Master en Big Data y Data Science

Aplicación de Técnicas de Aprendizaje Automático Supervisado para el Pronóstico del Número de Casos Nuevos de Covid19 en Bogotá-Colombia

Trabajo Fin de Máster

Alumno: FALLA GAITAN, CARLOS ANDRES

Dirección: BOGOTÁ

Tutor Trabajo Fin de Máster: PERALTA MARTÍN-PALOMINO, ARTURO

Edición octubre 2019 a octubre 2020

Índice

Índice de ilustraciones

Índice de tablas

Resumen

En este trabajo se abordan técnicas de aprendizaje supervisado, para realizar un pronóstico del número de nuevos casos positivos en la evolución del virus Covid-19 en la ciudad de Bogotá, capital de Colombia, y visualizar los resultados obtenidos mediante una dashboard basado en la herramienta Qlik Sense.

La selección del tema del TFM, se basa en la necesidad de aplicar en un caso práctico algunas de las técnicas, metodologías y herramientas de Big Data, aprendidas en el Máster. Así mismo, a partir del listado de temas propuestos por la Universidad para TFM, se seleccionó el tema T14 – *Análisis y monitoreo de variables influyentes en la propagación del Cov-19*, propuesto por el tutor Arturo Peralta Martín-Palomino; que, en el contexto de la pandemia vivida en los últimos meses, le da no sólo el sentido práctico al trabajo, sino también un sentido social, humano y colaborativo con las necesidades actuales que vive el planeta entero.

Una vez seleccionado el tema, se evalúa con el tutor la conveniencia de enfocar el trabajo en el estudio de la propagación del virus en una ciudad específica, valorando diferentes variables de contexto de la ciudad que puedan tener relación con la propagación del virus, pero que permitan hacer un pronóstico de esta. De esta forma el trabajo no se basa en trabajos previos realizados por el tutor, y aunque bajo la crisis que vive Colombia junto al resto del mundo, hay una gran cantidad de trabajos que persiguen establecer el número de casos de contagios nuevos, no se usa ningún trabajo como base de partida para este TFM.

Basados en este contexto, el trabajo propone identificar los datos diarios de casos positivos de Covid-19, datos de oferta de servicios de salud, datos de trasporte, clima y disposiciones del gobierno; para crear un conjunto de variables de estudio y establecer con ellas un modelo que permita pronosticar el número de casos nuevos con resultado positivo.

Como valor adicional, partiendo de que el conjunto de datos principal de casos de Covid-19, es un conjunto de datos cronológico que podría analizarse y estimarse a través de técnicas de series temporales, tal como lo ha hecho el grupo de investigación en Software Inteligente y Convergencia Tecnológica GISIC de la Facultad de Ingeniería de la Universidad Católica de Colombia (Facultad de Ingenieria, 2020); este TFM plantea realizar el pronóstico de manera alternativa mediante el uso de técnicas de Aprendizaje Supervisado como Árboles de Decisión o KNN (CART), encontrando la relación que las demás variables de contexto puedan tener con el número de casos y como con esa relación podemos llegar a predecir el número de casos para un periodo posterior.

1. Introducción

El Covid-19 es una nueva enfermad cuyo origen tuvo lugar en Wuhan (China) a finales del año 2019. Los primeros casos se identificaron con un diagnóstico de neumonía de origen desconocido, el patógeno se identificó como un nuevo betacoronavirus de ARN (Cender Quispe-Juli, 2020) que actualmente se ha denominado coronavirus del síndrome respiratorio agudo severo 2 (SARS-CoV-2), por su similitud con el SARS-CoV (Cender Quispe-Juli, 2020). En la actualidad se ha extendido por todo el mundo, hasta mayo del 2020 se reconocía su presencia en 184 países y se habían reportado cerca de 4 millones (María Matilde García Lorenzo, 2020) de diagnósticos positivos de la enfermedad y decenas de miles de muertes.

Por su novedad, el conocimiento que se tiene sobre la enfermedad es escaso. Por su rápida propagación en diferentes países se han venido abordando estudios que permitan obtener conocimiento acerca del comportamiento, medios de diagnósticos y posibles tratamientos, así como el desarrollo de la vacuna.

La inteligencia artificial es considerada como una de las herramientas con gran potencial para afrontar este reto. Durante el desarrollo que ha tenido la enfermedad, se ha observado la rápida respuesta del mundo y de los diferentes gremios para generar soluciones, en las que se cuenta como factor principal la tecnología, casos como el establecimiento del teletrabajo, telemedicina, el fortalecimiento de las clases online, en el sector retail ventas a través de medios no presenciales, así como también el seguimiento que se le ha dado a la enfermedad con datos estadísticos diarios, y allí un gran reto, y es el manejo de volúmenes de datos para interpretar la información y bajarla a la población en términos simples y precisos.

En esta contingencia, también se ha observado la respuesta de los gobiernos en los diferentes países ante la enfermedad, así como frente al reto de mantener a la población informada, identificar de manera oportuna los casos, hacer detección y seguimiento a las personas confirmadas con el diagnóstico y garantizar que los ciudadanos cuenten con el conocimiento y las herramientas para hacer contacto y reconocer síntomas de sospecha, tarea nada sencilla. Para ello se ha tenido que valer del descubrimiento del conocimiento en un dominio de aplicación de los datos existentes, en el que es posible gracias a las diversas fuentes de información que los generan y almacenan. En ese descubrimiento de conocimiento se han tenido que desarrollar diferentes técnicas desde perspectivas tales como la estadística y la inteligencia artificial, y es así como en la actualidad se han vuelto familiares para el mundo los términos de aprendizaje automático, minería de datos, análisis de redes sociales, ciencias de datos, entre otros.

Para el diagnostico se ha establecido como prueba de referencia la reacción en cadena de la polimerasa (PCR), que permite la detección de virus a pacientes en quienes se sospecha COVID-19. Dada la demanda excesiva de insumos y reactivos de PCR para COVID-19 en el mundo, éstos se han visto escasos ante el comportamiento del virus, situación que es más evidente en los países de Latinoamérica (D., 2020), donde la falta de plataformas tecnológicas y reactivos para el dispositivo dificultan el llamado de la

OMS respecto al fortalecimiento de las capacidades de diagnóstico masivo y oportuno y su consecuente aislamiento y búsqueda de contactos.

Ante la necesidad, las pruebas inmunológicas para buscar anticuerpos IgM/IgA e IgC dirigidos con el SARS-CoV-2, así como también las pruebas rápidas (PDR) son alternativas complementarias a la PCR y sirven para el rastreo de casos asintomáticos y contactos de los casos confirmados con la prueba PCR, lo que permite encontrar, aislar e interrumpir la cadena de transmisión en menor termino (D., 2020).

En Colombia, el primer caso identificado fue el 6 de marzo de 2020 y posteriormente de identificaron 608 casos confirmados y 6 muertes asociadas a la enfermedad. Para ese entonces el país contaba con un único laboratorio apto para efectuar las PCR, ubicado en el Instituto Nacional de Salud (INS) en la capital del país, Bogotá (D., 2020). Para el 22 de marzo, el INS capacitó 22 laboratorios de salud pública y centro de investigación el país para el diagnóstico de Covid-19 a través de PCR, y el 25 de marzo el Ministerio de Salud de Colombia, apoyado por la Asociación de Infectólogos y el INS, emitió la resolución 019, que permite aplicar pruebas de diagnóstico rápido para buscar anticuerpos como tamizaje previo a la PCR que sigue siendo la prueba de referencia.

Para el abordaje de la detección de casos, Colombia se ha propuesto un modelo de atención en salud para la pandemia por Covid-19 para países con baja capacidad tecnológica y científica de diagnóstico – imagen 1- en la que se espera que, el paciente acceda voluntariamente a una aplicación móvil habilitada y creada para reporte: tanto casos sintomáticos como los asintomáticos son registrados en la base de datos, inclusive ubicación geográfica. Con ayuda de algoritmos matemáticos y minería de datos, las personas de Covid-19 son seleccionadas, los profesionales pueden añadir o excluir datos atípicos en una sala situacional virtual, lo profesionales biomédicos junto con los ingenieros direccionan a los pacientes para que sean sometido a una prueba de diagnóstico en un horario y fecha asignados, los lugares de realización de pruebas diagnósticas pueden notificarse a SIVIGILA – Sistema de Vigilancia Publica de Colombia, resultados y recomendaciones pueden enviarse al paciente a través de la aplicación móvil, tras la confirmación de los casos, se envía la información al SIVIGILA y la información de los pacientes positivos regresa a la sala situacional virtual, donde se activa la ruta inversa de atención. Aplicando nuevamente ciencia de datos e información proporcionada por los casos índices, se establece un cerco epidemiológico y se activa una red de búsqueda activa de contactos con las personas infectados a través de pruebas de diagnóstico rápido (D., 2020).

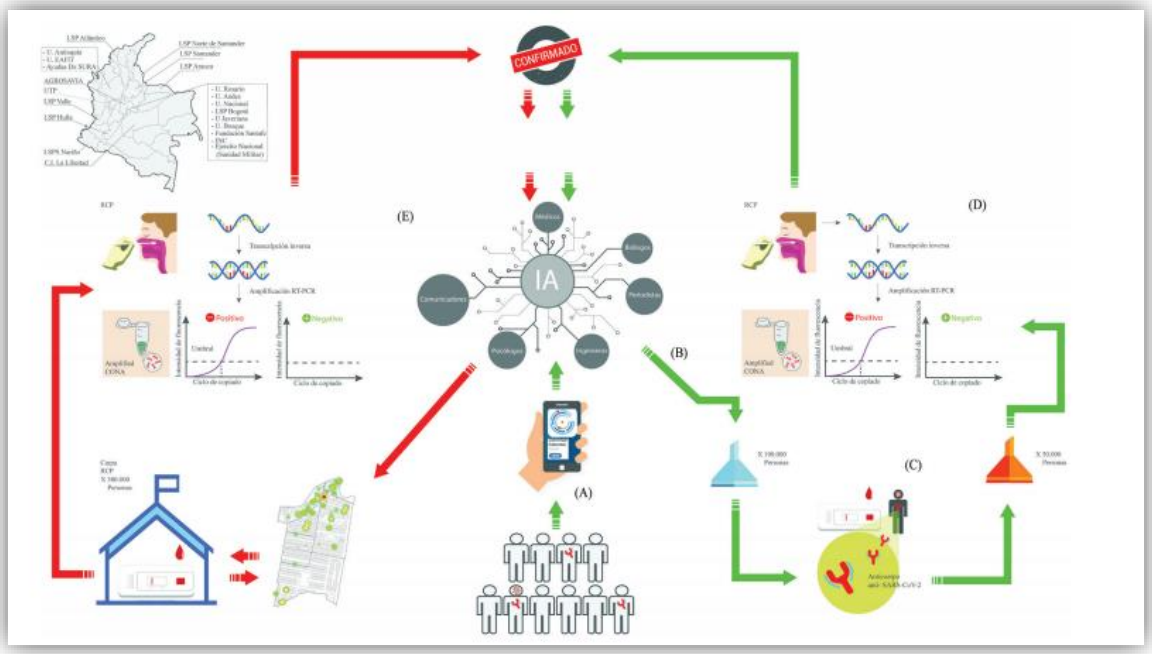

Ilustración 1 -Modelo de atención en salud para la pandemia por Covid-19 en Colombia. Fuente: Revistas Científicas y Boletines Pan American Health Organization - https://iris.paho.org/handle/10665.2/52559

Teniendo en cuenta este contexto situacional del Covid-19 y con los datos abiertos disponibles a través de la paginas web que se han dispuesto en Colombia, tales como Datos Abiertos Gov.co, se encuentran los conjuntos de datos recopilados de casos por Covid-19.

Se plantea en este trabajo la aplicación de técnicas de aprendizaje automático supervisado para el pronóstico del número de nuevos casos con diagnóstico positivo en la ciudad de Bogotá. Se ha seleccionado esta ciudad por ser la capital y la ciudad donde se concentra la mayor cantidad de población, que de acuerdo con el último censo del año 2018 son aproximadamente 7.413 millones de habilitantes, sin contar la población flotante (emigrantes no censados y personas habitantes de la calle).

Aplicando la metodología de minería de datos CRISP-DM [1] en cada una de sus fases: 1) Entendimiento del negocio, para el contexto del trabajo estaría enfocado en el contexto de la pandemia en Colombia, 2) Entendimiento de Datos: el conjunto de datos de fuentes de datos abiertos, 3) Preparación de datos: el proceso de limpieza, transformación y selección de variables, 4) Modelado: en el que se evaluará la selección de la técnica, construcción, ejecución y evaluación y finalmente un análisis de resultados, conclusiones y la identificación de oportunidades de mejora en cuanto al contexto del trabajo.

[1] CRIPS-DM : son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos (6)

2. Objetivos

2.1. Objetivo General

Plantear una alternativa de estimación del número de nuevos casos positivos para Covid-19 en la ciudad de Bogotá – Colombia, mediante el uso de técnicas de aprendizaje automático supervisado y la visualización de los resultados en un dashboard, que sirvan como herramientas de referencia para el seguimiento y control del virus, y la toma de decisiones.

2.2. Objetivos Específicos

- Realizar el análisis de datos a conjuntos de datos disponibles de los casos detectados de Covid-19 en Colombia específicamente en Bogotá utilizando la metodología CRISP-DM.

- Proponer e implementar un modelo de aprendizaje automático supervisado que acorde con el resultado del análisis de datos aplique de manera efectiva y contribuya para el pronóstico de nuevos de casos positivos de Covid-19 en la ciudad de Bogotá.

- Proponer oportunidades de mejora acorde a los resultados del modelo propuesto para el pronóstico de nuevos casos positivos de Covid- 19 en Bogotá.

- Construir un Dashboard para monitorear el comportamiento de las variables de contexto con las que se está realizando la estimación, y la evolución del número de casos, para seguimiento y apoyo en la toma de decisiones, de las entidades e instituciones sanitarias; compartiéndolo en el portal de datos abiertos www.datos.gov.co, en la sección de publicación de estudios, datos y visualizaciones.

3. Marco teórico

Ante la situación como la actual en la que como población nos enfrentamos a una pandemia por una nueva enfermedad de la cual no se tiene conocimiento y con el desarrollo de tecnología con el que contamos, surge una pregunta, ¿cómo puede la ciencia de datos ayudar en predicciones de pandemias como el Covid-19? Y ¿qué ha hecho la ciencia de datos durante los tiempos posteriores al reconocimiento de pandemia a la Covid-19?, cuestiones que tienen relevancia en el desarrollo de trabajo que se presenta en este documento en el que se pretende aportar técnicas para el pronóstico de casos bajo el método supervisado

La ciencia de datos es una disciplina que se puede utilizar en diferentes ámbitos para visualizar, predecir y solucionar problemas. También se ha demostrado que las barreras de entrada a la ciencia de datos son mínimas y que se pueden incorporar perfiles muy diversos que funcionan muy bien ante una situación tan delicada como la actual. Prueba de la versatilidad de esta tecnología se puede observar en Kaggle, unas de las comunidades para aficionados y expertos en datos más importantes del mundo, en donde se han llegado a publicar más de 44.000 (MIOTI, 2020) investigaciones profesionales relacionadas con el coronavirus.

Visualizaciones

La visualización de la información (Infographics o Data Visualization) es una disciplina relativamente nueva, que estudia como visualizar los datos de forma que la mera percepción genere conocimiento. Se usa visualización de información para explicar historias, simplificar, medir, comparar, explorar, descubrir (geographica.com, s.f.).

Instituciones, universidades, comunidades, empresas han creado decenas de visualizaciones de datos sobre la situación actual de la pandemia, tanto como información pública o simplemente como un interés personal con el ánimo de contribuir. Estos pueden ser desarrollados en lenguajes de programación como Python o con las aplicaciones de visualizaciones tales como PowerBI, Tableau, Qlik Sense, Prometeus, entre otras por nombrar ante una amplia variedad

A continuación, se presentan algunas visualizaciones que han destacado por su calidad y utilidad (MIOTI, 2020)

- Dashboard de CSSE de la Universidad Johns Hopkins

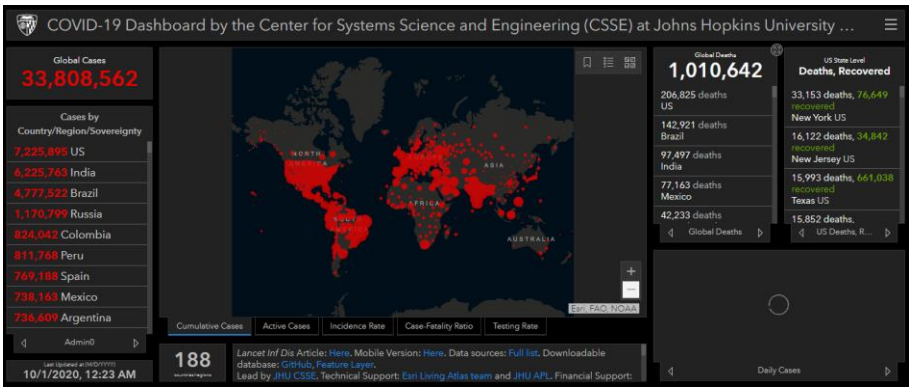

Ilustración 2- Dashboard de CSSE de la Universidad Johns Hopkins. Fuente: https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6 (Universidad Johns Hopkins, 2020)

- Dashboard del Proyecto Nextstrain

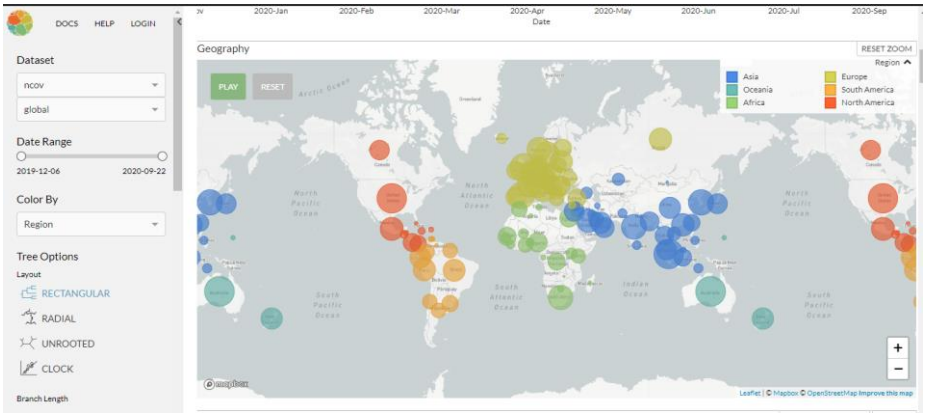

Ilustración 3- Dashboard del Proyecto Nextstrain. Fuente: (Proyecto Nextstrain, 2020)

- Dashboard creado por Avi Schiffmann

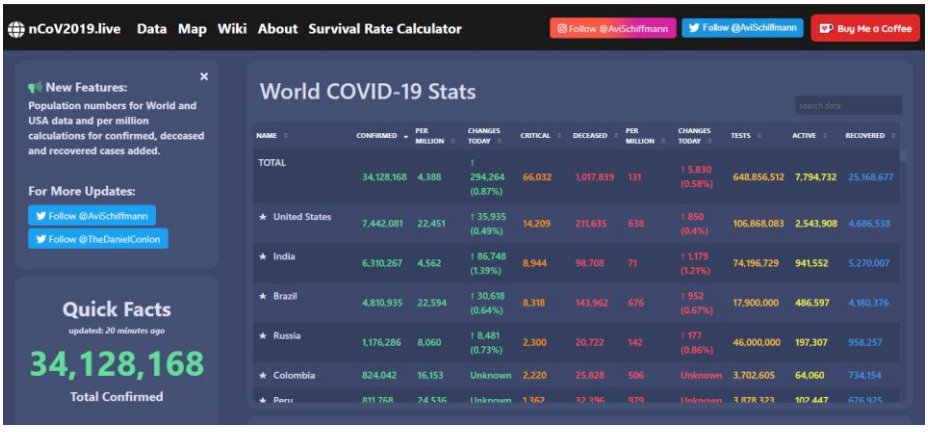

Ilustración 4 - Dashboard creado por Avi Schiffmann. Fuente: https://ncov2019.live/data (Avi Schiffmann, 2020)

Predecir su evolución (forecasting)

El descubrimiento de información, en un dominio de aplicación a través de las variables existentes, es una práctica que actualmente está en auge dado los altos volúmenes de datos que se manejan. Para esto las técnicas de aprendizaje automático resultan útiles en especial en aquellos casos donde todavía se cuenta con un conocimiento limitado como es la situación generada por la pandemia del Covid-19.

Técnicas de aprendizaje automático predictivo

El objetivo es predecir el valor particular de un atributo basado en otros atributos. El atributo a estimar es comúnmente llamado" clase" o variable dependiente, mientras que los atributos usados para hacer el pronóstico se llaman variables independientes, a continuación, se describen las principales técnicas:

Arboles de decisión: su funcionamiento general se basa en la aplicación de premisas que pueden ser cumplidas, o no, por un registro, el registro pasa a través del árbol de premisa en premisa hasta que se evalúa totalmente o hasta que encuentra un nodo terminal.

Las ventajas que tiene este algoritmo son:

- Fácil de entender: la salida del árbol es fácil de entender
- Útil en exploración de datos: los árboles es una de las formas para identificar variables significativas y la relación entre dos o mas
- Se requiere menos limpieza de datos: se requiere menos limpieza que otras técnicas, además que no se ve afectado de valores atípicos y faltantes en la data
- El tipo de datos no es una restricción: puede manejar variables numéricas y categóricas
- Método no paramétrico: esto quiere decir que no tienes suposiciones sobre la estructura del espacio y la estructura del clasificador

Algunas desventajas que se pueden presentar con esta técnica son:

- Sobreajuste: es una de las dificultades más comunes de este algoritmo, esto se puede resolver colocando restricciones en los parámetros del modelo y eliminando ramas en el análisis.
- Los modelos basados en árboles no están diseñados para funcionar con características muy dispersas. Cuando se trata de datos de entrada dispersos (por ejemplo, características categóricas con una gran dimensión), podemos preprocesar las características dispersas para generar estadísticas numéricas, o cambiar a un modelo lineal, que es más adecuado para dichos escenarios (Blog, Machine Learning, s.f.)

Redes Neuronales: consisten en "neuronas" o nodos interconectados que se organizan en capas, las redes neuronales aprenden en forma supervisada o no supervisada. En la modalidad supervisada, la red neuronal intenta predecir los resultados para ejemplos conocidos, compara sus predicciones con la respuesta objetivo y aprende de sus errores. Las redes neuronales supervisadas se emplean para pronósticos clasificación y modelos de series históricas. Una red neuronal de aprendizaje no supervisado es

eficaz para la descripción de datos, pero no para el pronóstico de resultados.

Las ventajas que tiene este algoritmo son:

- Aprendizaje: las RNA tienen la habilidad de aprender mediante la etapa de aprendizaje, donde se proporciona los datos como entrada y a su vez que se le indica cuál es la salida esperada
- Auto organización: una RNA crea su propia representación de la información en su interior
- Tolerancia a fallos: debido a que una RNA almacena la información de forma redundante, ésta puede seguir respondiendo de manera aceptable aun si se daña parcialmente.
- Flexibilidad: Una RNA puede manejar cambios no importantes en la información de entrada, como señales con ruido u otros cambios en la entrada

Algunas desventajas que se pueden presentar con esta técnica son:

- Complejidad de aprendizaje para grandes tareas, cuanto más cosas se necesiten que aprenda una red, más complicado será enseñarle
- Tiempo de aprendizaje elevado: esto depende de dos factores: primero si se incrementa la cantidad de patrones a identificar o clasificar y segundo si se requiere mayor flexibilidad o capacidad de adaptación de la red neuronal para reconocer patrones que sean sumamente parecidos, se deberá invertir más tiempo en lograr que la red converja a valores de pesos que representen lo que se quiera enseñar (Wikidot, s.f.).
- No permite interpretar lo que se ha aprendido, la red por si sola proporciona una salida, un número, que no puede ser interpretado por ella misma, sino que se requiere de la intervención del programador y de la aplicación en si para encontrarle un significado a la salida proporcionada (Wikidot, s.f.).
- Elevada cantidad de datos para el entrenamiento, cuanto más flexible se requiera que sea la red neuronal, más información tendrá que enseñarle para que realice de forma adecuada la identificación (Wikidot, s.f.).

Redes Bayesianas: se basa en el teorema estadístico de Bayes, el cual provee un cálculo para la probabilidad a posterior, una red bayesiana tiene una representación gráfica de dependencia entre un conjunto de atributos y se compone de dos elementos: un grafo acíclico que codifica la dependencia de las relaciones entre un conjunto de variables y una tabla de probabilidad aplicada a cada nodo

Las ventajas que tiene este algoritmo son:

- Es fácil y rápido predecir la clase de conjunto de datos de prueba
- Funciona bien en el pronóstico multiclase
- Cuando se mantiene la suposición de independencia, un clasificador Naive Bayes funciona mejor en comparación con otros modelos como la Regresión Logística y se necesitan menos datos de entrenamiento.

- Funciona bien en el caso de variables de entrada categóricas comparada con variables numéricas

Algunas desventajas que se pueden presentar con esta técnica son:

- Si la variable categórica tiene una categoría en el conjunto de datos de prueba, que no se observó en el conjunto de datos de entrenamiento, el modelo asignará una probabilidad de 0 y no podrá hacer un pronóstico . Esto se conoce a menudo como frecuencia cero. Para resolver esto, podemos utilizar la técnica de alisamiento (Blog, Machine Learning, s.f.).
- Otra limitación de Naive Bayes es la asunción de predictores independientes. En la vida real, es casi imposible que obtengamos un conjunto de predictores que sean completamente independiente (Blog, Machine Learning, s.f.)

Máquinas de soporte Vectorial (SVM): es un perceptrón (como una red neuronal) y es idealmente adecuado para la clasificación binaria de patrones que son linealmente separable, la idea principal de la SVM es obtener un único separador que maximice el margen entre la separación de dos clases.

Las ventajas que tiene este algoritmo son:

- Los clasificadores de Máquinas de Vectores de Soporte ofrecen una buena precisión y realizan predicciones más rápidas en comparación con el algoritmo de Naive Bayes (Blog, Machine Learning, s.f.)
- Utilizan menos memoria porque utilizan un subconjunto de puntos de entrenamiento en la fase de decisión
- Este algoritmo funciona bien con un claro margen de separación y con un espacio dimensional elevado (Blog, Machine Learning, s.f.).

Algunas desventajas que se pueden presentar con esta técnica son:

- Las Máquinas de Vectores de Soporte no son adecuadas para grandes conjuntos de datos debido a su alto tiempo de formación y también requiere más tiempo de formación en comparación con Naive Bayes.
- Funciona mal con clases superpuestas
- Es sensible al tipo de núcleo utilizado.

- ***Técnicas estadísticas de análisis multivariado para reducción de dimensionalidad PAC (Análisis de Componentes Principales)***

Los métodos de análisis multivariado se presentan de dos clases: los que suministran información de la interdependencia entre las variables y los que dan información sobre la dependencia entre una o más variables respecto a otras y otras (MONROY, 2007), el análisis de componentes principales se presenta como uno de los métodos de interdependencia.

El objetivo del análisis de componentes principales es transformar el conjunto de datos de variables originales en un conjunto más pequeño de variables, que retengan la mayor variabilidad contenida en los datos. Las nuevas variables poseen algunas características estadísticas "deseables", tales como independencia (bajo el supuesto de normalidad) y no correlación, utilizando este método se puede (MONROY, 2007):

  - Generar nuevas variables que expresen información contenida en un conjunto de datos
  - Reducir la dimensión del espacio donde están inscritos los datos
  - Eliminar las variables (si es posible) que aporten poco al estudio del problema
  - Facilitar la interpretación de la información contenida en los datos

En el caso de la no correlación entre las variables originales, el análisis de componentes principales no tiene mucho que hacer, pues las componentes se corresponderían con cada variable por orden de magnitud en la varianza; es decir, la primera componente coincide con la variable de mayor varianza, la segunda componente con la variable de segunda mayor varianza, y así sucesivamente

- **Aprendizaje Supervisado: Métodos de Clasificación con aprendizaje automático:**

  Para a aplicar un método de clasificación se requiere de la partición del conjunto de datos en dos conjuntos de datos más pequeños que serán utilizadas con los siguientes fines: entrenamiento y testing. El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de testing, se emplea para comprobar el comportamiento del modelo estimado (Parra, 2019).

  Cada registro de la base de datos debe de aparecer en uno de los dos subconjuntos, y para dividir el conjunto de datos en ambos subconjuntos, se

utiliza un procedimiento de muestreo: muestreo aleatorio simple o muestreo estratificado. Lo ideal es entrenar el modelo con un conjunto de datos independiente de los datos con los que realizamos el testing (Parra, 2019).

Como resultado de aplicar un método de clasificación, se cometerán dos errores, en el caso de una variable binaria que toma valores 0 y 1, habrá ceros que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros (Parra, 2019).

Un método para evaluar clasificadores, alternativo a la métrica expuesta, es la curva ROC (Receiver Operating Characteristic). La curva ROC es una representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos. La fracción de verdaderos positivos se conoce como sensibilidad, sería la probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo. La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea clasificado como negativo, esto es igual a restar uno de la fracción de falsos positivos (Parra, 2019).

La curva ROC también es conocida como la representación de sensibilidad (1-especificidad). Cada resultado de estimación representa un punto en el espacio ROC. El mejor método posible de estimación se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Una clasificación totalmente aleatoria daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación. En definitiva, se considera un modelo inútil, cuando la curva ROC recorre la diagonal positiva del gráfico. En tanto que, en un testing perfecto, la curva ROC recorre los bordes izquierdo y superior del gráfico. La curva ROC permite comparar modelos a través del área bajo su curva (Parra, 2019).

A continuación, se describen dos métodos de clasificación seleccionados para desarrollar el trabajo

- **Árbol de Decisión CART (Classification and Regression Trees)**

  El algoritmo CART es el acrónimo de Classification And Regression Trees (Árboles de Clasificación y de Regresión) fue diseñado por Breiman et al. (1984). Con este algoritmo, se generan árboles de decisión binarios, lo que quiere decir que cada nodo se divide en exactamente dos ramas

  Este modelo admite variables de entrada y de salida nominales, ordinales y continuas, por lo que se pueden resolver tanto problemas de clasificación como de regresión (Parra, 2019).

- **Algoritmo K-vecinos más cercanos KNN**

El método K-nn (K nearest neighbors Fix y Hodges, 1951) es un método de clasificación supervisada (Aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad F(x/Cj) de las predictora x por cada clase Cj (Parra, 2019).

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase Cj a partir de la información proporcionada por el conjunto de prototipos o ejemplos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras (Parra, 2019).

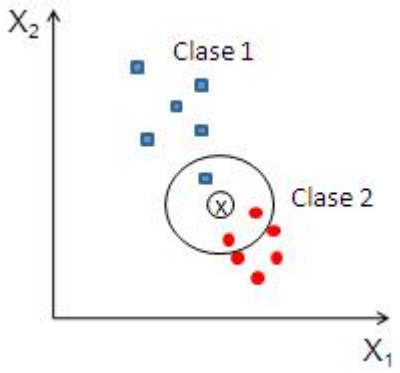

Ilustración 5 - K- vecinos. Fuente: https://bookdown.org/content/2274/metodos-de-clasificacion.html#algoritmo-k-vecinos-mas-cercanos (Parra, 2019)

La fase de entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de las clases de los ejemplos de entrenamiento. En la fase de testing, la evaluación del ejemplo (del que no se conoce su clase) es representada por un vector en el espacio característico. Se calcula la distancia entre los vectores almacenados y el nuevo vector, y se seleccionan los k ejemplos más cercanos. El nuevo ejemplo es clasificado con la clase que más se repite en los vectores seleccionados (Parra, 2019).

El método KNN supone que los vecinos más cercanos nos dan la mejor clasificación y esto se hace utilizando todos los atributos; el problema de dicha suposición es que es posible que se tengan muchos atributos irrelevantes que dominen sobre la clasificación, de manera que los atributos relevantes perderían peso entre otros veinte irrelevantes (Parra, 2019).

La mejor elección de k depende fundamentalmente de los datos; generalmente, valores grandes de k reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas. Un buen k puede ser seleccionado mediante un procedimiento de optimización. El caso especial en que la clase es predicha para ser la clase más cercana al

ejemplo de entrenamiento (cuando k=1) es llamada Nearest Neighbor Algorithm, Algoritmo del vecino más cercano (Parra, 2019).

- **Aprendizaje No Supervisado: Agrupación de Información (medidas de / distancia/proximidad)**

  Tanto las técnicas de reducción de dimensiones como las de agrupamiento, están basadas en determinar la semejanza (proximidad, similaridad) o disparidad (distancia, disimilaridad) existente; entre las variables las primeras, entre los individuos/variables las segundas (Parra, 2019).

  Lo primero a decidir será, pues, si optamos por centrar el análisis en medir disparidad o semejanza, lo cual dependerá en buena parte de los objetivos planteados en la investigación (Parra, 2019).

  Se llaman de distancia o de disimilaridad, porque cuanto mayor es el valor de la medida, mayor será la diferencia entre los individuos. Los distintos tipos que existen dependen de la escala en la que éstas estén formuladas. Cuando la variable está medida en una escala de intervalo las distancias más empleadas son las siguientes (Parra, 2019):

  Distancia Euclidea: Es la raíz cuadrada de la suma de las diferencias al cuadrado entre los dos elementos en la variable o variables consideradas.

  Distancia métrica de Chebychev: Es la distancia máxima en valores absolutos entre los valores de los elementos.

  Distancia de Manhattan o Bloque: La suma de las diferencias absolutas entre los valores de los elementos.

  Distancia de Minkowski: La raíz p-ésima de la suma de las diferencias absolutas elevada a la potencia p-ésima entre los valores de los elementos.

  Dentro de los métodos de agrupación de información se encuentra el análisis de clúster sobre el cual se puntualizará en el desarrollo del trabajo bajo las siguientes técnicas:

  - **Clustering Jerárquico (HAC)**

    En el análisis clúster jerárquico se parte del número de individuos (países, empresas etc.) y posteriormente se van uniendo en función de la mayor o menor proximidad de los individuos entre sí, formando grupos. Éstos a su vez se van uniendo entre sí hasta llegar a un único grupo. Las dos decisiones que existen son (Parra, 2019):

1. La determinación de la medida de distancia o proximidad a usar: como hemos visto, la opción entre una u otra vendrá determinada por la medida en que los datos estén referidos.

2. El método que determinará el modo de unión sucesiva de los distintos grupos entre sí. Es decir, el que determinará la distancia existente entre los sucesivos grupos. Entre ellos encontramos los siguientes:

- **Clustering de Densidad DBSCAN**

Los algoritmos para clustering basado en densidad identifican regiones de alta densidad que están rodeadas de ´áreas poco densas, cada una de las regiones densas identificadas se corresponde con un clúster. El clustering basado en densidad es apropiado cuando los clústeres no tienen una forma geométrica definida (Martin Ester, 1996).

Algoritmo DBSCAN: es un algoritmo de densidad simple que implementa la noción de densidad por medio de un procedimiento basado en centro, La densidad de cada punto se estima contando el número de puntos, incluido el, que se encuentran a una distancia no mayor que Eps del punto, dependiendo de dicho valor, cada puntos es clasificado como central (core), frontera (border) o ruido (noise) (Martin Ester, 1996).

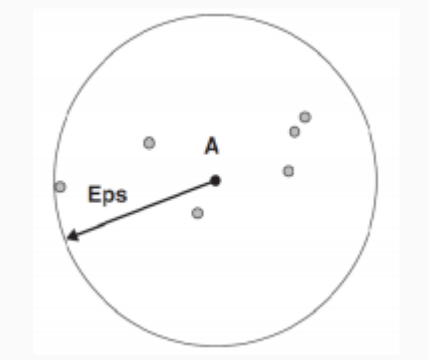

Ilustración 6 - Algoritmo DBSCAN - Densidad basada en centro. Fuente: https://dl.acm.org/doi/10.5555/3001460.3001507 (Martin Ester, 1996)

Trabajos entorno al descubrimiento del conocimiento de Covid-19 con ciencia de datos

La ciencia de datos y los grandes volúmenes de datos pueden utilizarse para comprender mejor los acontecimientos y lugares clave relacionados con el coronavirus, señalar su origen y medir la tasa de propagación. Tales técnicas pueden incluir el uso de análisis de datos impulsados por la ciencia de datos para obtener información del comportamiento en línea, como consultas de búsqueda en línea y conversaciones de medios sociales para identificar señales de una población específica que puedan proporcionar información sobre el movimiento del coronavirus.

- La empresa canadiense BlueDot, utiliza toneladas de datos para evaluar los riesgos para la salud pública. Utilizando el Procesamiento del Lenguaje Natural (NLP) y Machine Learning, navega y analiza unos 100.000 artículos en 65 idiomas diariamente para rastrear más de 100 enfermedades infecciosas. Fueron los primeros, por delante de los Centros de Control y Prevención de Enfermedades de los Estados Unidos, en notificar a sus clientes sobre el brote de coronavirus, en diciembre de 2019. Para predecir la propagación de la enfermedad, también utilizan datos como información sobre el itinerario del viajero, los medios sociales y los informes de noticias (AprendeIA, 2020)

- La Universidad Johns Hopkins ha estado usando Twitter para reunir información en tiempo real sobre dónde ocurren las enfermedades. Pero este tipo de datos es bastante ruidoso. Otros tipos de datos que pueden ser recopilados y analizados para predecir la propagación de enfermedades son las compras al por menor, los patrones de navegación y las palabras claves en los mensajes privados (AprendeIA, 2020)

- Por su parte, Baidu ha aprovechado los sistemas de mapeo impulsados por la Inteligencia Artificial para identificar el flujo de viajes a través de las zonas de alto riesgo utilizando la “Gran Plataforma de Datos de Migración” de Baidu Maps. Los movimientos de población de Wuhan, epicentro de la enfermedad, pueden rastrear ampliamente la propagación temprana del coronavirus. La Inteligencia Artificial está ayudando a los epidemiológicos a construir un cuadro aproximado de la migración de la gente con algunos portadores del coronavirus (AprendeIA, 2020)

- DeepMind, una subsidiaria de la empresa Google, compartió sus predicciones sobre las estructuras de la proteína del coronavirus, que generó usando su sistema AlphaFold. Asimismo, Baidu puso a disposición de los investigadores su algoritmo de inteligencia Artificial Linearfold para predecir la estructura del virus (AprendeIA, 2020)

- Atomwise, utiliza redes neuronales convolucionales que encuentran patrones en los datos de las pruebas que las personas nunca serían capaces de ver. Iktos, por su parte, utiliza redes neuronales generativas profundas para acelerar el proceso de descubrimiento de fármacos a través del diseño automático de moléculas virtuales con las características requeridas de un nuevo candidato a fármaco (AprendeIA, 2020)

4. Desarrollo del proyecto y resultados

4.1. Metodología

4.1.1. Arquitectura del proyecto

Para desarrollar el trabajo y poder estimar el número de nuevos casos positivos en la evolución del virus Covid-19 en la ciudad de Bogotá, capital de Colombia se ha definido la siguiente arquitectura de proyecto con las siguientes capas y metodología de desarrollo:

Capa de datos: Se obtienen fuentes de información desde la plataforma de datos abiertos de Colombia en los que se han publicado los conjuntos de datos con las variables requeridas para este proyecto tales como:

- El consolidado de los casos positivos de Coronavirus COVID-19 en Colombia reportados por el Instituto Nacional de Salud (INS)
- Pruebas PCR procesadas de COVID-19 en Colombia por los laboratorios
- Información de pasajeros de transporte masivo reportada por el ministerio de transporte de Colombia
- Consolidado de la cantidad de carga aérea que ha ingresado por cada uno de los aeropuertos del país desde el aislamiento reportada por el ministerio de transporte de Colombia
- información del delito de hurto en Colombia a través de las modalidades de Comercio y Motocicletas desde el 01 de enero del año 2010 al 31 de agosto del año 2020 reportada por la Policía Nacional de Colombia
- Información del delito de lesiones personales y lesiones en accidente de Tránsito desde el 01 de enero del año 2010 al 31 de agosto del año 2020

Ilustración 7 - Capa de datos. Fuente: creación propia para el proyecto

Capa de Información y Conocimiento: Luego de seleccionar las fuentes de información se define realizar el proceso extracción de la información, limpieza, transformación, construcción, selección de variables y preparación de datos de entrada al modelo con el entorno R que es un lenguaje de programación con enfoque estadístico.

Ilustración 8 - Capa de información y conocimiento. Fuente: Creación propia para el proyecto

Modelado: En este punto es donde la capa de datos y la capa de información y conocimiento cobra sentido ya que es aquí donde se integran las técnicas de analítica seleccionadas, para llevar a cabo este desarrollo utilizará el entorno de Python un lenguaje de programación abierto que cuenta con librerías para ciencia de datos

Ilustración 9 - Capa de Modelado. Fuente: Creación propia para el proyecto

Visualización: En esta capa esta de cara al usuario y es el punto donde se presenta gráficamente los resultados del análisis de datos y el resultado de la aplicación del modelo de analítica seleccionado. Para hacer esta representación gráfica se ha elegido el entorno Qlik Sense que es una plataforma de análisis de datos y visualizaciones

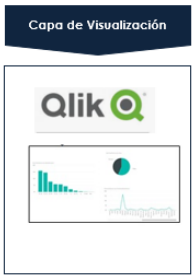

Ilustración 10 - Capa de visualización. Fuente: Creación propia para el proyecto

Metodología CRISP-DM

Transversalmente a la arquitectura del proyecto se utilizará los estándares de la metodología CRISP-DM para proyectos de analítica que apoyará el entendimiento, análisis y desarrollo del trabajo a través de sus distintas fases:

Entendimiento del negocio o situación: es la fase donde se evalúa la situación, requisitos y supuestos y el tipo de datos con el que se cuenta para el análisis y determinación de objetivos.

Entendimiento de los datos: la comprensión de datos implica tener disponible los datos y explorarlos para determinar la calidad de los datos.

Preparación de los datos: selección y limpieza de datos, construcción de nuevos datos, integración de datos y validar el formato de los datos que se usarán como variables para el modelo.

Modelado: selección de modelo, construcción y pruebas.

Evaluación: durante esta fase se realiza el análisis de los resultados, se determina si se han realizado descubrimientos especiales o particulares relevantes para resaltar y se determina si el modelo utilizado es viable o no.

En la ilustración 5 - Modelo de Arquitectura del proyecto se presenta la arquitectura del proyecto seleccionada:

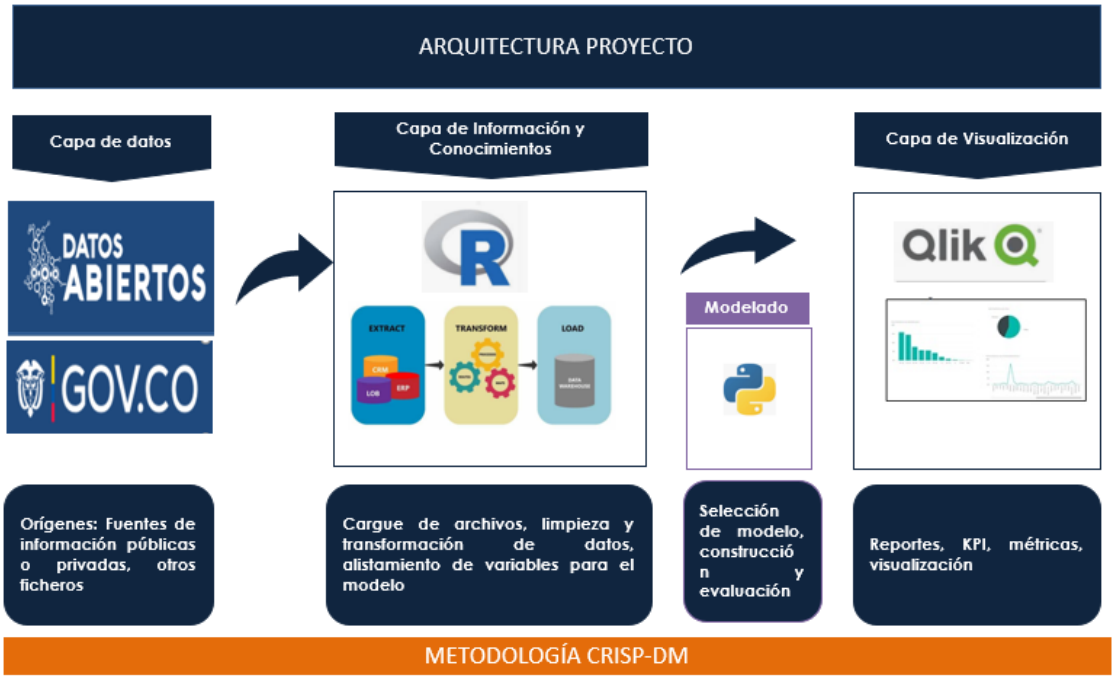

Ilustración 11 - Modelo de Arquitectura del proyecto. Fuente propia para el proyecto

4.1.2. Entendimiento del negocio o situación

El 6 de marzo Colombia se sumó a los 191 países que confirmaron la presencia del virus Covid-19 en su población. Un mes después del primer contagio en el país, se habían reportado más de 1400 casos en 24 departamentos, 35 muertes, 88 pacientes recuperados y más de 20 mil casos descartados. El 4.3% (Llorente y Cuenca, 2020)del total de los pacientes diagnosticados con el virus se encontraban en Unidades de Cuidado Intensivo mientras que el 82% (Llorente y Cuenca, 2020) están en sus casas. Hasta esa primera etapa la red hospitalaria había respondido de acuerdo con su capacidad instalada sin verse aún, colapsada. Igualmente, el 80% (Llorente y Cuenca, 2020) de los eventos se concentran en Bogotá, Valle del Cauca, Antioquia, Cundinamarca y Atlántico y hasta ese entonces, según las cifras reportadas, la región de la Orinoquia era la que menos diagnósticos de coronavirus había registrado

Desde que llegó el virus a Colombia, las miradas han estado dirigidas a las medidas adelantadas por el Gobierno Nacional ante un escenario sorpresivo, de incertidumbre y que no estaba contemplado en la política pública. Los planes sobre la marcha, la poca información respecto al virus y la situación que estaban viviendo países como Italia y España, trasladaron la conversación a las recomendaciones técnicas y científicas de la Organización Mundial de la Salud que fueron acogidas y lideradas por el Ministerio de Salud y Protección Social y el Instituto Nacional de Salud (INS) (Llorente y Cuenca, 2020)

Contexto

Al inicio de la pandemia los epidemiólogos y científicos de datos trabajan en tratar de establecer la curva de evolución del virus. El problema era que se enfrentaban a una situación nueva por las características de un virus desconocido y que las medidas para controlar la propagación no tienen precedentes. Al pasar los meses y ante la evolución en la adquisición del conocimiento de la enfermedad, se plantean nuevos retos como averiguar si producirá un rebrote importante y cuándo llegará, y aunque hay avances y equipos interdisciplinares que continúan trabajando para conocer mejor el comportamiento del virus, siguen en el ambiente grandes incertidumbres.

Muchos expertos coinciden en que estimar la evolución de la pandemia merece el esfuerzo ya que esto aporta información vital para la preparación de los servicios sanitarios. Es por ello, por lo que, tanto para realizar predicciones como para la lucha de la pandemia de forma efectiva, resulta crucial identificar y rastrear nuevos casos.

Análisis FODA – Fortalezas, Oportunidades, Debilidades, Amenazas

Fortalezas	Oportunidades
• Actualmente se cuenta con más herramientas • Se conoce mejor el comportamiento del virus • Se tiene visión de lo que está pasando en todo el mundo y la reacción de la sociedad	• Investigaciones promovidas por las diferentes universidades y empresas • A pesar de que el covid-19 tenga características únicas, hay patrones de otras enfermedades que se repiten • La capacidad para realizar test y rastrear contactos
Debilidades	**Amenazas**
• Esta situación es nueva y caótica • Aún existe grandes incertidumbres sobre el virus • La variabilidad de los datos oficiales complica la labor de análisis y predicciones • Se vuelve complejo rastrear casos, uno de los puntos importantes solicitados por OMS para la lucha del virus • Es difícil anticipar un rebrote de coronavirus con mucha antelación	• La dinámica de propagación de un virus humano ha cambiado • La posibilidad que se producirá un rebrote importante • Existen personas asintomáticas que pueden estar infectado a otras • Los científicos han establecido que existen “super-contagiadores” (Carlos Eduardo Álvarez Cabrera, 2015) • La gente está más conecta e informada y se genera pánico más

Definición del problema

Predecir nuevos casos teniendo en cuenta diferente variables del entorno social, con el fin de poder anticipar un rebrote o un incremento relevante en su volumen en la ciudad Bogotá, y de cada una de sus localidades para así aportar conocimiento al sistema de salud y que este se prepare para atender más casos, o por conocimiento de los ciudadanos y que éstos tomen medidas que refuercen la protección.

4.1.3. Entendimiento de los datos

Para el análisis se obtiene información de los datos abiertos de fuentes públicas de Colombia, seleccionando los siguientes conjuntos de datos, para realizar el pronóstico de nuevos casos:

Tabla 1 - Inventario de fuentes seleccionadas para análisis de datos

No	Grupo	Descripción del conjunto de datos	Fuente
1	Salud	Casos_positivos_de_COVID-19_en_Colombia.csv	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data
2	Salud	Pruebas PCR procesadas de COVID-19 en Colombia (Departamental)	https://www.datos.gov.co/Salud-y-Protecci-n-Social/Pruebas-PCR-procesadas-de-COVID-19-en-Colombia-Dep/8835-5baf/data
3	Salud	Casos confirmados de COVID-19 en los Trabajadores del sector salud en Bogotá D.C.	http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/salud-laboral/covid-19-trabajadores-salud/
4	Salud	Número de Reproducción Efectivo (Rt)	http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/covid19/
5	Salud	Porcentaje de ocupación de Hospitalización general para la atención del COVID-19	http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/ocupacion-ucis/
6	Salud	Porcentaje de uso de Unidades Cuidado Intensivo para la atención del COVID-19	http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/ocupacion-ucis/
7	Salud	Disposición de cadáveres por COVID – 19 en Bogotá D.C.	http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/enfermedades-trasmisibles/disposicion-de-cadaveres-por-covid-19/
8	Clima	Temperatura mínima diaria	http://dhime.ideam.gov.co/atencionciudadano/
9	Clima	Temperatura máxima diaria	http://dhime.ideam.gov.co/atencionciudadano/
10	Transporte	Información Pasajeros Transporte Masivo en Colombia	https://www.datos.gov.co/Transporte/Informaci-n-Pasajeros-Transporte-Masivo/2h8t-2zik

11	Transporte	Informe de Carga Aérea	https://www.datos.gov.co/Transporte/Informe-de-carga-a-rea/4wwa-qb9a
12	Transporte	Impacto de la congestión del tráfico por coronavirus en América Latina	https://github.com/EL-BID/IDB-IDB-Invest-Coronavirus-Impact-Dashboard
13	Seguridad	Reporte Hurto por Modalidades Policía Nacional	https://www.datos.gov.co/Seguridad-y-Defensa/Reporte-Hurto-por-Modalidades-Polic-a-Nacional/6z45-aexg
13	Seguridad	Lesiones Personales y Lesiones en accidente de tránsito	https://www.datos.gov.co/Seguridad-y-Defensa/Reporte-Lesiones-Personales-y-Lesiones-en-accident/673k-7cs2
15	Finanzas	Tasa de cambio representativa del mercado (TRM)	https://totoro.banrep.gov.co/analytics/saw.dll?Go&Action=prompt&Path=%2fshared%2fSeries%20Estad%C3%ADsticas_T%2f1.%20Tasa%20de%20Cambio%20Peso%20Colombiano%2f1.1%20TRM%20-%20Disponible%20desde%20el%2027%20de%20noviembre%20de%201991%2f1.1.2.TCM_Para%20rango%20de%20fechas%20dado&Options=rdf&lang=es
16	Finanzas	Tasa de intervención Banco de la República (Tasa de intervención de política monetaria)	http://www.superfinanciera.gov.co/

Conjunto de Datos 1: Casos positivos de COVID-19 en Colombia

Contiene el consolidado de los casos positivos de Coronavirus COVID-19 en Colombia reportados por el Instituto Nacional de Salud (INS). Incluye variables como género, departamento, grupo etario, entre otras.

Aspectos para tener en cuenta en el conjunto de datos:

- Para las ciudades que son distritos (Cartagena, Bogotá, Santa Marta, Buenaventura y Barranquilla), sus cifras son independientes a las cifras del departamento al cual pertenecen, en concordancia con la división oficial de Colombia.

- Los casos marcados como en estudio están sujetos a modificación una vez se identifique el origen (importado o relacionado)
- Recuperado es paciente con segunda prueba negativa para el virus
- El paciente puede permanecer en el hospital por otras razones.
- Por seguridad de las personas, algunos datos serán limitados evitando así la exposición y posible identificación en determinados municipios

Motivaciones para la selección del conjunto de datos:

El propósito de esta fuente de datos, descrita en al Tabla 2, es proporcionar al conjunto de datos de estudio la variable dependiente a estimar: el número de casos positivos de Covid-19 en la ciudad de Bogotá.

Tabla 2 - Estructura de datos - Casos Positivos Covid19 Colombia. Fuente: Datos abiertos Colombia

Nombre del Campo	Tipo de dato	Descripción
Id Caso	texto	Número asignado
Fecha Notificación	texto	Fecha de notificación a SIVIGILA
Código Divipola	texto	Código estándar de la división política de Colombia
Ciudad de Ubicación	texto	Nombre de la ciudad
Departamento	texto	Nombre del departamento
Atención	texto	Corresponde a muertes no relacionadas con COVID-19, aún si eran casos activos **Hay pacientes recuperados para COVID-19, que pueden permanecer en hospitalización por otras comorbilidades
Edad	texto	
Sexo	texto	
Tipo	texto	Identifica si el caso es importado o local
Estado	texto	
País de procedencia	texto	
Fis	texto	Fecha de inicio de síntomas
Fecha de muerte	texto	
Fecha de diagnostico	texto	

Fecha recuperado	texto	
Fecha reporte web	texto	Fecha de publicación en sitio web
Tipo de recuperación	texto	Se refiere a la variable de tipo de recuperación que tiene dos opciones: PCR y tiempo. PCR indica que la persona se encuentra recuperada por segunda muestra, en donde dio negativo para el virus; mientras que tiempo significa que son personas que cumplieron 30 días posteriores al inicio de síntomas o toma de muestras que no tienen síntomas, que no tengan más de 70 años ni que estén hospitalizados.
Código del departamento	texto	Código estándar asignado en la división política de Colombia
Código del país	texto	Código estándar de identificación de los países
Nombre del grupo étnico	texto	La variable etnia depende totalmente de tres cosas: - El correcto diligenciamiento de la variable Etnia por los profesionales de salud que notifican en más de 10.000 instituciones de salud en todos los municipios y departamentos. - Del autorreconocimiento de la persona cuando se le pregunta por esta variable. - Del listado censal que haga y mantenga actualizado cada departamento. No depende del Instituto Nacional de Salud, y por lo tanto, es responsabilidad de las autoridades de cada municipio, departamento y distrito de Colombia; la calidad y consistencia de dicha variable
Ubicación del recuperado	texto	Algunos casos recuperados permanecen en hospitalización por otras comorbilidades. Se discrimina su ubicación independiente del estado.

Conjunto de datos 2: Pruebas PCR procesadas de COVID-19 en Colombia (Departamental)

Corresponde a la información que se ha incorporado al sistema SisMuestras que entró en funcionamiento a partir del 15 de abril de 2020, recopilando y reportando la información que cargan los laboratorios pertenecientes a la red. Los valores acumulados pueden variar de acuerdo con la información histórica que se ajuste. La información cargada corresponde a lo dispuesto por otros actores diferentes al INS, por tal razón, existe la columna de "Procedencia desconocida". Las muestras cargadas en esta columna serán asignadas a sus departamentos una vez se corrobore la información.

Motivaciones para la selección del conjunto de datos:

El número de pruebas PCR que se realizan a los ciudadanos, permite detectar con antelación casos que estén circulando. Por esta razón es probable que el número de casos encontrado dependa fuertemente del número de pruebas realizadas, y desde el sector salud, resulta una variable de interés para involucrar al estudio, como variable independiente.

Tabla 3 – Estructura de datos de Pruebas PCR procesadas. Fuente: Datos abiertos Colombia

Nombre del Campo	Tipo de dato	Descripción
Fecha	text	Fecha de reporte según corte
Acumuladas	Numérico	Número de PCR acumuladas
Amazonas	Numérico	Pruebas PCR procesadas
Antioquia	Numérico	Pruebas PCR procesadas
Arauca	Numérico	Pruebas PCR procesadas
Atlántico	Numérico	Pruebas PCR procesadas
Bogotá	Numérico	Pruebas PCR procesadas
Bolívar	Numérico	Pruebas PCR procesadas
Boyacá	Numérico	Pruebas PCR procesadas
Caldas	Numérico	Pruebas PCR procesadas
Caquetá	Numérico	Pruebas PCR procesadas
Casanare	Numérico	Pruebas PCR procesadas
Cauca	Numérico	Pruebas PCR procesadas
Cesar	Numérico	Pruebas PCR procesadas

Choco	Numérico	Pruebas PCR procesadas
Córdoba	Numérico	Pruebas PCR procesadas
Cundinamarca	Numérico	Pruebas PCR procesadas
Guainía	Numérico	Pruebas PCR procesadas
Guajira	Numérico	Pruebas PCR procesadas
Guaviare	Numérico	Pruebas PCR procesadas
Huila	Numérico	Pruebas PCR procesadas
Magdalena	Numérico	Pruebas PCR procesadas
Meta	Numérico	Pruebas PCR procesadas
Nariño	Numérico	Pruebas PCR procesadas
Norte de Santander	Numérico	Pruebas PCR procesadas
Putumayo	Numérico	Pruebas PCR procesadas
Quindío	Numérico	Pruebas PCR procesadas
Risaralda	Numérico	Pruebas PCR procesadas
San Andrés	Numérico	Pruebas PCR procesadas
Santander	Numérico	Pruebas PCR procesadas
Sucre	Numérico	Pruebas PCR procesadas
Tolima	Numérico	Pruebas PCR procesadas
Valle del Cauca	Numérico	Pruebas PCR procesadas
Vaupes	Numérico	Pruebas PCR procesadas
Vichada	Numérico	Pruebas PCR procesadas
Procedencia desconocida	Numérico	Pruebas PCR procesadas
Barranquilla	Numérico	Pruebas PCR procesadas
Cartagena	Numérico	Pruebas PCR procesadas
Santa Marta	Numérico	Pruebas PCR procesadas
Positiva Acumuladas	Numérico	Corresponden a muestras que tienen resultado positivo para PCR - COVID19, pero que incluyen segundas (o más

		pruebas) realizadas a casos activos, que aún no se recuperan.
Negativas Acumuladas	Numérico	Corresponden a las muestras con resultado negativo en PCR - COVID19 en el país; a diferencia de los casos descartados que corresponden a las personas que no tienen el virus.
Positividad Acumulada	Texto	Valor obtenido entre muestras positivas acumuladas / muestras procesadas acumuladas
Indeterminadas	Numérico	Requieren nueva muestra para obtener resultado

Conjunto de datos 3: Casos confirmados de COVID-19 en los Trabajadores del sector salud en Bogotá D.C

Motivaciones para la selección del conjunto de datos:

Otra variable del sector salud que pareciera tener una relación directa con la presencia de casos nuevos, es el número de casos nuevos en personas del sector salud, extraído de la fuente descrita en la Tabla 4. Si los trabajadores de la salud terminan siendo víctimas del virus, al estar prestando sus servicios a la ciudadanía, es probable que funcionen como dispersadores del virus y por tanto tengan influencia en la propagación.

Tabla 4 - Estructura de datos - Casos confirmados de Covid-19 en trabajadores de sector salud en Bogotá

Nombre del Campo	Tipo de dato	Descripción
Caso	texto	Número consecutivo del caso de acuerdo con el orden de registro
Fecha de inicio de síntomas	texto	Fecha de inicio de síntomas de acuerdo con información del trabajador
Fecha de Diagnóstico	texto	Fecha de diagnóstico medica
Edad	texto	Edad del trabajador
Sexo	texto	Género: Masculino o femenino
Localidad de Residencia	texto	Localidad de la ciudad de Bogotá

Fuente de contagio	texto	No asociado a la prestación del servicio/ Asociado a la prestación de servicios/ Importado
Ubicación	texto	Lugar físico donde se encuentra el trabajador luego del diagnóstico: Hospital/ Casa
Estado	texto	Recuperado, Fallecido, Leve, Moderado, Leve. Grave
Ocupación	texto	Profesión u oficio a la que se dedica el trabajador
Servicio	texto	Ubicación por donde ingresa el trabajador a consulta

Conjunto de datos 4: Número de Reproducción Efectivo (Rt)

Motivaciones para la selección del conjunto de datos:

El Rt también conocido como número de reproducción efectiva es un valor que mide el número de infectados secundarios por un infectado primario, en función del tiempo. Este número es sensible a las restricciones o aperturas que se dan en una población y cambia en el tiempo según la intensidad y duración de cada medida por lo que conocer su valor es esencial para el control y la vigilancia epidemiológica (Instituto Nacional de Salud (INS), s.f.)

Si este número aumenta, implica un incremento del número de casos nuevos, siendo entonces un candidato para estimarlos.

Tabla 5 - Estructura de datos - Número de reproducción Efectivo (Rt)

Nombre del Campo	Tipo de dato	Descripción
Fecha Inicio de Ventana	texto	Fecha Inicio de registro
Fecha_Fin_Ventana	texto	Fecha fin de registro
Mean(R)	Numérico	Mediana del número de reproducción efectivo
Std(R)	Numérico	Desviación Estándar del número de reproducción efectivo

Cuantil.0.025(R)	Numérico	Cuartil 1
Cuantil.0.05(R)	Numérico	Cuartil 2
Cuantil.0.25(R)	Numérico	Cuartil 3
Median(R)	Numérico	Mediada del número de reproducción efectivo
Cuantil.0.75(R)	Numérico	Cuartil 4
Cuantil.0.95(R)	Numérico	Cuartil 5
Cuantil.0.975(R)	Numérico	Cuartil 6
Localidad	Texto	Nombre de la localidad de Bogotá

Conjunto de datos 5: Porcentaje de ocupación de Hospitalización general para la atención del COVID-19

Motivaciones para la selección del conjunto de datos:

El número de camas disponibles para atención del Covid, es un recurso del sector salud que a medida que se vea ocupado por más pacientes, indicará un mayor aumento de casos. Por eso resulta de interés para ser parte del conjunto de variables independientes para la estimación del trabajo.

Tabla 6 - Estructura de datos - Porcentaje de ocupación hospitalización general para atención Covid-19 en Bogotá

Nombre del Campo	Tipo de dato	Descripción
Fecha	texto	Fecha de información
Camas asignadas COVID 19	Numérico	Número de camas asignadas en hospitales de Bogotá
Camas ocupadas COVID 19	Numérico	Número de camas ocupadas en hospitales de Bogotá
Ocupación	Numérico	Porcentaje de ocupación

Conjunto de datos 6: Porcentaje de uso de Unidades Cuidado Intensivo para la atención del COVID-19

Motivaciones para la selección del conjunto de datos:

Al igual que el número de camas, el nivel de ocupación de Unidades de Cuidado intensivo indicará un mayor número de casos de congio, y por consecuencia un mayor riesgo a contagios nuevos.

Tabla 7- Estructura de datos- Porcentaje de uso Unidades de Cuidados Intensivos para atención de Covid-19

Nombre del Campo	Tipo de dato	Descripción
Fecha	texto	Fecha de información
Camas UCI ocupadas Covid-19	Numérico	Número de camas asignadas en hospitales de Bogotá
Camas UCI Disponibles COVID 19	Numérico	Número de camas ocupadas en hospitales de Bogotá
Ocupación UCI	Numérico	Porcentaje de ocupación en UCI

Conjunto de datos 7: Disposición de cadáveres por COVID – 19 en Bogotá D.C

Motivaciones para la selección del conjunto de datos:

La variable de porcentaje de ocupación de salas crematorias para cadáveres de víctimas del Covid-19, también resulta de interés para ser parte del conjunto de datos final, ya que si hay mayor ocupación de dichas salas, es porque hay mayor volumen de mortalidad por el virus y esto puede indicar un riesgo alto de mayor número de contagios.

Tabla 8 - Estructura de datos - Disposición de cadáveres por Covid-19 en Bogotá

Nombre del Campo	Tipo de dato	Descripción
Fecha	texto	Fecha de información
Licencias cremación NO COVID - 19	Numérico	Número de licencias por causas diferentes a Covid-19
Licencias por COVID - 19	Numérico	Número de licencias crematorios por Covid-19

Capacidad servicios crematorios	Numérico	Número de cupos para servicios de crematorios
ocupación servicios crematorios	Numérico	Porcentaje de ocupación servicios crematorio

Conjunto de datos 8: Temperatura mínima diaria

Motivaciones para la selección del conjunto de datos:

Considerando variables climáticas como posibles factores colaboradores en la expansión del virus, se selecciona la variable de temperatura media mínima en la ciudad de Bogotá para ser parte del conjunto de variables del estudio.

Tabla 9 - Estructura de datos - Temperatura mínima diaria

Nombre del Campo	Tipo de dato	Descripción
CodigoEstacion	texto	Código de la estación meteorológica
NombreEstacion	texto	Nombre de la estación
Latitud	texto	Coordenada geográfica
Longitud	texto	Coordenada geográfica
Altitud	texto	Coordenadas selenográficas
Categoria	texto	Climática Ordinaria
Entidad	texto	Instituto de Hidrologia Meteorologia y Estudios Ambientales
AreaOperativa	texto	Área geográfica sobre la que se toma la medida
Departamento	texto	Bogotá
Municipio	texto	Bogotá
FechaInstalacion	texto	Fecha de instalación de la estación
FechaSuspension	texto	Fecha de suspensión de la estación
IdParametro	texto	Temperatura
Etiqueta	texto	Código de etiqueta asociado al parámetro
DescripcionSerie	texto	Temperatura mínima diaria

Frecuencia	texto	Diario
Fecha	texto	Fecha de información
Valor	texto	
Grado	texto	
Calificador	texto	
NivelAprobacion	texto	

Conjunto de datos 9: Temperatura máxima diaria

Motivaciones para la selección del conjunto de datos:

Al igual que el conjunto de datos anterior, este conjunto de datos corresponde a la una variable climática, en este caso la temperatura media máxima en la ciudad de Bogotá.

Tabla 10 - Estructura de datos - Temperatura máxima diaria

Nombre del Campo	Tipo de dato	Descripción
CodigoEstacion	texto	Código de la estación meteorológica
NombreEstacion	texto	Nombre de la estación
Latitud	texto	Coordenada geográfica
Longitud	texto	Coordenada geográfica
Altitud	texto	Coordenadas selenográficas
Categoria	texto	Climática Ordinaria
Entidad	texto	Instituto de Hidrologia Meteorologia y Estudios Ambientales
AreaOperativa	texto	Área geográfica sobre la que se toma la medida
Departamento	texto	Bogotá
Municipio	texto	Bogotá
FechaInstalacion	texto	Fecha de instalación de la estación
FechaSuspension	texto	Fecha de suspensión de la estación

IdParametro	texto	Temperatura
Etiqueta	texto	Código de etiqueta asociado al parametro
DescripcionSerie	texto	Temperatura máxima diaria
Frecuencia	texto	Diario
Fecha	texto	Fecha de información
Valor	texto	
Grado	texto	
Calificador	texto	
NivelAprobacion	texto	

Conjunto de datos 10: Información Pasajeros Transporte Masivo

Conjunto de datos que contiene información relacionada con la cantidad de pasajeros que se transportan en el transporte masivo en Colombia.

Motivaciones para la selección del conjunto de datos:

Aunque la ciudad desde los primeros días de la aparición del Covid-19, entró en periodo de cuarentena estricta, muchos sectores de servicios básicos, mensajería y sanitarios; siguieron funcionando.

Muchos de los trabajadores de estos sectores, estando en riesgo, debían movilizarse por la ciudad en trasporte público. Por esta razón el volumen de personas que se movilizan a diario en el transporte masivo, puede ser un foco de contagios, siendo entonces este número un buen candidato a variable independiente para la estimación del estudio.

Tabla 11 -Estructura de datos de Información Pasajeros Transporte Masivo. Fuente: Datos abiertos Colombia

Nombre del Campo	Tipo de dato	Descripción
Fecha	texto	Fecha de información
Ciudad	texto	Ciudad de acuerdo con la división política de Colombia
Sistema	texto	Nombre del sistema de transporte de la ciudad
Pasajeros/día	Numérico	Número de pasajeros reportados por día

Variación Transmilenio	texto	
Pasajeros día típico laboral	Numérico	Número pasajeros en días laborales (lunes a viernes, días hábiles)
Pasajeros día sábado	Numérico	Número pasajeros sábados
Pasajero día festivo	Numérico	Número de pasajeros en días festivos
Días Semana	Numérico	Días de la semana analizados

Conjunto de datos 11: Informe de Carga Aérea

Conjunto de datos que contiene información de la cantidad de carga aérea que ha ingresado por cada uno de los aeropuertos del país desde el aislamiento

Motivaciones para la selección del conjunto de datos:

Como otra de las variables del sector transporte, entendiendo que éste seguía activo y que podría ser un foco de contagio; están los vuelos de carga, humanitarios y de ambulancias, resumidos en la variable cantidad total de vuelos. Esta variable es entonces de interés para el estudio del trabajo, como una posible influenciadora al número de casos nuevos.

Tabla 12 - Estructura de datos - Informe de Carga Aérea. Fuente: Datos abiertos Colombia

Nombre del Campo	Tipo de dato	Descripción
Fecha	texto	Fecha de información
Regiones	texto	Regiones de Colombia de acuerdo con la división política de Colombia
Departamento	texto	Departamento de Colombia
Latitud	texto	Coordenadas geográficas
Longitud	texto	Coordenadas geográficas
Aeropuerto	texto	Nombre del aeropuerto del cual se obtienen los datos
Carga llegada nacionales	numérico	Número de cargas
Carga Salidas Nacionales	numérico	Número de cargas
Carga llegadas internacionales	numérico	Número de cargas

Carga salidas internacionales	texto	Número de cargas
Vuelos carga	numérico	Número de vuelos
Vuelos ambulancia	numérico	Número de vuelos
Total Vuelos	numérico	Total de vuelos
Total carga transporte toneladas	numérico	Total de la carga
Carga vuelos nacionales toneladas	numérico	Número de toneladas de carga
Carga vuelos internacionales toneladas	texto	Número de toneladas de carga

Conjunto de datos 12: Impacto de la congestión del tráfico por coronavirus en América Latina – Conjuntos de datos en idioma ingles

Los conjuntos de datos tienen como objetivo rastrear una variedad de variables de interés con el fin de proporcionar a los legisladores, epidemiólogos y el público en general de la región medidas del impacto que las restricciones y recomendaciones de "distanciamiento social" debido al brote de coronavirus están teniendo en la población y sobre la actividad económica.

Motivaciones para la selección del conjunto de datos:

De este conjunto, otra variable representativa del sector transporte, el % de trafico respecto a la primera semana del mes de marzo al inicio de la pandemia; es una variable que habla del movimiento de personas en la ciudad en los diferentes medios de transporte, incluyendo medio particular. Por esta razón podría ser una variable explicativa del volumen de casos nuevos: a mayor variación porcentual, quiere decir que el tráfico respecto al tráfico normal antes de la cuarentena, aumenta y por consecuencia aumenta el riesgo de contagio.

Tabla 13 - Estructura de Datos - Diario Impacto de la congestión del tráfico por coronavirus en América Latina

Column Name	Description	Type
last_updated_utc	Last updated date in UTC time	timestamp
region_slug	Region unique name	string
region_name	Region human readable name	string
country_name	Country name in english	string
country_iso_code	Country code in ISO-2 standard	string
country_idb_code	Country code in IDB standards	string
region_type	Region type, e.g. city, country	string
population	Population of the region	int

timezone	Timezone of the region	string
month	Month in 2020	int
day	Day in 2020.	int
dow	Day of the week	int
ratio_20	Percentage change in Traffic Congestion Intensity (TCI) is change_TCI = (ratio_20 - 1) * 100	float
tcp	Percentage change in Traffic Congestion Intensity (TCI)	float

Conjunto de datos 13: Reporte Hurto por Modalidades Policía Nacional

En este conjunto de datos la ciudadanía puede encontrará información del delito de hurto en Colombia a través de las modalidades de Comercio y Motocicletas desde el 01 de enero del año 2010 al 31 de agosto del año 2020.

Fuente: DIJIN - Policía Nacional. Datos extraídos el día 08 de septiembre del año 2020 a las 14:00 horas. Cifras sujetas a variación, en proceso de integración y consolidación con información de fiscalía general de la nación.

Motivaciones para la selección del conjunto de datos:

Las conductas delictivas de y comportamiento al margen de una vida pacífica en una ciudad metropolitana, ni tienen vacaciones o excusas para descansar. Es por esto por lo que es común que ante eventos tales como el presente Covid-19 aún se presenten casos de inseguridad, hurto, violencia y lesiones personales; como lo presenta esta fuente de datos y la siguiente.

Buscando el aporte de otros sectores al estudio, la seguridad, expresada en número de hurtos y lesiones personales, se convierten en variables de interés, ya que estos eventos requieren de algún tipo de contacto entre víctimas y victimarios, y es posible que estos momentos sean focos de propagación del virus.

Tabla 14 - Estructura de datos -Reporte Hurto por Modalidades Policía Nacional . Fuente: Datos abiertos de Colombia

Nombre del Campo	Tipo de dato	Descripción
Departamento	texto	Nombre del departamento
Municipio	texto	Nombre del municipio
Código DANE	texto	Código estándar del Departamento Administrativo Nacional de Estadística DANE
Armas Medios	texto	Tipo de artefacto con el que se produce la agresión
Fecha Hecho	texto	Fecha en la que ocurrió el evento

Genero	texto	Identifica el género de la persona atacada
Grupo Etario	texto	Grupo por edades
Tipo de Hurto	Texto	Clasificación del Hurto
Cantidad	texto	Número hurtos reportados

Conjunto de datos 14: Reporte Lesiones Personales y Lesiones en accidente de tránsito Policía Nacional

En este conjunto de datos la ciudadanía puede encontrará información del delito de Lesiones personales y Lesiones en accidente de Tránsito desde el 01 de enero del año 2010 al 31 de agosto del año 2020

Fuente: DIJIN - Policía Nacional. Datos extraídos el día 08 de septiembre del año 2020 a las 14:00 horas. Cifras sujetas a variación, en proceso de integración y consolidación con información de fiscalía general de la nación

Motivaciones para la selección del conjunto de datos:

Las conductas delictivas de y comportamiento al margen de una vida pacífica en una ciudad metropolitana, ni tienen vacaciones o excusas para descansar. Es por esto por lo que es común que ante eventos tales como el presente Covid-19 aún se presenten casos de inseguridad, hurto, violencia y lesiones personales; como lo presenta esta fuente de datos y la siguiente.

Buscando el aporte de otros sectores al estudio, la seguridad, expresada en número de hurtos y lesiones personales, se convierten en variables de interés, ya que estos eventos requieren de algún tipo de contacto entre víctimas y victimarios, y es posible que estos momentos sean focos de propagación del virus.

Tabla 15 - Estructura de datos - Reporte de lesiones personales y lesiones en accidente de tránsito Policía Nacional

Nombre del Campo	Tipo de dato	Descripción
Departamento	texto	Nombre del departamento
Municipio	texto	Nombre del municipio
Código DANE	texto	Código estándar del Departamento Administrativo Nacional de Estadística DANE
Armas Medios	texto	Tipo de artefacto con el que se produce la agresión

Fecha Hecho	texto	Fecha en la que ocurrió el evento
Genero	texto	Identifica el género de la persona atacada
Grupo Etario	texto	Grupo por edades
Descripción de la conducta	texto	Lesiones Culposas / Lesiones personales
Cantidad	texto	Número de lesiones reportados

Conjunto de datos 15: Tasa de cambio representativa del mercado (TRM)

Motivaciones para la selección del conjunto de datos:

El Covid-19 afectando drásticamente la dinámica del mundo entero, al interior de cada ciudad se desdibuja la cotidianidad y específicamente el mercado financiero. La incertidumbre, la falta de demanda de servicios y la afectación laboral, impactan directamente tanto el bolsillo de los ciudadanos, como los bancos y los gigantes financieros. Este contexto mueve las tasas de interés, las tasas de cambio y el valor de las monedas.

La Tasa Representativa de Mercado (valor del cambio de pesos colombianos a dólares) y la Tasa de Intervención de Política Monetaria, se toman como variables representativas del sector financiero que podrían influenciar en las negociaciones de la ciudad, las condiciones de bienestar de los ciudadanos y por ende en sus decisiones, incluyendo el estar tranquilo en casa o tener que salir a buscar el sustento diario y aumentar el riesgo de contagio.

Tabla 16 - Estructura de datos - Tasa de cambio representativa del mercado TMR

Nombre del Campo	Tipo de dato	Descripción
Año	texto	Año
Fecha	texto	Fecha de la información
TRM	texto	Valor de referencia para el día de información
Día del mes	texto	Día del mes
Mes	texto	Mes del año
Id Mes	texto	Número del mes

Conjunto de datos 16: Tasa de intervención Banco de la República (Tasa de intervención de política monetaria)

Motivaciones para la selección del conjunto de datos:

El Covid-19 afectando drásticamente la dinámica del mundo entero, al interior de cada ciudad se desdibuja la cotidianidad y específicamente el mercado financiero. La incertidumbre, la falta de demanda de servicios y la afectación laboral, impactan directamente tanto el bolsillo de los ciudadanos, como los bancos y los gigantes financieros. Este contexto mueve las tasas de interés, las tasas de cambio y el valor de las monedas.

La Tasa Representativa de Mercado (valor del cambio de pesos colombianos a dólares) y la Tasa de Intervención de Política Monetaria, se toman como variables representativas del sector financiero que podrían influenciar en las negociaciones de la ciudad, las condiciones de bienestar de los ciudadanos y por ende en sus decisiones, incluyendo el estar tranquilo en casa o tener que salir a buscar el sustento diario y aumentar el riesgo de contagio.

Tabla 17 - Estructura de datos - Tasa de interés de política monetaria - Banco de la República Colombia

Nombre del Campo	Tipo de dato	Descripción
Año	texto	Año
TIP	texto	Tasa de interés política monetaria

4.1.4. Preparación de los datos

Para la estimación del número de nuevos casos de Covid en la ciudad de Bogotá, la base de datos a analizar debe contener una estructura que permita tener un registro por día, en el rango de fechas a estudiar.

Rango de Fechas: 01/03/2020 hasta 31/08/2020

Variable Dependiente: campo ‘Casos’

Variables Independientes: serán las variables seleccionadas durante el proceso de limpieza, transformación y construcción de nuevas variables

Para llevar a cabo el proceso de preparación se considerará los siguientes pasos que se desarrollarán en el entorno R:

1. Preparación del entorno y cargue de paquetes
2. Lectura y carga de cada fuente de datos
3. Normalización de nombres de variables (eliminación de espacios, tildes, mayúsculas, y caracteres especiales)

4. Selección de datos para la ciudad de Bogotá y variables de interés
5. Selección de registros dentro del rango de tiempo de estudio
6. Creación de variables
7. Transformación de las variables seleccionadas al formato requerido

A continuación, se describen los pasos que se llevaron a cabo durante la preparación de datos.

1. **Preparación del entorno y cargue de paquetes**

Se realizar el cargue de los siguientes paquetes de R

Librería data.table: es una versión mejorada de un dataframe que ofrece ventajas como mejoras en los tiempos de ejecución y que ocupan menos espacio que un dataframe
Librería stringr: se utiliza para trabajar con strings
Librería bit64: son útiles para manejar las claves de la base de datos y contar con exactitud
Librería tidyverse: es una librería que resume la mayor parte de las tareas que tiene que realizar un científico de datos
Librería dplyr: se utiliza para manipulación de dataframe
Librería lubridate: facilitan el análisis y manipulación de fechas

2. **Lectura y carga de cada fuente de datos**

Entrada: 16 fuentes de datos definidas durante el proceso de entendimiento de datos

Creación de tabla maestra de datos:

Paso 1: Definir mes de inicio y fin del periodo

Tabla 18 - Datos entrada y salida paso 1 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Rango de fechas definidas	Variable vRng

Paso 2: Crear una paramétrica de Periodos

Tabla 19 - Datos entrada y salida paso 2 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Variable vRng	Variable Periodos

Paso 3: Crear base final de registros

Tabla 20 - Datos entrada y salida paso 3 de proceso de limpieza, selección y construcción de variables

Entrada	Salida

Variable Periodos, Rango de fechas	Data.table BCovid

Carga de datos:

Paso 4: Se fija ruta de extracción

Tabla 21 -Datos entrada y salida paso 4 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Ruta de directorio	Parámetro ruta de directorio

Paso 5: Con la función 'fread' se realiza el cargue de las fuentes

Se utiliza la función 'fread': que se utiliza para crear data.table

Tabla 22 - Datos entrada y salida paso 5 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Fuentes de datos	Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP

3. Normalización de nombres de variables (eliminación de espacios, tildes, mayúsculas, y caracteres especiales)

Paso 6: Normalizar nombres de variables

Por cada conjunto de datos se aplican las funciones: 'chartr'; para quitar tildes de mayúsculas y minúsculas, función 'gsub': para quitar espacios en las palabras, y la función' tolower': convierte los caracteres a minúscula

Tabla 23 - Datos entrada y salida paso 6 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP	Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP con eliminación de espacios, tildes, mayúsculas, y caracteres especiales

4. Selección de datos para la ciudad de Bogotá y variables de interés

Paso 7: Selección de información para Bogotá y selección de variables

Para cada conjunto de datos que lo requiere se realiza un subconjunto con la información para la ciudad de Bogotá y se seleccionan las variables de interés

Tabla 24 - Datos entrada y salida paso 7 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP	Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP con las variables seleccionadas para cada conjunto de datos de la información para la ciudad de Bogotá

5. Creación de variables

Paso 8: Creación de variables para completar variables de interés

Para cada conjunto de datos se realiza la creación de la variable diaNum y los espacios vacíos se reemplazan por ceros utilizando la función is.na()

Tabla 25 - Datos entrada y salida paso 8 de proceso de limpieza, selección y construcción de variables

Entrada	Salida
Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP	Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP con una nueva variable díaNum y datos sin vaciós

6. Transformación de las variables seleccionadas al formato requerido

Paso 9: Agrupación de los conjuntos de datos en la tabla maestra de datos

Cada conjunto de una vez terminado el proceso de limpieza, selección y creación de variables se consolida en la tabla maestra de datos *BCovid* utilizando la función *merge*

Tabla 26 - Datos entrada y salida paso 9 de proceso de limpieza, selección y construcción de variables

Entrada	Salida

Variables de tipo dataframe por cada archivo: Covid, PCR, CovidSalud, RT, OcuCamas, OcuUCI, OcuCremat, TempMin, TempMax, TransMasv, Vuelos, Trafic, Hurtos, Lesiones, TRM, TIP	Tabla maestra de datos BCovid con los datos preparados para construcción del modelo dashboard

4.1.5. Modelado

Selección de la técnica

Supuestos para la selección técnica

Para la selección de la técnica se tienen en cuenta los siguientes supuestos:

- Para la estimación del número de casos de Covid-19 en la ciudad de Bogotá se parte de considerar fuentes datos en diferentes aspectos del entorno social que luego de evaluarlos se consideran relevantes frente al análisis de estimación de nuevo casos positivos, tales aspectos son: Salud, Clima, Seguridad, Transporte y Financiero.
- La OMS ha establecido que el periodo de incubación del coronavirus COVID-19 es de 2 a 14 días (Coranavirus Colombia Gov, 2020), por esta razón, si se quiere estimar el número de casos nuevos para un día en particular, debería hacerse evaluando las aspectos considerados en un periodo entre 2 a 14 días atrás, es decir, por la naturaleza del virus no es conveniente pronosticar el número de casos nuevos con las condiciones del día anterior.
- Para efecto de este trabajo se seleccionó un tiempo medio del periodo de incubación de 8 días, así a los datos de las variables independientes de una fecha específica se le asoció el valor de número de casos nuevos de 8 días adelante para la variable dependiente.
- El conjunto de datos trabajado luego de realizar el proceso de transformación y limpieza contempla una variable a pronosticar (Casos) y 15 variables independientes.
- Todas las vas variables seleccionadas para el conjunto de datos, son de tipo numérico.

Selección de técnicas y justificación para ser utilizadas en el proyecto

La técnica comúnmente utilizada para realizar pronósticos numéricos (variable a estimar numérica) es la Regresión Lineal. Cuando el contexto de los datos implica adicionalmente un componente de temporalidad, los análisis de series de tiempo son técnicas bastante extensas y profundas con alto contenido estadístico y matemático.

Sin embargo, en el marco del Máster, métodos para el análisis e interpretación de datos masivos, basados en el aprendizaje estadístico y la minería de datos, y la aplicación de técnicas de Machine Learning; sugieren otros caminos interesantes para abordar la

estimación objeto de este trabajo. Por esta razón se eligen las técnicas regresión con modelos KNN y Árboles de Decisión para realizar el pronóstico de nuevos casos de Covid-19, diferentes a la Regresión Lineal Genera (GLM) que también hace parte de este grupo.

Preparación de entorno para análisis de datos y modelado

Entorno de desarrollo → Python

Paso 1: Cargue de librerías

Librería Pandas: librería de python destinada al análisis de datos, que proporciona unas estructuras de datos flexibles y eficiente (Pandas, 2020)

Librería scikit-learn: librería que proporciona funciones para desarrollar aprendizaje automático y análisis de datos (Scikit-learn.org, 2020)

- Función sklearn.tree.DecisionTreeRegressor: función que pertenece al modulo de arboles de decisión utilizando un regresor (Scikit-learn.org, 2020)
- Función sklearn.neighbors.KNeighborsRegressor: función que pertenece al módulo del algoritmo de K vecinos cercanos (Scikit-learn.org, 2020)

Análisis de Componentes Principales (PAC) y Análisis de Clúster

Entorno de desarrollo → Python

Paso 1: **Obtener datos para trabajar**

Paso 2: **Preparación de conjunto de datos para realizar análisis de características**

Para hacer el análisis de características se excluye la variable de pronóstico “Casos” y se realiza el transposición del conjunto de datos utilizando el función ‘transpose()’.

Paso 3: **Identificar el grado de asociación entre variables independientes para reducir el problema de maldición de dimensionalidad**

El primer análisis que se propone es identificar la colinealidad o dependencia entre las variables explicativas del conjunto de datos. Para esto se plantea el análisis de correlaciones multivariado (Matriz de Correlaciones) con el coeficiente de Pearson.

Paso 4: **Normalización de datos previo al Análisis de Componentes Principales (PAC) de las variables**

Se realiza la normalización de los datos mediante las funciones de ‘RobustScaler, StandardScaler’ y ‘MinMaxScaler’, siendo esta última de la que obtuvo el mejor resultado al usar los datos normalizados en los pasos siguientes.

Paso 5: **Análisis de Componentes Principales (PAC) de las variables**

Se realiza un análisis de componentes principales (PAC) sobre las variables del conjunto de datos con el fin de reducir la dimensionalidad y graficarlas en un espacio bidimensional, evaluando el criterio de 'variance ratio[2]' para verificar el aporte de variabilidad de las variables seleccionas, en un gráfico de dispersión.

Paso 6: **Análisis de clustering jerárquico**

Teniendo en cuenta que las variables representan valores pequeños se realiza análisis de clustering jerárquico sobre estas para observar sus relaciones de similitud.

Para evaluar estas relaciones se utilizó la matriz de distancias, basada en una medida de distancia euclidiana y un dendograma para verificar grupos de características similares. Se utilizan las funciones 'Sigle Link' y 'Complete Link', para la generación de los grupos.

Paso 7: **Análisis de clustering de densidad DBSCAN**

Se realiza análisis de clustering de densidad DBSCAN, con el fin de determinar si alguna de las características se comporta de forma diferente a las demás (outliers). Para esto se parametriza del algoritmo seleccionando con un mínimo de puntos 'minPtos' y encontrando para diferentes 'épsilons', las distancias a los puntos más lejanos.

Mediante el gráfico de distancias vs. épsilons se estableció un rango de pruebas para simular para cada uno de sus valores, el número de clústeres y especialmente los 'outliers' presentes en las variables.

Paso 8: **Selección de variables para el pronóstico**

Luego de realizar los pasos anteriormente descritos determina un criterio de selección de variables para incluir en el pronóstico de casos de Covid-19.

Regresión Mediante Arboles de Decisión (CART)

Entorno de desarrollo → Python

Paso 1: Cargue de conjunto de datos con la selección de variables para pronóstico obtenido mediante análisis de componentes principales (PAC) y análisis de clustering

Paso 2: Análisis de Correlaciones

[2] Variace Ratio: en español la razón de varianza corresponde a la carga de variabilidad en el comportamiento del conjunto de datos, asociable a cada variable o componente proyectado en el PAC. Se espera que los 2 primeros componentes acumules más del 95% de la varaibilidad.

Se realiza análisis de correlaciones basado en el coeficiente de Pearson, en la que se tiene en cuenta la variable objetivo 'Casos', con el fin de verificar la relación entre las variables independientes con la variable objetivo.

Paso 3: Partición del conjunto de datos en training y test

Se realiza partición del conjunto de datos: training y test. Para esta partición se tiene en cuenta que el conjunto de datos corresponde a una estructura cronológica, así que se mantiene la naturaleza cronológica del conjunto de datos con el fin de evitar realizar predicciones para datos del pasado usando datos futuros.

Considerando que la definición del rango de tiempo para el análisis de datos corresponde a un periodo de 6 meses comprendido entre el 01/03/2020 al 31/08/2020 y que este no es muy amplio, ya que, en términos de información, la duración de la presencia de virus es muy corta; la partición de archivos se hace 80% para training (comprendido en el rango de 01/03/2020 al 31/07/2020) y 20% testing (comprendido en el rango de 01/08/2020 al 31/08/2020)

Paso 4: Regresión mediante arboles de decisión

Se aplica la técnica de regresión mediante arboles de decisión con la función 'DecisionTreeRegressor' de la librería 'scikit-learn'

Paso 5: Parametrización del modelo

Como criterio de evaluación se eligió el error absoluto medio (MAE). Adicionalmente se procedió a hacer la simulación del parámetro clave de máxima profundidad 'max_depth' con valores entre 2 y 30.

Paso 6: Selección del criterio de evaluación

Mediante un gráfico de dispersión de profundidad vs. valores del error absoluto medio (MAE), se analizó la curva para escoger los candidatos de profundidad que garantizaran un error absoluto medio más pequeño posible, mejorando así rendimiento.

Paso 7: Validación cruzada mediante el método 'KFold[3]' - Cross-Validation

Se realiza la validación cruzada mediante el método 'KFold', estableciendo un numero de subgrupos 'splits' acorde al tamaño de la base de datos de training, y niveles de profundidad del árbol simulados en un intervalo; se escoge la profundidad adecuada con un MAE razonablemente pequeño.

Paso 8: Construcción del modelo

Se construye el modelo con la función 'DecisionTreeRegressor', con el parámetro de profundidad seleccionado en el paso anterior. Se ejecuta el modelo y se realizan las

[3] El método K-Fold Cross-Validation consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, el proceso genera k estimaciones del test-error, cuyo promedio se emplea como estimación final (Parra, 2019)

pruebas y el cálculo final del error absoluto medio (MAE) contrastando la estimación con los datos de testing.

Paso 9: Calcular la importancia de las variables seleccionadas para el pronóstico

Utilizando la función de ajuste 'fit' se calcula la importancia de las variables seleccionadas para el pronóstico con el fin de avaluar el peso en el modelo.

Paso 10: Construcción de la visualización de estimación

Se grafican los valores estimados vs. los valores reales para visualizar la estimación.

Paso 11: Optimar resultados para obtener el mejor ajuste y profundidad

El proceso se repite varias veces probando la mejora de la parametrización hasta llegar a un nivel adecuado de los resultados, un error absoluto medio (MAE) lo más pequeño posible, con un buen ajuste y un nivel de profundidad adecuado, cuidando que no llegar al sobreajuste.

Regresión Mediante K-NN

Paso 1: Selección de conjunto datos

Para esta regresión se utiliza el conjunto de datos utilizado para la regresión con árboles de decisión del apartado anterior, así como también se mantiene la partición de archivos 80% para training (comprendido en el rango de 01/03/2020 al 31/07/2020) y 20% testing (comprendido en el rango de 01/08/2020 al 31/08/2020)

Paso 2: Parametrización del modelo

Se establecen 3 parámetros:

- 'KNeighbors' (*k* vecinos más cercanos) ó 'RadiusNeighboors' (vecinos en un radio)
- *k* ó Radio
- El peso: uniforme o por distancia

Paso 3: Selección 'KNeighbors' – *k* vecinos más cercanos

Se escoge 'KNeighbors', *k* y simulando el *peso* para los dos tipos de ponderación, por validación cruzada.

Paso 4: Selección del criterio de evaluación

Al igual que en el modelo anterior, se considerará el Error Absoluto Medio (MAE) como criterio de evaluación del modelo.

Paso 5: Validación cruzada mediante el método 'KFold ' – Cross-Validation

Mediante el método de validación cruzada 'KFold' se establecen un numero de subgrupos acorde al tamaño de la base de datos de training para las validaciones, y el número de vecinos a simular en un intervalo; se encuentra el menor valor del MAE entre los obtenidos por los dos métodos de pesos: uniforme y distancia.

Paso 6: Construcción del modelo

Escogida la mejor parametrización del paso anterior: vecinos y método de peso, se realiza construcción del modelo utilizando la función 'KNeighborsRegressor' de la librería 'scikit-learn'.

Paso 7: Ejecución de training y test

Se ejecuta el modelo y se realizan las pruebas con los datos del conjunto de datos 'testing' para obtener valores reales y los valores estimados. Luego se elabora grafica de las variables con el fin de visualizar los resultados del pronóstico.

4.1.6. Construcción de Dashboard

Las visualizaciones del análisis de datos y resultados se construyen en Qlik Sense como se tenía definido en la arquitectura del proyecto, a continuación se detallan los pasos para su elaboración:

Paso 1: **Alistamiento de datos para las visualizaciones y cargue de datos**

Para la construcción del dashboad se utilizan las siguientes fuentes:

- Conjunto de datos 1: Casos positivos de COVID-19 en Colombia de datos, en este conjunto se encuentra los casos diagnosticados por Covid-19 a nivel nacional,
- La tabla maestra BCovid, en la que se encuentran las variables luego del proceso de transformación, limpieza y selección de variables de las fuentes de datos de los tópicos elegidos
- Del proceso de modelamiento se obtienen tres salidas de datos estimados: Estimacion_DT que corresponde al resultado de la ejecución del algoritmos de árboles de decisión , y de los resultados del algoritmo para KNN las salidas: Estimacion_KNN1 y Estimacion_KNN2

Paso 2- Preparación modelo de visualización

Una vez se obtienen los datos descritos en el paso 1 se disponen los datos en un modelo estrella el campo referencia será el 'Fecha' para todas las tablas. El campo 'Fecha' corresponde a los días de estudio de los casos de Covid-19

Paso 3- Ajuste a los datos obtenidos desde lo modelos

En los datos que se obtienen en los modelos en los que se encuentra el periodo y la estimación se realiza el ajuste del valor de predicción a 8 días adelante, esto se realiza porque durante la etapa de preparación y análisis de datos se había traído el valor del Covid-19 de 8 días adelante a un periodo anterior para poder

estimar con una media de 8 días de incubación es por ello que para la preparación de datos se debe ajustar al estado original, cabe aclarar, que los valores de las variables explicativas de una fecha se deben pronosticar a 8 días promedio de incubación del virus

Paso 4 –Transformación de datos en Qlik Sense

Con los datos cargados y ajustados según la necesidad se realizan las transformaciones para garantizar línea de tiempo y calidad de los datos

Paso 5 –Definición de los indicadores

Tabla 27 -Definición de indicadores para visualización

Definición de Indicadores para visualización del pronóstico de casos nuevos positivos en Bogotá, Colombia	
Audiencia:	Está dirigido a todos aquellos ciudadanos de Colombia e interesados de otras regiones del mundo, entidades públicas y privadas y centros de investigación de seguir en cifras el comportamiento del Covid-19 en la ciudad de Bogotá, Colombia
Objetivo	Utilizar la visualización de datos con el fin de presentar las principales cifras que describen el comportamiento del virus en la ciudad de Bogotá, y contrastarlo con el comportamiento de las variables de los tópicos seleccionados para el análisis y pronostico, para que de manera visual se puedan observar las relaciones encontradas entre ellas y el número de casos de Covid -19. Adicionalmente, presentar los resultados de la estimación obtenida en cada uno de los modelos y así orientar a la audiencia
Mensaje	El impacto que sobre la ciudad de Bogotá pueden tener las dinámicas de diferentes sectores del país, en la propagación del virus y medir su impacto mediante la estimación de nuevos casos.
Definición de métricas	
Número de casos a nivel nacional	Conteo Número de casos
Número de Casos a nivel Bogotá	Conteo Número de casos filtrado por la ciudad de Bogotá
% Casos Bogotá respecto a Colombia	$\frac{Número\ de\ casos\ a\ nivel\ nacional}{Número\ de\ casos\ a\ nivel\ Bogotá}$
Número estimado de nuevos casos Modelo DT	Conteo de nuevos casos estimados

Número estimado de nuevos casos Modelo KNN1	Conteo de nuevos casos estimados
Número estimado de nuevos casos Modelo KNN2	Conteo de nuevos casos estimados
Casos Bogotá por Genero	Conteo Número de casos filtrado por la ciudad por cada uno de los géneros
% Casos Bogotá por estado **(Recuperado, Fallecido, Leve, Moderado, Leve. Grave)**	$\frac{Número\ de\ casos\ a\ nivel\ Bogotá\ x\ estado}{Número\ de\ casos\ a\ nivel\ Bogotá}$
Total de Casos por cada una de estas variables: Pruebas_PCR, Casos_Salud, RT, Ocupacion_Camas, Ocupacion_UCI, Ocupacion_Crematorios, Temp_Min, Temp_Max, Pasajeros_TM, Vuelos, Trafico, Hurtos, Lesiones_Personales, TRM, TIP	Conteo Número de casos / por cada variable

Resultado de la construcción de la visualización

Luego de preparar los datos y establecer el objetivo, audiencia y mensaje que se pretende trasmitir a través de la visualización a continuación se presenta la propuesta de dashboard para el proyecto

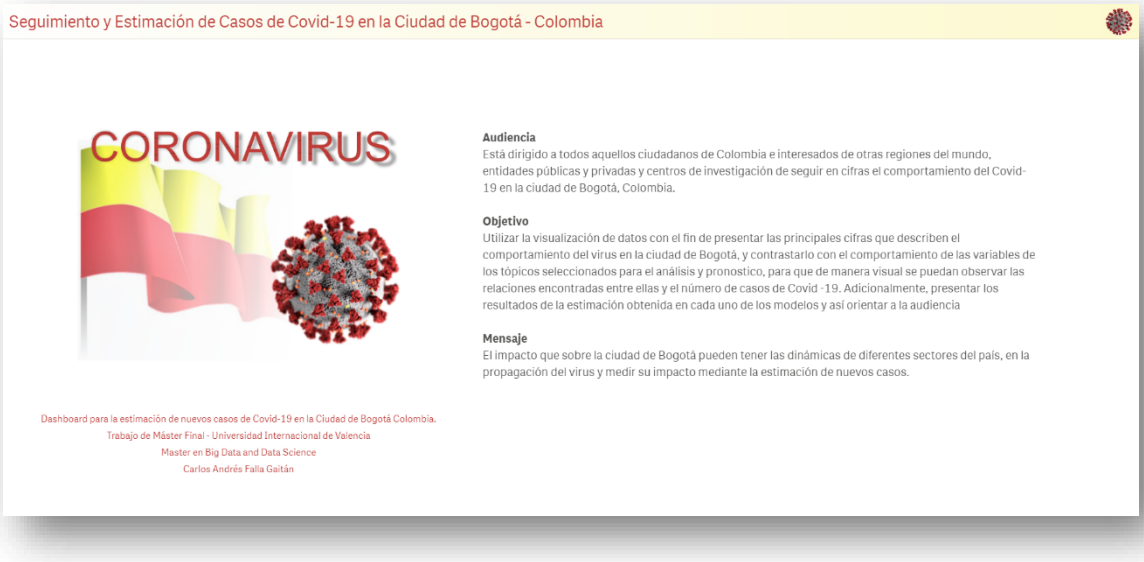

Ilustración 12 - Presentación de introducción al dashboard de seguimiento y control. Fuente: Elaboración propia en Qlik para el proyecto

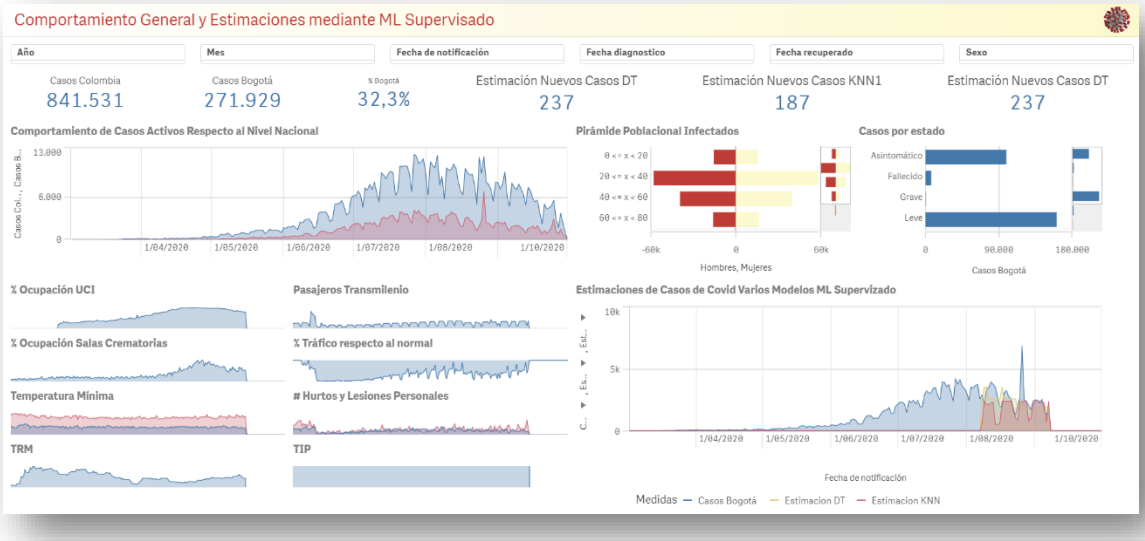

Ilustración 13 - Presentación dashboard seguimiento y control. Fuente: creación propia en Qlik para el proyecto

4.2. Resultados

Selección de Características Mediante Análisis de Correlaciones, Análisis de Componentes Principales (PAC) y Análisis Clúster

Análisis de Correlación para selección de características

Para verificar la dependencia entre características, y entre ellas y la variable a estimar (Casos), se usa la matriz de correlaciones de Pearson, y su representación gráfica con un mapa de calor.

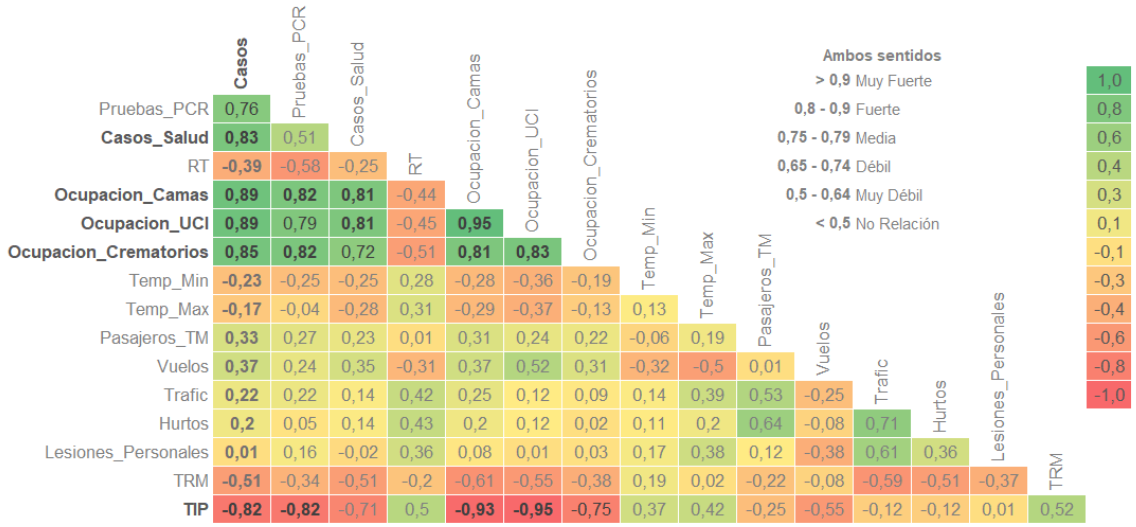

Ilustración 14 - Grafica Matriz de correlaciones. Fuente: Creación propia para el proyecto

Partiendo de la clasificación de niveles de correlación presentado en la Ilustración 12, se encontró que respecto a la relación de la variable a estimar (Casos) con las demás variables del conjunto de datos:

- No existen relaciones de nivel **muy fuerte** entre las variables independientes.
- Se encuentra relación **fuerte positiva** con la mayoría de las variables del grupo de Salud: Casos_Salud, Ocupación_Camas, Ocupación_UCI, Ocupación_Crematorios
- Presenta una relación **media positiva** con la variable Pruebas_PCR.
- Existe una relación **fuerte negativa** con la variable económica TIP.
- La relación con la TRM es una relación **positiva muy débil**.
- Las demás variables del conjunto de datos no tienen una relación significativa con la variable de estudio.

Desde el punto de vista de la colinealidad o del problema de la "maldición de la dimensionalidad[4]", analizando la Ilistración1 2, se muestra un patrón claro de correlaciones entre algunas variables:

- Es evidente que las variables referentes al grupo de fuentes de Salud (porcentaje de ocupación de camas, UCIs y salas crematorias) excepto por el número de retransmisión RT, presentan una dependencia **media** y **fuerte** especialmente entre ellas.
- La variable RT pareciera no tener relación significativa con nadie, es decir, resulta ser una variable totalmente independiente.
- El bloque de variables de los grupos de clima, seguridad y transporte tampoco presentan relación significativa con ninguna de las demás variables del conjunto de datos, por lo que se asumen independientes.
- En cuanto a las variables financieras, la tasa representativa de mercado TRM, también presenta un comportamiento independiente de las demás variables.
- La variable de Tasa Impuesta del Gobierno (TIP) mantiene una correlación **muy fuerte** con las variables de ocupación de Camas y UCIs, y **fuerte** y **media**, con Pruebas_PCR y Ocupación_Crematorios respectivamente.

Análisis de Componentes Principales

El análisis PAC sobre las características, se obtuvo como resultado del 'variance ratio' para los dos primeros componentes, los valores de 0.937 y 0.063, lo que significa un muy buen ajuste.

La visualización de los puntos de datos (características en este caso) en el plano de las dos nuevas dimensiones, como se aprecia en la Ilustración 13, muestra un resultado bastante diferente al comportamiento visto en la matriz de correlaciones.

[4] En Machine Learning, el número de dimensiones se puede equiparar al número de variables o características (features) que estemos utilizando. La maldición de la dimensión se «manifiesta» de dos maneras: 1) La distancia media entre los datos aumenta con el número de dimensiones y 2) La variabilidad de la distancia disminuye exponencialmente con el número de dimensiones (éste es el verdadero problema)

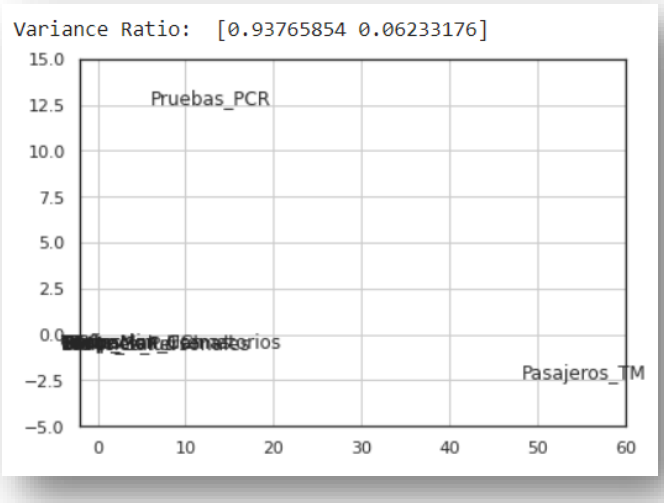

Ilustración 15 - Grafica de Variance Ratio para Analisis de Componentes Principales (PAC). Fuente: creación propia para el proyecto

Claramente sugiere que las variables Pruebas_PCR y Pasajeros_TM, tienen comportamientos totalmente diferentes al de las demás variables, y que las demás prácticamente se comportan de manera similar.

Esto podría interpretarse como que todas ellas deberían estar muy correlacionadas entre sí, lo que contrasta con los hallazgos anteriores en donde un grupo de ellas prácticamente es totalmente excluyente del comportamiento de las demás.

Por otro lado, la variable Pruebas_PCR que se presenta como variable atípica, tiene una correlación alta con las variables de ocupación, lo que no concuerda con PAC que la aísla completamente; mientras la variable Pasajeros_TM no presenta relación significativa con nadie en el análisis de correlación, es decir, se alinea con los resultados del PAC.

Es probable que la consistencia y variabilidad de datos en términos de número de características y en calidad de la información influya significativamente en los resultados de los modelos, lo que sería un punto de mejora a tener en cuenta en el ejercicio.

De manera general, adelante se hace la verificación con las técnicas de clustering jerárquico y de densidad DBSCAN, y los resultados se refuerzan con lo encontrado en el PAC.

Análisis de clustering jerárquico

A partir de los datos normalizados, se aplica un algoritmo de clustering jerárquico, en la ilustración 14 se presenta la matriz de distancias con un mapa de calor las distancias entre características

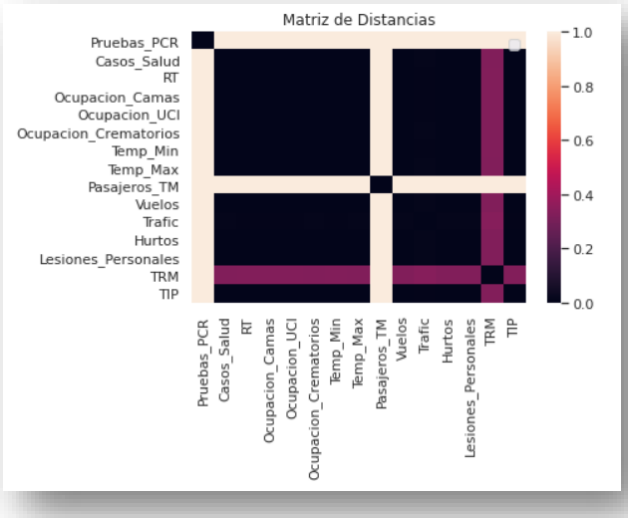

Ilustración 16 - Matriz de Distancias para Análisis de Clustering Jerárquico. Fuente: creación propia para el proyecto

Es evidente que la distancia de las dos variables ya descritas en el PAC, es mucho mayor que las del resto de características, y que la distancia de las demás características es muy pequeña y parecida entre sí, a excepción de la variable TRM que presenta una distancia menos parecida pero aun así pequeña.

Para terminar de verificar este hecho, se procede a realizar la agrupación de las características mediante los métodos de distancias entre clústeres 'Single Link' y 'Complete Link'

En las ilustraciones 15-'Complete Link' e ilustración 16 –'Single Link' se presentan los dendogramas correspondientes para evaluar el umbral de distancia, evidencian de nuevo que la distancia de los puntos establecidos entre la mayoría de las características es muy parecida, por lo que solo se distinguen las dos variables analizadas previamente.

Así sin importar la parametrización, el resultado es el mismo, es decir tres clústeres, uno para Pruebas_PCR, uno para Pasajeros_MT y otro para el resto de las características.

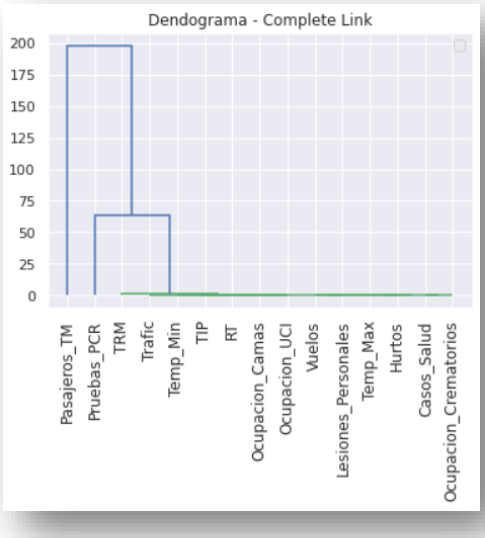

Ilustración 17 - Dendograma Complete Link - Clustering Jerárquico. Fuente: creación propia para el proyecto

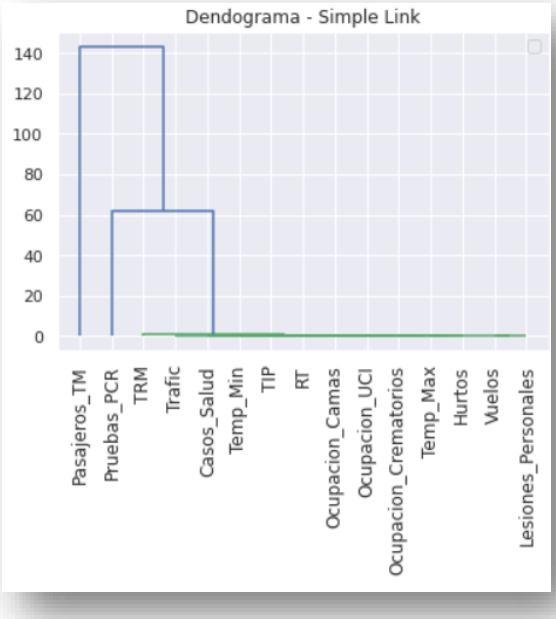

Ilustración 18 - Dendograma Simple Link - Clustering Jerárquico. Fuente: creación propia para el proyecto

Análisis de clustering densidad (DBSCAN)

Para complementar el ejercicio, se realiza un análisis por densidad mediante el método DBSCAN, con el fin de encontrar características atípicas. Sin embargo, por los resultados obtenidos previamente es posible deducir cuales variables van a salir como atípicas en el análisis y cuántos clústeres se van a encontrar.

Se inicia parametrizando el modelo con un número mínimo de 2 vecinos y se ejecuta la simulación de distancias lejanas para varios valores de 'épsilon'. Se grafican, y el resultado es el siguiente diagrama:

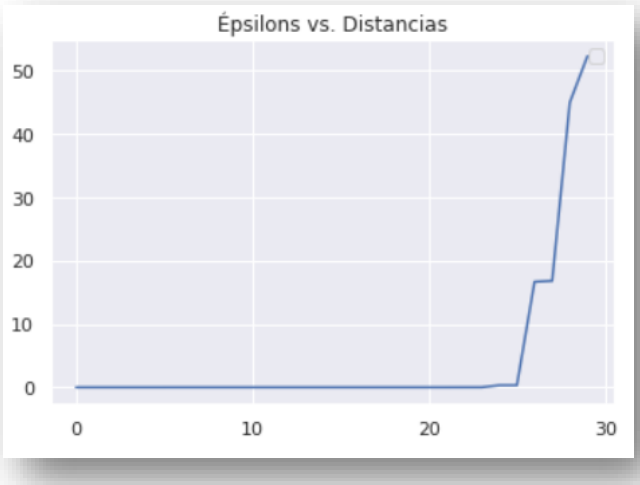

Ilustración 19 - Grafica Épsilon Vs Distancias - Clustering de Densidad DBSCAN. Fuente: creación propia para el proyecto

Es claro como la curva presenta un arrastre del valor de 'épsilon' hasta una distancia de 25, por lo que el punto de quiebre se da a un 'épsilon' alrededor de 1.

Con este parámetro definido, y realizando varias pruebas se establece el épsilon en un intervalo de 0.5 y 1.8 con incrementos de 0.2, para generar los escenarios de numero de clústeres, número de 'outliers' y los valores de 'épsilon'. El resultado es como sigue:

```
from sklearn.cluster import DBSCAN

for eps in np.arange(1, 5, 0.5):
  db = DBSCAN(eps, min_samples=minPts).fit(features_norm)
  core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
  core_samples_mask[db.core_sample_indices_] = True
  labels = db.labels_
  n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
  n_outliers = list(labels).count(-1)
  print ("%6.2f, %d, %d" % (eps, n_clusters_, n_outliers))

#labels
```

```
  1.00, 1, 2
  1.50, 1, 2
  2.00, 1, 2
  2.50, 1, 2
  3.00, 1, 2
  3.50, 1, 2
  4.00, 1, 2
  4.50, 1, 2
```

Ilustración 20 - Resultado de agrupación 'outliers' - Clustering de Densidad DBSCAN. Fuente: creación propia para el proyecto

Se corrobora entonces que la mejor selección implica dos características atípicas y 1 clúster para el resto de ellas. Mismos resultados vistos en el análisis PAC y el clustering Jerárquico.

A manera de resumen, consolidando los resultados vistos en todas las metodologías propuestas, las sugerencias de selección de variables para los modelos propuestos adelante son:

Tabla 28 - Aplicación de metodologías para cada variable

Características	Correlación	PAC/HAC/DBSCAN
Pruebas_PCR	x	x
Casos_Salud		
RT		
Ocupacion_Camas		
Ocupacion_UCI	x	x
Ocupacion_Crematorios		
Temp_Min	x	
Temp_Max	x	
Pasajeros_TM	x	x
Vuelos	x	
Trafic	x	
Hurtos	x	
Lesiones_Personales	x	
TRM	x	
TIP		

A continuación, se procede a modelar los datos para estimar los Casos de Covid, mediante dos métodos de regresión por aprendizaje supervisado: Arboles de Decisión y KNN.

La elección de las características en cada uno se hará mediante la combinación de los escenarios presentados en la Tabla 27

Regresión Mediante Arboles de Decisión (CART)

Para el desarrollo del modelo, se realizó la partición del archivo de datos en las bases de training y testing, 80% para training (comprendido en el rango de 01/03/2020 al 31/07/2020) y 20% testing (comprendido en el rango de 01/08/2020 al 31/08/2020).

Parametrización del modelo

Para parametrizar el modelo, inicialmente se ejecuta el 'DecisionTreeRegressor' con todas las viables del archivo de training varias veces, iterando el parámetro de

profundidad del árbol entre 2 y 30, con el fin de simular los valores de criterio de evaluación MAE y tener una idea de la curva de “ganancia”, es decir de los errores versus la profundidad.

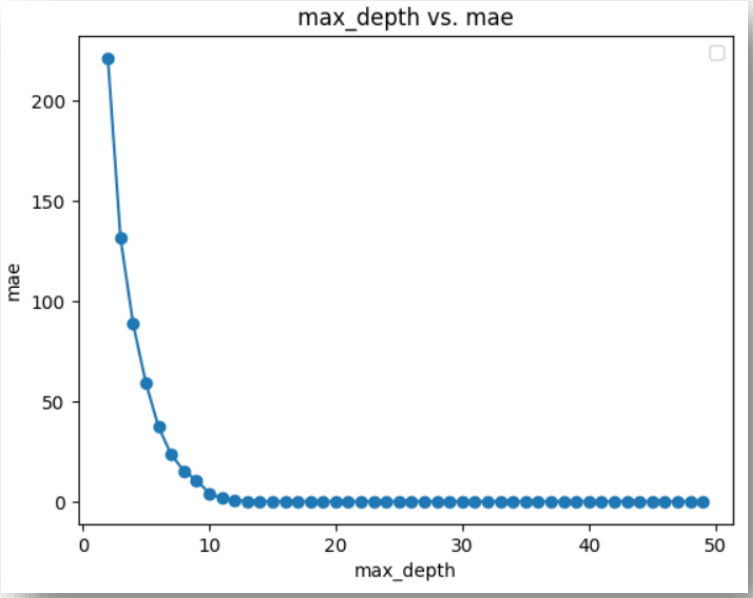

Ilustración 21 - Grafica de profundidad vs. MAE - Regresión Mediante Arboles de Decisión (CART). Fuente: creación propia para el proyecto

Este resultado indica que para estos datos una profundidad de más de 10 en el árbol podría minimizar el error, pero apuntando a un riesgo de sobreajuste, es decir que al validar el modelo contra testing, probablemente no va a ajustar bien.

Para optimizar esta apreciación de los parámetros, se realizó validación cruzada (KFold), guardando del conjunto de iteraciones, la profundidad con el menor valor posible; y usándolos para ejecutar el modelo final.

En la validación cruzada se utilizó un parámetro de poda ‘n_split’ en variaciones de 5 a 10 grupos, obteniendo mejores resultados en 5. Esto claramente puede deberse a que el conjunto de datos no es muy grande y por tanto muchas particiones tienen menos datos para entrenar.

Construcción y ejecución del modelo

Este ejercicio se ejecutó con las dos opciones de características seleccionadas en el apartado anterior, y se incluyeron dos ejercicios exploratorios adicionales, que surgen de las iteraciones hechas con variaciones de los parámetros, resultados de la validación cruzada. A continuación, los escenarios de características seleccionadas en cada caso,

los parámetros usados y los resultados del criterio de evaluación mae, después de haber ejecutado el modelo con los datos de testing:

Tabla 29 - Relación Variables, Criterios MAE, parámetros de algoritmos

Características	Caso Selección de variables			
	Correlación	PAC - HAC-DBSCAN	Exploración 1	Exploración 2
Pruebas_PCR		x		
Casos_Salud				x
RT				x
Ocupacion_Camas				
Ocupacion_UCI	x	x	x	x
Ocupacion_Crematorios				
Temp_Min	x		x	x
Temp_Max	x		x	x
Pasajeros_TM	x	x	x	x
Vuelos	x		x	x
Trafic	x			x
Hurtos	x			x
Lesiones_Personales				x
TRM	x		x	x
TIP			x	x
Parámetros				
Cross Validation - n_splits	5	5	5	7
Profundidad del árbol	7	6	9	7
Resultados				
MAE	685,505	847,516	680,967	875,58

En la tabla 28, se observan los ajustes reales versus las estimaciones para los 4 escenarios.

El escenario de correlación, correspondiente a la selección de variables por los criterios de asociación encontrados previamente, arrojan un modelo un modelo optimo en el estudio con un error bajo respecto a los otros modelos, con una profundidad de 7 para el árbol.

El escenario que contempla análisis de las técnicas no supervisadas y el PAC, se obtiene un modelo con un error mayor al primero, y una profundidad menor. Al ver el ajuste de la Tabla 28, se observa que no se adhiere al comportamiento de los datos originales y al hacer una estimación moderada y centrada como si fuera un promedio, pierde sensibilidad en los cambios de la serie original. Con estos resultados, se puede considera que la profundidad puede incidir en el resultado generando sobreajuste y por ende un aumento en el valor del error

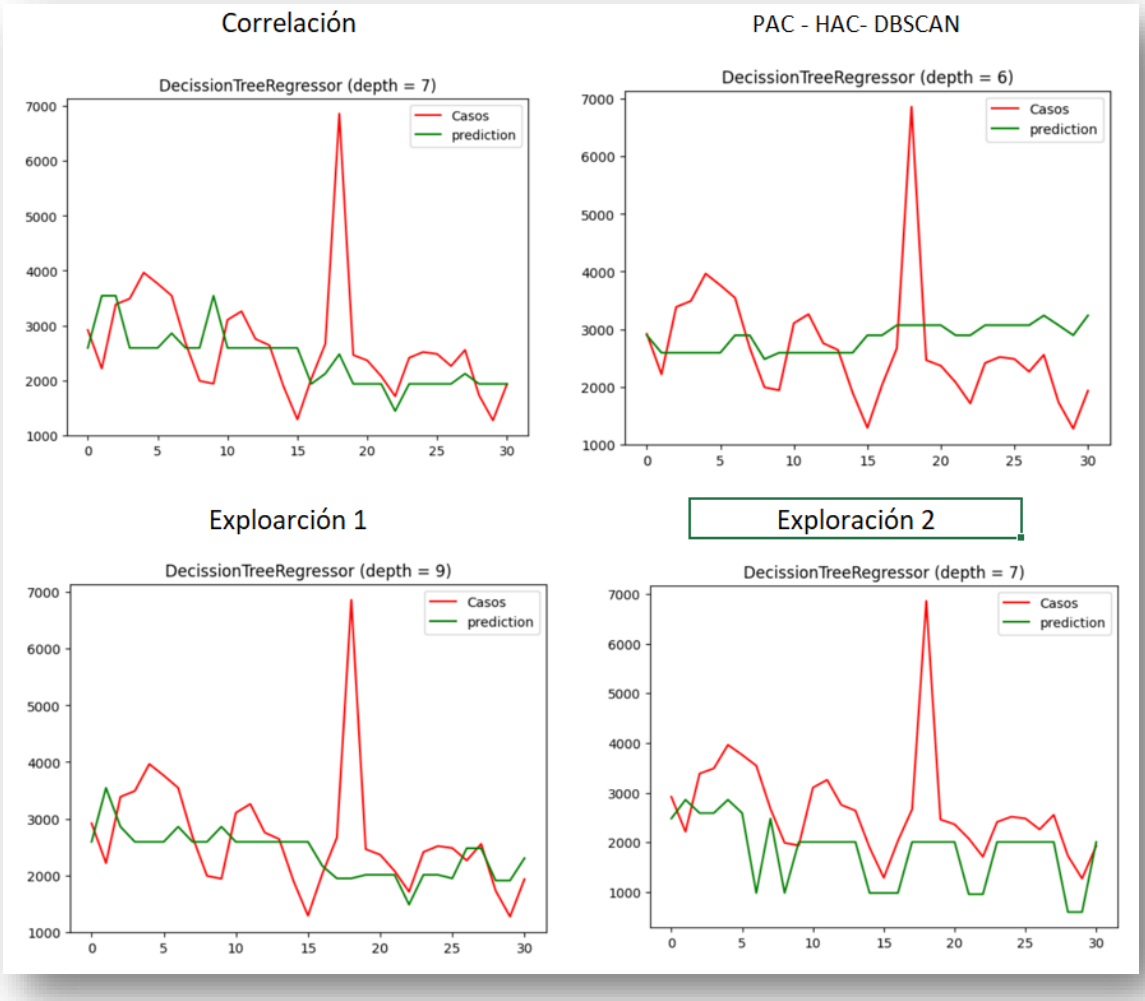

Ilustración 22 -Análisis de las técnicas no supervisadas y el PAC. Fuente: creación propia para el proyecto

El tercer escenario, exploración 1, mejora el MAE respecto al primer ejercicio, pero a otro costo, ya que aumenta la profundidad arriesgando al modelo a un sobreajuste. Sin embargo, la profundidad de 9 aun no es significativa, respecto a las profundidades obtenidas en las diferentes salidas de las iteraciones de la validación cruzada, que presentaban valores de 13, 17, 19, 22, 28.

Al ver el ajuste de la tabla 28, se obtiene un valor similar al primer ajuste presentado, pero pierde sensibilidad en ciertos puntos, es decir se adhiere menos en los picos bruscos, lo que hace considerar que la profundidad incide significativamente en sobreajuste y sin mucha ganancia del error respecto al primer modelo, pues solo disminuye 5 unidades, aún así es modelo con gran similitud con el primero

El cuarto escenario, exploración 2, resulta ser otra variación del primer escenario, agregando las variables Casos_Salud, RT y Lesiones personales. Estas dos últimas tal como se concluyó en el análisis inicial no tiene relación a la variable objetivo, adicional que no generan impacto para el modelo ya que son las variables de Salud las que representan mayor impacto al modelo:

Feature Relevances

	Attributes	Decision Tree
0	Casos_Salud	0.834641
1	RT	0.011036
2	Ocupacion_UCI	0.062774
3	Ocupacion_Crematorios	0.052558
4	Temp_Min	0.000135
5	Temp_Max	0.000984
6	Pasajeros_TM	0.009692
7	Vuelos	0.004510
8	Trafic	0.000350
9	Hurtos	0.000001
10	Lesiones_Personales	0.010438
11	TRM	0.004301
12	TIP	0.008579

Ilustración 23 - Valor de relevancia de las variables en el análisis de árboles. Fuente: creación propia para el modelo

Este modelo conserva el nivel de profundidad del árbol en 7, sin embargo, el resultado obtenido desmejora significativamente el error de estimación, a pesar de que se observa adherido a la curva de los datos reales. Para esto se considera que las estimaciones siempre están por debajo lo que implica que esto podría incidir de manera negativa en los resultados de los modelos y los resultados puedan llegar a subestimar el número de casos.

Regresión Mediante K-NN

Para modelar el comportamiento del Covid-19 por regresión KNN se realizan con el mismo conjunto de datos para el modelado con árboles.

Para el análisis de resultados se evalúan los escenarios presentando con en la Tabla 29 para cada variable.

Tabla 30 - análisis de resultados Regresión KNN

Características	Selección de variables Correlación 1	Correlación 2	Exploración
Pruebas_PCR			
Casos_Salud			
RT			
Ocupacion_Camas			
Ocupacion_UCI	x	x	x
Ocupacion_Crematorios	x	x	
Temp_Min	x	x	x
Temp_Max	x	x	x
Pasajeros_TM	x	x	x
Vuelos	x	x	x
Trafic	x	x	
Hurtos	x	x	
Lesiones_Personales	x	x	
TRM	x	x	x
TIP	x	x	x
Parámetros			
Cross Validation - n_splits	5	5	10
Vecinos k	7	6	17
Peso	Uniforme	Uniforme	Uniforme
Resultados			
MAE	984,562	976,167	902,869

Cabe resaltar que para este modelo no se normalizaron los datos, aunque el algoritmo al ser un método de distancia requiere este paso, normalmente se usa cuando se hace una clasificación, pero para el caso de este trabajo que es regresión, normalizarlas puede implicar perdida de interpretación al final del modelo o del aporte real de las variables.

Parametrización del Modelo - Modelo 1

A través de la validación cruzada, estableciendo para cada escenario (Correlación 1, Correlación 2 y Exploración) un número de subgrupos para prueba con la base de training (Kfold = 5, 5, 10 respectivamente); se procedió a simular los valores del criterio de evaluación MAE**,** para cada una de las variaciones del número de vecinos entre 1 y 50.

Los resultados de la validación cruzada se presentan en la siguiente Ilustración:

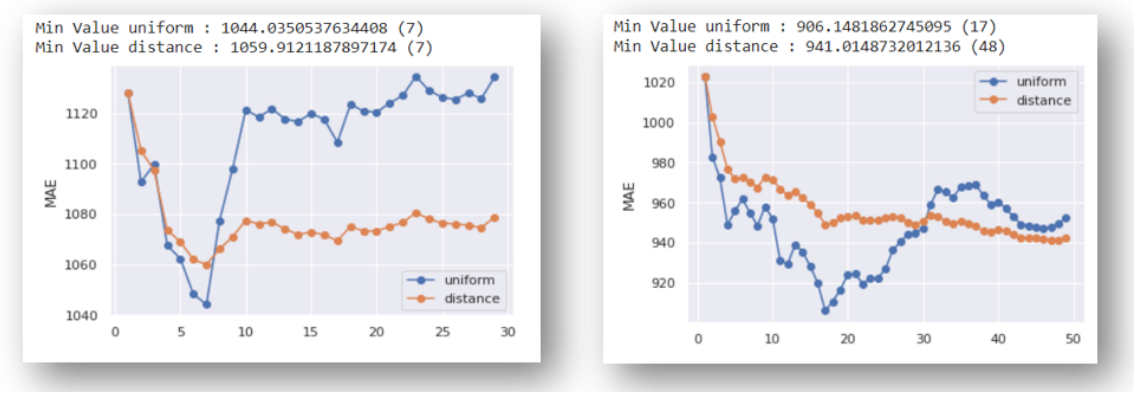

Ilustración 24 - Grafico resultado MAE por los pesos uniforme y distancia. Regresión Mediante K-NN. Fuente: Creación propia para el proyecto

La elección del parámetro número de vecinos se hace entonces basados en estos resultados. En la ilustración 22, gráfico de la izquierda muestra que la mejor escogencia de k para garantizar un mae mínimo con el conjunto de datos de training es 7, generando una mae=984,562. Sin embargo, haciendo pruebas de valores alrededor de 7, se encuentra que con un k=6 el mae puede mejorar un poco disminuyendo en 12 unidades (976,167).

Para el ejercicio exploratorio, se opta por excluir las variables Ocupación_Crematorios, Trafic, Hurtos y Lesiones personales, ya que las tres últimas no tienen una correlación alta con la variable objetivo, y la primera está altamente correlacionada con la variable Ocupación UCI involucrada en el modelo.

Adicionalmente, se ejecuta la validación cruzada permitiendo que el tamaño de los subgrupos de prueba sea más grande, teniendo como resultado que el valor mae mejoró. La elección de los 17 vecinos se hace evaluado el gráfico de la derecha de la Ilustración 22, con los resultados de la validación cruzada.

Construcción y ejecución del modelo

Habiendo escogido los parámetros para cada ejercicio, se procede a ejecutar los modelos obteniendo los valores mae (estos parámetros y resultados se encuentran en la Tabla 29).

El primer modelo, escenario Correlación 1, obtiene un mae de 984,562 con un número de vecinos igual a 7 y un peso uniforme, gráficamente no se diferencia del modelo del escenario Correlación 2, sin embargo, al bajar en este el número de vecinos a 6, el criterio de evaluación mae, mejora en 976,167.

Luego de este análisis se concluye que la diferencia entre los dos modelos es mínima, de igual manera se hace la comparación para efectos prácticos de del trabajo.

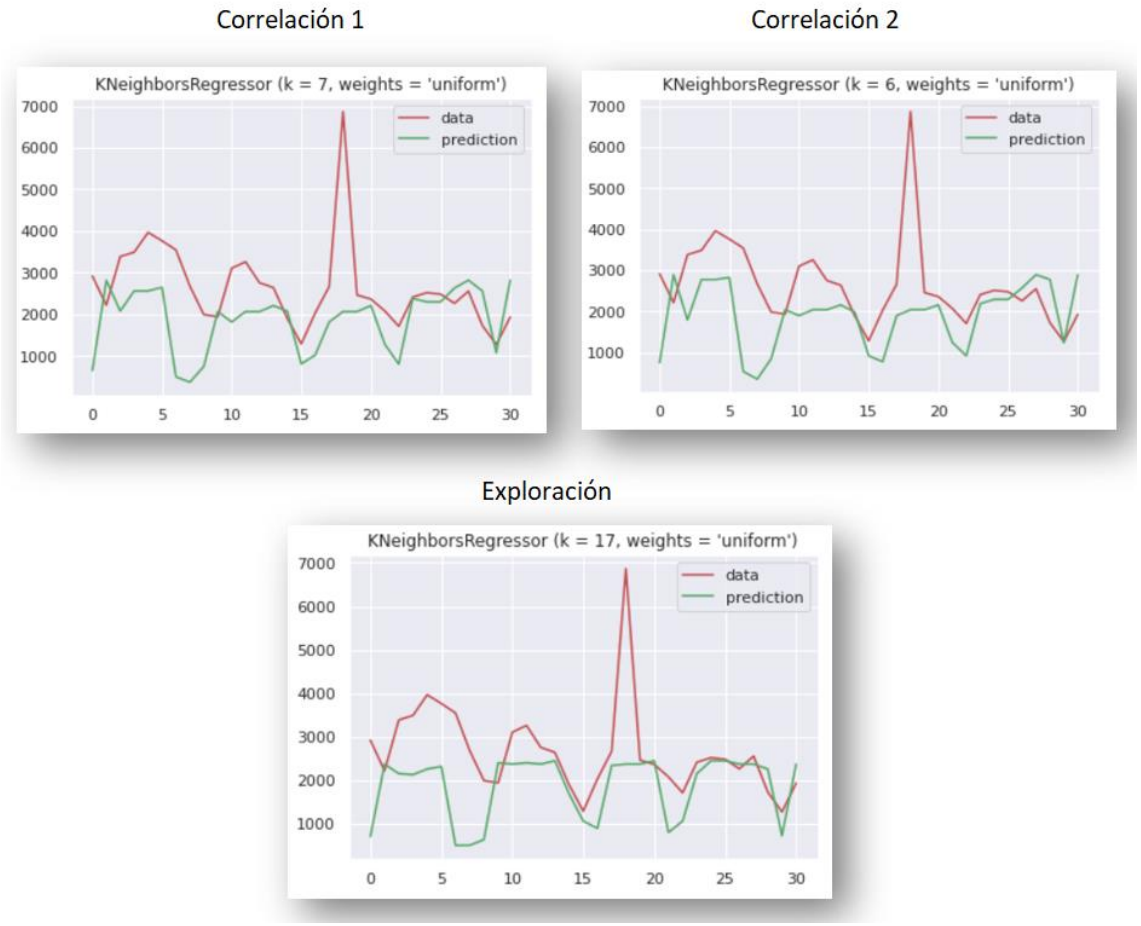

Ilustración 25 - Resultados del pronóstico - Regresión Mediante K-NN. Fuente: Creación propia para el proyecto

En el caso del tercer escenario exploración, se logra reducir el valor de mae a 902,869, es decir es un modelo con mejor criterio. Sin embargo, esto tiene un costo y es el hecho de crear menos grupos con más vecinos, en este caso 17. En consecuencia, también se está corriendo el riesgo de no ver los datos atípicos y que estos afecten la estimación.

En el tercer grafico de la Ilustración 28, se puede observar que el pronóstico realizado se parece mucho a los dos escenarios anteriores. Sin embargo, también se observa una menor sensibilidad a los picos altos.

5. Conclusiones

La variabilidad de los datos oficiales complica la labor de análisis y predicciones

Tal como se identificó en el análisis FODA del apartado de la definición del problema, la calidad de los datos influyó significativamente en los resultados de los modelos, durante el proceso de preparación de datos se encontraron para la fuentes de datos seleccionadas incompletitud de la información así como también que no existía homogenización de datos entre los diferentes conjuntos de datos, durante la fase de preparado de datos se procuró crear una tabla maestra de datos homogénea que permitiera integrar los diferentes tópicos escogidos para el análisis de datos y posterior modelamiento, sin embargo, la incompletitud de datos impacto significativamente en los resultados.

Del análisis de variables de los tópicos seleccionados para el modelo

De los tópicos seleccionados en el análisis de las variables por las técnicas de del análisis de componente principales (PAC) y análisis de Clustering, se resalta las siguientes conclusiones:

1. Es claro que las variables provenientes de las fuentes del grupo de Salud tienen una relación positiva significativa con el número de casos (excepto por el indicador RT del número de casos secundarios), esto era de esperarse ya que la ocupación de camas, UCIs y salas crematorias destinadas a la pandemia, dependen de los casos, para este trabajo el propósito es aprovechar esa relación en el sentido contrario, observar cuando ese comportamiento ayuda a detectar los casos a futuro.
2. Del grupo de variables del sector financiero parecen tener menos incidencia, pero sin dejar de presentar relación. En cuanto a las variables de las fuentes de los grupos de clima, transporte y seguridad parecen no tener ninguna incidencia significativa respecto al número de casos nuevos de Covid-19 en la ciudad.

Respecto al análisis de las variables se concluye que aquellas que fueron detectadas como independientes, incluyendo independencia con la variable objetivo del conjunto de datos, se encuentran las siguientes: RT, Temp_Min, Temp_Max, Humd_Relativa, Pasajeros_TM, Vuelos, Trafico, Hurtos, Lesiones_Personales, TRM

Para propósito de este trabajo se plantea la opción de excluir las variables de datos independientes para la entrada al modelo, excluirlas en teoría no aportan al resultado de la estimación, sin embargo, antes de excluirlas se tienen en cuenta las siguientes consideraciones:

1. Al ser independientes entre sí, sí se excluyen como variables del modelo no generarán colinealidad en las estimaciones.
2. Al incluirlas en el modelo por la información que aportan estas variables se puede generar entre ellas un efecto combinado que aportaría a la estimación buscada.
3. Las demás variables explicativas del conjunto de datos presentan un nivel alto de asociación entre ellas, de modo que al excluir algunas de ellas el número de variables explicativas en el modelo sería muy reducido.

Finalmente, de estas consideraciones se concluye que todas las variables seleccionadas pese a su independencia o dependencia de la variable objetivo serán consideradas.

En cuanto al grupo de variables con una asociación significativa entre ellas y la variable objetivo, se encuentran: Pruebas_PCR, Casos_Salud, Ocupacion_Camas, Ocupacion_UCI, Ocupacion_Crematorios y TIP.

De las técnicas seleccionadas y sus resultados

De las técnicas seleccionadas para el trabajo y de acuerdo con su aplicación y resultados obtenidos se puede concluir que modelos de regresión basados en algoritmos de árboles de decisión presentaron un menor error de estimación que los algoritmos basados en distancia KNN.

Para el caso de los aboles de decisión; se concluyó que el modelo efectuado con las variables características escogidas por su correlación se obtuvo un resultado optimo frente a los modelos efectuados con las variables obtenidas del análisis de componente principales (PAC) y las obtenidas del análisis de clústeres

Para el caso de regresión mediante K-NN, se concluyo que entre los modelos aplicados para el trabajo se obtienen mejores resultados con la variables seleccionadas a través de la Correlación 1, sin embargo, comparado con las otras dos ejecuciones que se realizaron con la variables de Correlación 2 y escenario exploratorio, la diferencia no es significativa para establecer con una diferencia importante que unos de estos escenarios es mejor que el otro

Cabe resaltar que dentro del análisis se detectó un día del mes de agosto en el que el volumen de nuevos casos positivos de Covid-19 aumento significativamente respecto a los demás días lo que tuvo incidencia en las representaciones graficas de los resultados y que incidieron en los resultados de los modelos ya que no se acercaron a este comportamiento, esto puede valerse a las variables seleccionadas para el análisis.

Oportunidades futuras sobre el planteamiento del proyecto

Como parte del desarrollo del trabajo se han identificado algunas oportunidades futuras para optimización de los resultados tales como:

- Mejorar la consecución de fuentes de información
- Probar más técnicas y comparar
- Probar las técnicas de conjuntos de modelos
- Realizar un proceso de optimización de parámetros

6. Referencias

AprendeIA. (30 de 09 de 2020). Recuperado el 30 de 09 de 2020, de https://aprendeia.com/aplicaciones-de-la-inteligencia-artificial-para-detectar-y-controlar-el-coronavirus/

Avi Schiffmann. (30 de 09 de 2020). Recuperado el 30 de 09 de 2020, de https://ncov2019.live/data

Blog, Machine Learning. (s.f.). *Blog, Machine Learning*. Recuperado el 30 de 09 de 2020

Carlos Eduardo Álvarez Cabrera, E. J. (2015). Modelos epidemiológicos en redes: una presentación introductoria. *22*(1).

Cender Quispe-Juli, P. V.-A.-R.-A. (2020). COVID-19: Una Pandemia en la era de la salud digital.

Coranavirus Colombia Gov. (05 de 10 de 2020). Obtenido de https://coronaviruscolombia.gov.co/Covid19/preguntas-frecuentes.html#:~:text=La%20OMS%20ha%20establecido%20que,2%20a%2014%20d%C3%ADas

D., M.-L. (2020). Uso de tecnologías en el lugar de atención para el manejo de la pandemia por COVID-19 en Colombia. *44*(97).

Facultad de Ingenieria. (2020). *Universidad Catolica de Colombia*. Recuperado el 30 de 09 de 2020, de https://www.ucatolica.edu.co/portal/facultad-de-ingenieria-desarrolla-plataforma-de-analisis-y-prediccion-de-datos-del-covid-19-en-colombia/

geographica.com. (s.f.). Recuperado el 20 de 09 de 2020, de https://geographica.com/es/blog/data-visualization/

Instituto Nacional de Salud (INS). (s.f.). Recuperado el 03 de 10 de 2020, de https://www.ins.gov.co/Noticias/Paginas/Coronavirus-rt.aspx

Llorente y Cuenca. (2020). Situación Colombia COVID-19: gobernanza, visiones y tendencias. *LLORENTE Y CUENCA*, 1 -6.

María Matilde García Lorenzo, Y. R.-H.-G. (2020). Adquisición de conocimiento sobre la letalidad de la COVMaría Matilde García LorenzoID-19 mediante técnicas de inteligencia artificial. *10*(3).

Martin Ester, H.-P. K. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. (226-231).

MIOTI. (2020). *MIOTI*. Recuperado el 30 de 09 de 2020, de https://www.mioti.es/es/data-science-predecir-covid-19/

MONROY, L. G. (2007). *Estadística Multivariada: Inferencia y métodos.* Bogotá: Universidad Nacional de Colombia. Facultad de Ciencias.

Pandas. (03 de 10 de 2020). Obtenido de https://pandas.pydata.org/

Parra, F. (2019). *Estadística y Machine Learning con R.* Editorial Academica Española.

Proyecto Nextstrain. (2020). Recuperado el 30 de 09 de 2020, de https://nextstrain.org/ncov/global

Scikit-learn.org. (03 de 10 de 2020). Obtenido de https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html?highlight=decisiontreeregressor#sklearn.tree.DecisionTreeRegressor

Universidad Johns Hopkins. (2020). Recuperado el 30 de 09 de 2020, de https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Wikidot. (s.f.). Recuperado el 30 de 09 de 2020, de http://redes-neuronales.wikidot.com/definicion-ventajas-desventajas