizada (mensual).

Figura 11. Análisis de serie temporal para PM2.5 con media diaria

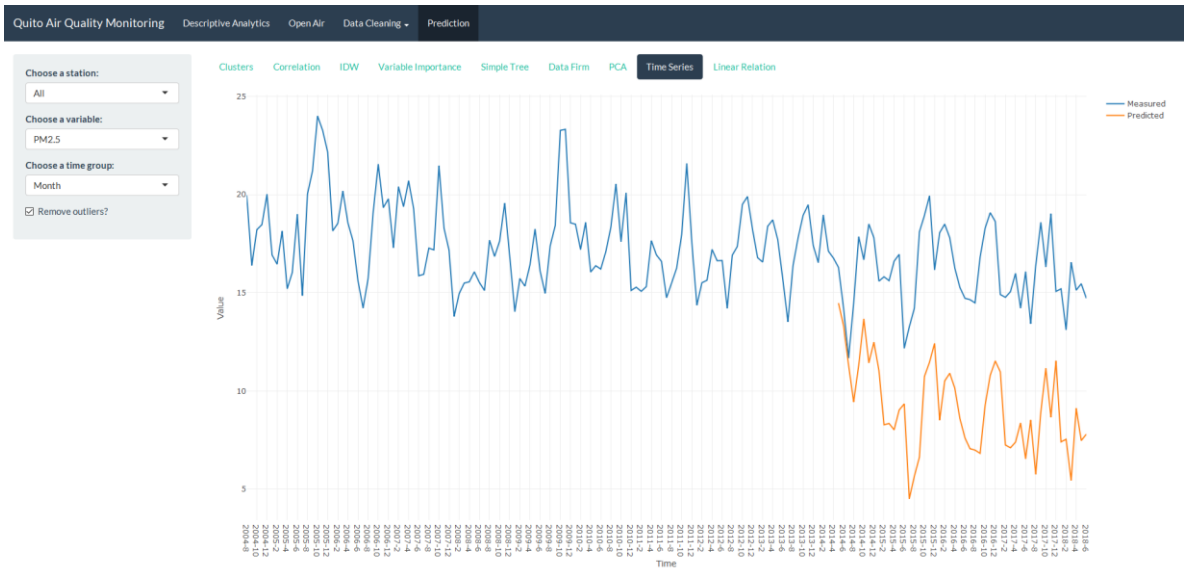

Figura 12. Análisis de serie temporal para PM2.5 con media mensual

Cuando se utiliza la media mensual para la construcción del modelo se observa que, aunque la figura se mantiene igual, los datos varían notablemente.

Finalmente, la pestaña llamada “Linear Relation” es una guía para obtener la relación lineal entre dos variables seleccionadas y filtradas por estación. Sirve como herramienta para conocer que variables son aptas para aplicar el algoritmo de regresión lineal al construir un modelo de predicción. En Fig. 13 se puede observar la relación lineal entre PM2.5 y PM10, sin remover valores atípicos.

Figura 13. Relación lineal entre PM2.5 y PM10 con todos los datos

Por otro lado, al remover datos atípicos para generar el gráfico se obtiene lo mostrado en Fig. 14. Los datos atípicos verdaderamente cambian al análisis de relación lineal e influyen considerablemente en el resultado esperado.

Figura 14. Relación lineal entre PM2.5 y PM10 sin datos atípicos

Todas estas son una serie de herramientas que tienen como propósito facilitar el desarrollo de Machine Learning enfocado a la contaminación del municipio de Quito. Se espera contar con un servidor dedicado para montar la aplicación y poder acceder a ella dentro de la red de la Universidad de las Américas. De este modo se podría ampliar la funcionalidad de la aplicación AQ2 para analizar mayores volúmenes de datos y realizar algoritmos de predicción más complejos sin las limitantes del servidor local en el que se desarrolla actualmente.

7. CONCLUSIONES y RECOMENDACIONES

7.1. Conclusiones

El desarrollo de análisis de series temporales montado en una aplicación Shiny permite ver dinámicamente el resultado de las predicciones del algoritmo implementado. Gracias a esto se observaron las grandes diferencias entre construir el modelo con datos detallados (diario), medianamente detallados (semanal) y generalizados (mensual). Con estos resultados se concluye que el modelo cuya construcción se realizó con datos detallados es el más preciso de los tres. Lo mismo se puede apreciar para las diferentes combinaciones de contaminante, estación y grupo de tiempo.

Para determinar si es factible utilizar los algoritmos de regresión lineal en la construcción de un modelo de predicción es necesario analizar la relación entre dos variables y observar una clara relación lineal. No siempre es el caso, y realizar este análisis antes de implementar regresión lineal puede ahorrar mucho tiempo de desarrollo desperdiciado al desarrollador.

Al realizar interpolación espacial se puede estimar el valor de un contaminante en un punto de la ciudad basado en la lectura de este contaminante en puntos conocidos (estaciones de REMMAQ). Lo que se implementó en AirQ2 es una función que pueda dividir la ciudad en cuadrantes y calcular una media del contaminante seleccionado en cada cuadrante. En base a este valor estimado se asigna un color a cada cuadrante, por lo que se puede concluir que al tener más divisiones sobre la ciudad el mapa de calor va a ser más exacto. Así mismo, se debe tomar en cuenta que solo se cubre el área dentro de las estaciones que cuentan con datos del contaminante seleccionado y que cumplan con los filtros de la aplicación.

Después de realizar el desarrollo del sistema web para la predicción de contaminación del aire de la ciudad de Quito se concluye que el mismo constituye una herramienta que facilita los procesos de análisis multidisciplinarios.

7.2. Recomendaciones

Durante el desarrollo de las nuevas funcionalidades de AirQ2 fue evidente la limitante del servidor local utilizado. Es muy recomendable el uso de controles dinámicos para la configuración de parámetros que puedan afectar al rendimiento del servidor utilizado.

Con respecto a la construcción de modelos de predicción, se recomienda utilizar datos detallados (diarios en el caso de AirQ2), ya que al generalizar los datos se pierde precisión de los datos a predecir, aunque se mantenga la forma de la curva.

Es muy recomendable también realizar análisis de relación lineal previo a la implementación de un algoritmo de regresión lineal, ya que depende de esta relación la precisión del algoritmo de predicción.

En cuestión de series temporales es recomendable utilizar valores semanales debido a la precisión de los datos y facilidad de lectura de los gráficos. Esto se recomienda en lugar de utilizar datos mensuales o datos diarios, ya que se pierde la precisión de la predicción o no se aprecian los resultados del modelo predicho respectivamente.

REFERENCIAS

Anniss, M. (2014). *What Is a Website and How Do I Use It?* Nueva York, Estados Unidos: Britannica Educational. Recuperado el 10 de Abril de 2019, de https://books.google.es/books?isbn= 1622750748

Baumer, B. S., Kaplan, D., & Horton, N. J. (2017). *Modern Data Science with R.* Boca Raton, Estados Unidos: CRS Press. Recuperado el 20 de Marzo de 2019, de https://books.google.es/books?isbn=1498724493

Cuesta, H. (2013). *Practical data analysis.* Birmingham: Packt Publishing.

Iafrate, F. (2015). *From big data to smart data.* Hoboken, Estados Unidos: John Wiley & Sons.

IBM. (2016). *IBM SPSS Modeler CRISP-DM Guide.* Nueva York, Estados Unidos: IBM.

Jugulum, R., & Gray, D. H. (2014). *Competing with Data Quality.* Hoboken, Estados Unidos: John Wiley & Sons.

Lutz, M. (2013). *Leaning Python: Powerful Objetc-Oriented Programming.* Sebastopol, Canada: O´ Reilly Media. Recuperado el 20 de Marzo de 2019, de https://books.google.es/books?isbn=9781449355692

Olmeda-Gómez, C. (2014). Visualización de Información. El profesional de la información, 23(3), 213-219.

R Project. (2018). *What is R?* Recuperado el 12 de Mayo de 2019, de https://www.r-project.org/about.html

Ramírez, O., Mura, I., & Franco, J. (2017). *How do people understand urban air pollution? Exploring citizens' perception on air quality, its causes and impacts in Colombian cities. Open Journal of Air Pollution, 6*(1), 1-1.

Resnizky, H. G. (2015). *Learning Shiny.* Birmingham, Reino Unido: Packt *Publishing*.

Schwaber, K., & Sutherland, J. (2017). La Guía Definitiva de Scrum: Las Reglas del Juego. Recuperado el 5 de Junio de 2019, de https://www.scrumguides.org/index.html

Scrum, T. A. (2015). Vanderjack, Brian. Nueva York, Nueva York: *Business Expert Press*. Recuperado el 5 de Junio de 2019, de https://ebookcentral.proquest.com/lib/udlap/detail.action?docID=2145193

Secretaría de Ambiente del Municipio del Distrito Metropolitano de Quito. (2018). Generalidades:Red de Monitoreo Atmosférico. Recuperado el 5 de Marzo de 2019, de http://www.quitoambiente.gob.ec/ambiente/index.php/generalidades

Sinharay, R., Gong, J., Barratt, B., Ohman-Strickland, P., Ernst, S., Kelly, F. J., . . . Chung, K. F. (2018). *Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls*. *The Lancet, 391*(10118), 339-349.

Spinu, V., Grolemund, G., Wickham, H., Lyttle, I., Constigan, I., Law, J., . . . Lee, C. H. (11 de Abril de 2018). *Package 'lubridate'.* Recuperado el 5 de Junio de 2019, de R Project: https://cran.r-project.org/web/packages/lubridate/lubridate.pdf

Torres Ponjuán, D., & Herrero-Solana, V. (2010). La visualización de la información en el entorno de la Ciencia de la Información. La Habana: Universidad de La Habana.

Vanderjack, B. (2015). *The agile edge: managing projects effectively using agile scrum.* Nueva York, Estados Unidos: Business Expert Press.

Zalakeviciute, R., López-Villada, J., & Rybarczyk, Y. (2018). *Contrasted effects of relative humidity and precipitation on urban pm 2.5 pollution in high elevation urban areas*. *Sustainability, 10*(6), 1-21.

Zalakeviciute, R., Rybarczyk, Y., López-Villada, J., & Suarez, M. V. (2018). *Quantifying decade-long effects of fuel and traffic regulations on urban ambient pm2.5 pollution in a mid-size south american city*. *Atmospheric Pollution Research, 9*(1), 66-75.