

Prototipo para analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México

Trabajo Terminal No. _____.-_____-_____-

Alumnos: *Lopez Hernandez David, Escamilla Sánchez Alejandro, Escobedo Domínguez Nadia Gabriela.

Directores: Dr. Zagal Flores Roberto Eswart

*e-mail: dlopezh1702@alumno.ipn.mx

Resumen - En el presente protocolo de trabajo terminal se propone crear un prototipo para el análisis de comportamiento de enfermedades respiratorias en México, a través de técnicas de minería de datos.

Palabras clave – enfermedades respiratorias, data mining, data integration, GIS.

1. Introducción

El impacto de enfermedades respiratorias puede tener consecuencias en el sistema hospitalario, como el nivel de atención a la población, como ha ocurrido durante la propagación de la influenza y la pandemia de Covid-19, es dada por un virus nombrado como SARS-CoV-2. [1] [2] [3], por el que se registraron 4291 decesos y 118,000 casos en 144 países, esto en la fecha de 11 de marzo del 2020. [4] Desde el 2020 hasta la actualidad, la pandemia en México ha provocado diversos efectos en diferentes tipos de industrias y sectores del país. Uno de ellos es el sector salud, siendo el segmento de los servicios médicos especializados el más afectado, mismo que ha presentado una caída en sus ingresos que le condujo a reducir sus gastos hasta en un 30% comparado con 2019 tal como lo muestra un estudio realizado por la Fintech Mexicana Konfío, la cual nos confirma que la alta demanda de los sectores de salud no ha sido completamente beneficiada por esta crisis sanitaria. [5] Durante la pandemia de SARS-COV-2 en la ciudad de México y área metropolitana se presentaron colapsos en algunas instituciones del sistema público de salud [6], obsérvese Gráfico 1, y un deficiente control en algunos estados con esta enfermedad [7]

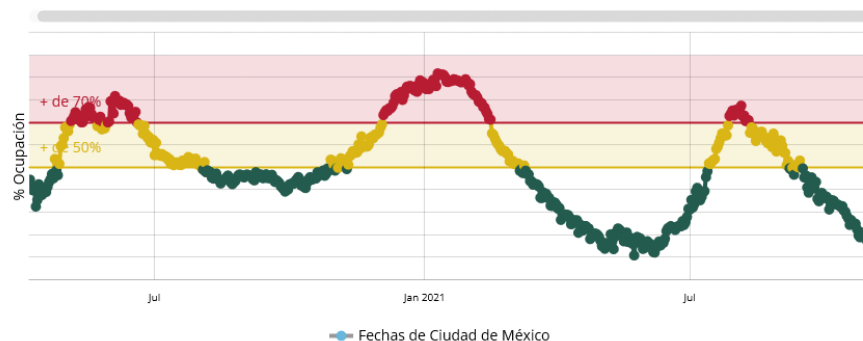


Gráfico 1 – Ocupación hospitalaria en la Ciudad de México [6]

La influenza A, conocida comúnmente como gripe porcina se produce principalmente por la cepa H1N1 del virus de la influenza [8]. Este virus se encuentra dentro de los virus que pueden causar gripe estacional y produce síntomas como la fiebre, tos seca, dolor de garganta, irritación de los ojos, dolores de cuerpo, congestión nasal, fatiga, diarrea, náuseas y vómitos. En 2009 se reconoció por primera vez esta cepa, los primeros casos de influenza en México se detectaron el 11 de abril en el estado de Veracruz, tan solo en los meses de marzo y abril se detectaron más de 1000 casos sospechosos en México y el Suroeste de Estados Unidos [9]. Para finales de abril de 2009 la Organización Mundial de la Salud (OMS) indicaba alerta nivel 5, misma que representaba la transmisión del virus de persona a persona y alertaba a la población de una pandemia inminente. [10]. Hoy se sabe que este virus se trasmite de persona a persona por gotas de saliva o por secreciones respiratorias que viajan en el aire cuando la persona tose, estornuda o escupe o al contacto con superficies contaminadas. Obsérvese Tabla 1 donde se muestran el desarrollo de la enfermedad a nivel mundial a fecha del 2009.

Country, territory and area	Cumulative total		Newly confirmed since the last reporting period	
	Cases	Deaths	Cases	Deaths
United States of America	27717	127	0	0
Mexico	8680	116	401	0
Canada	7983	25	208	4
United Kingdom	6538	3	2288	2
Grand Total	77201	332	6308	21

Tabla 1 – Estadísticas Influenza 2009 [11]

En la actualidad existen diferentes prototipos que muestran la incidencia de la influenza u otras enfermedades respiratorias como por ejemplo tenemos el trabajo que desarrollo Víctor Hugo Borja Aburto, Concepción Grajales Muñiz, Margot González León y Juan Manuel Mejia Aranguré en conjunto con la Coordinación de vigilancia Epidemiológica y Apoyo en contingencias que en conjunto con el Instituto Mexicano del Seguro Social que crearon con base en los métodos de la CDC y por la incidencia por encuestas de seroprevalencia en Londres. [12]

Otro trabajo de investigación que tenemos como parte de estos prototipos es una tesis de grado realizada por los alumnos Erika Andrea Rojas Gutiérrez y Juan Sebastián Aguilar, de la Universidad Católica de Colombia titulado “Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá Colombia”. Donde en este trabajo ellos usaron técnicas de minería de datos y algoritmos de clustering para poder encontrar patrones de incidencia acerca de las enfermedades respiratorias en la ciudad de Bogotá. [13]

1.1 Planteamiento del problema

Existen datos de las enfermedades respiratorias que han impacto en la Salud Pública del Valle de México, algunos datan desde los años 2000's, al integrar estos datos es posible conocer cuál ha sido su comportamiento histórico en el tiempo y en el espacio, cómo ha sido la movilidad de una enfermedad en ciertas regiones y cuáles son las variables asociadas principales asociadas como la edad o el género (dimensiones intrínsecas asociadas a los datos), detectar estos patrones y comportamientos que ayudarían al desarrollo de medidas de prevención, reacción y estimación de un posible impacto de brotes de enfermedades similares en el futuro. En este sentido analizar estos datos de manera multidimensional e histórica implica un esfuerzo considerable en la extracción, transformación y carga de los datos en estructuras que permitan su análisis histórico, en este sentido muchas de las estructuras de estas fuentes varían en su forma en diferentes periodos de tiempo por el cambio organizacional natural en las instituciones de salud.

2. Objetivo General

El presente trabajo tiene como objetivo la elaboración de un prototipo para caracterizar el comportamiento histórico en espacio y tiempo de enfermedades respiratorias que se suscitan en el Valle de México mediante la detección de patrones y tendencias. Analizando enfermedades como la influenza y el COVID tomando información del Sistema abierto de datos de la Dirección General de Información en Salud (DGIS) y de la plataforma de datos abiertos de la Ciudad de México, aplicando algoritmos de Data Mining y enfoques de Sistemas de Información Geográfica.

2.1 Objetivos específicos

- Definir procesos para integrar datos relacionados a enfermedades respiratorias desde fuentes heterogéneas.
- Analizar que fuentes relacionadas se requieren integrar.
- Construir una estructura de almacén de datos para Data Mining.
- Definir procesos de análisis de datos en espacio y tiempo,
- Seleccionar e implementar algoritmos de Data Mining para análisis multidimensional y detección de tendencias y patrones.
- Definir mecanismos de visualización de datos de enfermedades respiratorias.
- Definir una arquitectura de Data Mining que orqueste procesos de integración de datos, análisis multidimensional y detección de patrones, y visualización de datos en web.

3. Justificación

Se sabe que las enfermedades respiratorias están presentes y están en constante cambio y con la pandemia de Covid 19. Este trabajo busca brindar herramientas de apoyo para el sector salud de la República mexicana e investigadores al poner a disposición información que podría ser de utilidad para la creación, revisión e implementación de estrategias que se consideren pertinentes para controlar las enfermedades respiratorias como es la influenza.

Mediante el uso de DataSets abiertos obtenidos de páginas como la Dirección General de Información en Salud, se recopilará información. Se usarán técnicas de minería de datos y con ayuda de algoritmos que permitan observar el comportamiento de este virus en el valle de México tomando en cuenta un periodo razonable de tiempo para obtener datos más concisos y presentar estos resultados en una herramienta que facilite su visualización.

4. Productos o Resultados esperados

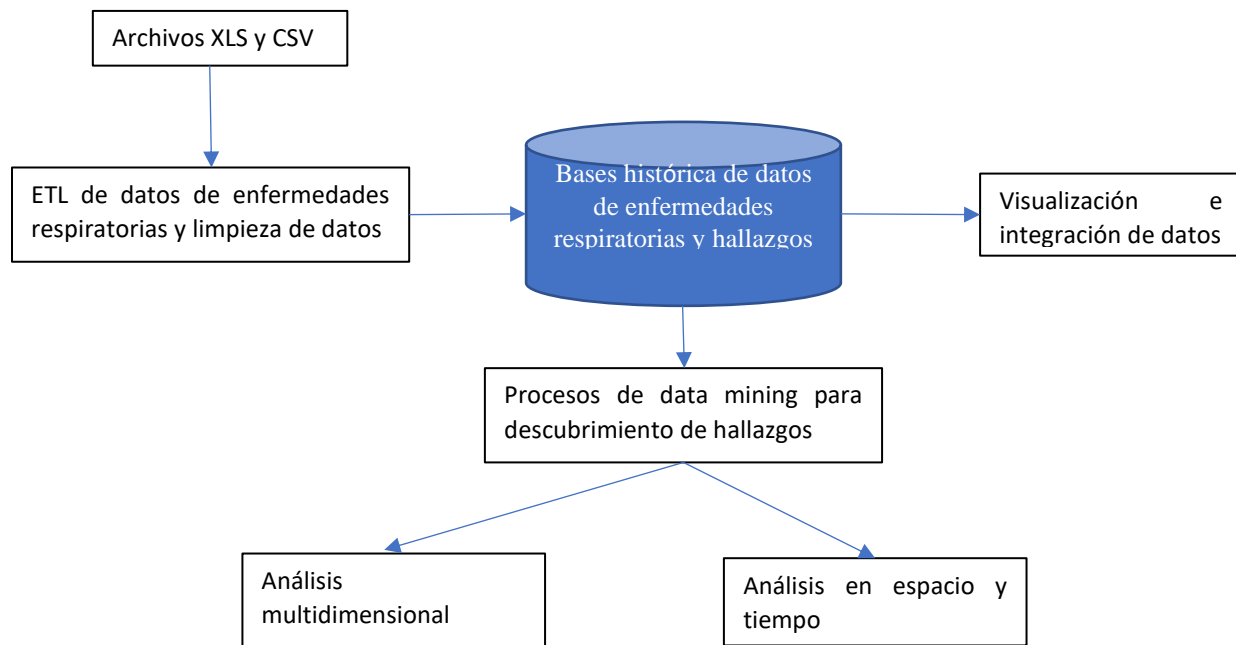


Figura 1 . Arquitectura de sistema propuesta

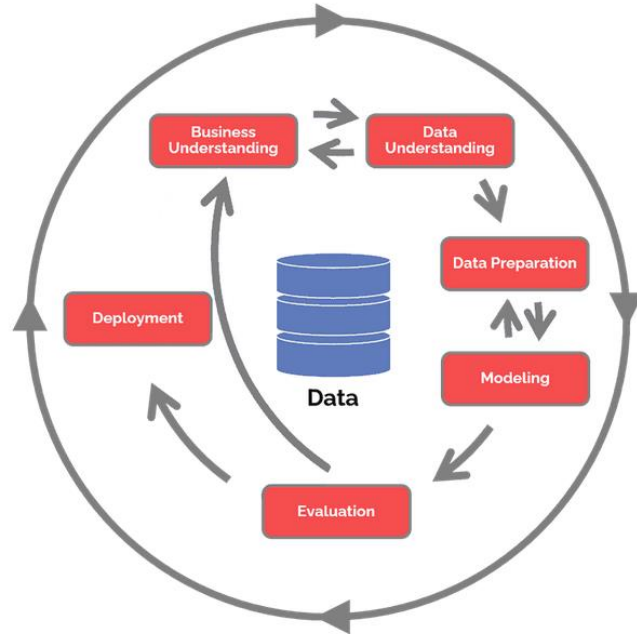
La Figura 1 muestra los componentes propuestos, los datos originalmente están almacenados en archivos XLS y CSV con diferente estructura, para integrarlos se ocupará un proceso ETL que también se encargará de la limpieza y filtrado de datos. Los datos transformados se almacenan en un base de datos histórica, la cual alimenta a los módulos de análisis con data mining para detectar tendencias y patrones en espacio y tiempo. Finalmente, los datos son mostrados en un dashboard.

Productos esperados

- Código fuente del sistema terminado.
- Sistema funcional
- Manual de usuario
- Documentación técnica del sistema
- Aplicación para la visualización del dashboard y prototipo de análisis

5. Metodología

Se utilizará el modelo CRISP-DM que lo conforman las siguientes fases, esto lo vemos conveniente pues es una metodología orientada hacia la minería de datos y que encontramos parecido a algunas metodologías que conocemos ya como Scrum, pero esta al ser especializada en el tema en el que trabajaremos la consideramos una estrategia más sólida al ser un modelo utilizado y con unas bases específicas en la especialización de data mining:



1. **Comprensión del negocio.** En esta fase se centra en entender los objetivos y requerimientos del proyecto y se define el problema de minería de datos. Esto se da en 4 fases
 - a. **Determinar los objetivos:** Para esto se deberá de tomar en cuenta un análisis profundo con respecto a lo que queremos obtener.
 - b. **Evaluar la situación:** Determinar la disponibilidad de recursos, requerimientos de proyectos, evaluar riesgos y contingencias y determinar un análisis de costo-beneficio.
 - c. **Determinar las metas de data mining:** En busca de definir los objetivos financieros, también deberemos de definir cuál será el éxito desde la perspectiva técnica de data mining.
 - d. **Producir el plan de proyecto:** Aquí se selecciona las tecnologías y herramientas y se define detalladamente los planes por cada fase del proyecto.
2. **Comprensión de los datos.** En la segunda fase se identifican, coleccionan y analizan los data sets con los que se planea alcanzar el objetivo del proyecto. Esta etapa consta de cuatro fases.
 - a. **Determinar los objetivos comerciales:** Es decir, se deberá establecer desde un punto de vista comercial el objetivo que se quiere alcanzar y luego se deberán definir los criterios de éxito.
 - b. **Evaluar la situación:** En esta fase se determina la disponibilidad de los recursos, los requisitos, se evalúan los riesgos y contingencias y se realiza un análisis costo-beneficio.
 - c. **Determinar las metas de la minería de datos:** en esta fase también se debe definir como se verá el éxito desde la perspectiva técnica de la minería de datos.

- d. Elaborar un plan de proyecto:** Se deberán seleccionar tecnologías y herramientas y se deberán definir planes detallados para cada una de las fases del proyecto.
- 3. Preparación de datos.** La tercera fase consta de preparar los o el data set final antes del modelado, para lo que se necesitan realizar las siguientes cinco tareas a el conjunto de datos: selección de datos, limpieza de datos, construcción de datos, integración de datos y por último dar formato a los datos.
 - a. Selección de datos:** En esta etapa se determina que data sets será usados y se documentará las razones de la elección o exclusión de un data set.
 - b. Limpieza de datos:** Esta será la fase más larga del proceso y donde se buscará hacer una limpieza lo más eficaz posible evitando los datos basura y el ciclo de garbage-in, garbage-out, que consiste en que cuando sale basura, se crea más basura.
 - c. Construcción de datos:** Es el proceso en el cual se busca crear nuevos datos útiles tomando en cuenta los datos con los que ya contamos.
 - d. Integración de datos:** Creación de nuevos data sets hechos por la combinación de varios datos pertenecientes a diferentes recursos.
 - e. Formato a los datos:** Este será darle el formato necesario para poder interactuar con los datos de manera correcta, está más orientado a los tipos de datos que tenemos y sus compatibilidades.
- 4. Modelado.** Durante la cuarta fase se crearán y evaluarán diferentes modelos basados en técnicas de modelado diferente realizando las siguientes cuatro tareas:
 - a. Seleccionar la técnica de modelamiento:** Determinar que algoritmos se usaran.
 - b. Generar el diseño de prueba:** Es posible que se necesite dividir los datos en conjuntos de entrenamiento prueba y validación.
 - c. Construcción del modelo:** Poner en marcha el modelo.
 - d. Evaluación del modelo:** Se deberá interpretar los resultados del modelo en función a los conocimientos y a los criterios de éxito anteriormente definidos.

Esta fase se repetirá hasta que se encuentre un modelo lo suficientemente bueno para continuar con el modelo CRISP-DM y en el futuro mejorar el modelado.
- 5. Evaluación.** En esta fase se analiza de una manera más amplia que modelo se adapta mejor al negocio y que hacer al final, se realiza en los siguientes 3 pasos.
 - a. Evalúan los resultados:** Se realiza la pregunta “¿Los modelos cumplen con el criterio de aceptación?” y también “¿Cuáles de estos deberíamos aprobar para el negocio?”.
 - b. Se revisa el proceso:** Se revisa el trabajo realizado, se toma en cuenta todo el proceso por el que se transcurrió buscando que si hubo una falla se corrija si es necesario.

- c. **Se determinan los siguientes pasos:** Tomando en cuenta los pasos anteriores, se toman en cuenta tres posibilidades, se procede a hacer un despliegue, se trabaja en la siguiente iteración necesaria o se inicia un nuevo proyecto.
6. **Despliegue.** La última fase del modelo se despliega el plan de proyecto, se monitorea y mantiene y se hace una revisión del proyecto
 - a. **Planificación de la implementación:** Se deberá desarrollar un plan de implementación del modelo.
 - b. **Planificación del seguimiento y mantenimiento:** Se deberá desarrollar un plan de seguimiento y mantenimiento para evitar problemas durante la fase operativa de un modelo.
 - c. **Elaboración de un reporte final:** Se deberá documentar en una recopilación del proyecto donde se puede incluir una presentación final de los resultados obtenidos.
 - d. **Revisión del proyecto:** Se realizará una retrospectiva del proyecto acerca de los buenos resultados, los que pudo salir mejor y como se podría mejorar en el futuro.

Nombre del alumno(a): Escamilla Sánchez Alejandro

TT No.:

Título del TT: Prototipo para la analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México

[illegible]

Nombre del alumno(a): Escobedo Domínguez Nadia Gabriela

TT No.:

Título del TT: Prototipo para la analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México

Actividad	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV
Definición de procesos para integrar los datos.											
Construcción de la estructura de almacén de datos para Data Mining.											
Primera revisión											
Selección de procesos de análisis de datos en espacio y tiempo											
Selección de los algoritmos de Data Mining											
Implementación de los algoritmos de Data Mining											
Evaluación de TT1											
Definir mecanismos de visualización de datos											
Implementación de mecanismos de visualización de datos											
Segunda revisión											
Evaluación de TT2											

7. Referencias

- [1] BBC News, «Coronavirus disease named Covid-19,» BBC News, 11 2 2020. [En línea]. Available: <https://www.bbc.com/news/world-asia-china-51466362>. [Último acceso: 09 11 2021].
- [2] A. E. Gorbalenya, S. C. Baker, R. S. Baric, R. J. de Groot, C. Drosten, B. L. Haagmans, C. Lauber, A. M. Leontovich, B. W. Neuman, D. Penzar, S. Perlma, L. L. Poon, D. Samborskiy, I. A. Sidorov, I. Sola y J. Ziebuhr, «Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group,» *Nature Microbiology*, p. 20, 11 2 2020.
- [3] C. Huang*, Y. Wang*, X. Li*, L. Ren*, J. Zhao*, Y. Hu*, L. Zhang, G. Fan, J. Xu, X. Gu, Z. Cheng, T. Yu, J. Xia, Y. Wei, W. Wu, X. Xie, W. Yin, H. Li, M. Liu, Y. Xiao, H. Gao y L. Guo, «Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,» <https://doi.org/10.1016/>, p. 10, 2020.
- [4] WHO | World Health Organization, «Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19 celebrada el 11 de marzo de 2020,» WHO | World Health Organization, 11 Marzo 2020. [En línea]. Available: <https://www.who.int/es/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>.
- [5] H. Cueto, «Estos son los efectos del Covid-19 en el Sector Salud de México,» Business Insider México | Noticias pensadas para ti, 7 Mayo 2020. [En línea]. Available: <https://businessinsider.mx/estos-son-los-efectos-del-covid-19-en-el-sector-salud-de-mexico/>. [Último acceso: 9 11 2021].
- [6] Desarrollado por el Laboratorio Internacional de Tecnología e Investigación Espacial (iSTAR Lab) del Instituto de Geografía de la UNAM, para la Secretaría de Salud. México ©2020, «Sistema de Información de la Red IRAG,» 2019. [En línea]. Available: <https://www.gits.igg.unam.mx/red-irag-dashboard/reviewHome>. [Último acceso: 9 11 2021].
- [7] IIGEA A.C., «COVID-19 en México - IIGEA A.C.,» IIGEA A.C., [En línea]. Available: <http://iigea.com/amag/covid-19/>. [Último acceso: 9 11 2021].
- [8] Mayo Clinic - Mayo Clinic, «Gripe H1N1 (gripe porcina) - Síntomas y causas - Mayo Clinic,» Mayo Clinic - Mayo Clinic, 29 Julio 2021. [En línea]. Available: <https://www.mayoclinic.org/es-es/diseases-conditions/swine-flu/symptoms-causes/syc-20378103>. [Último acceso: 11 9 2021].
- [9] P. LML., «La influenza A H1N1 en México. Diagnóstico, tratamiento y prevención. Rev Esp Cienc Salud.,» <https://www.medigraphic.com/pdfs/vertientes/vre-2009/vre091-2b.pdf>, p. 12, 2009.
- [10] c. A. Reynoso, «La influenza A (H1N1) y las medidas adoptadas por las autoridades sanitarias,» http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1607-050X2010000100004, p. 18, 2010.
- [11] WHO | World Health Organization, «Pandemic (H1N1) 2009 - update 56,» WHO | World Health Organization, 1 Julio 2009. [En línea]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2009_07_01a-en. [Último acceso: 9 11 2020].
- [12] V. H. B. Aburto*, C. G. Muñoz y M. G. León, «Estimación de la incidencia de influenza pandémica A(H1N1),» https://www.anmm.org.mx/GMM/2011/n4/12_GMM_Vol_147_-_4_2011.pdf, p. 8, 2011.
- [13] E. A. Gutiérrez Rojas y J. S. Aguilar, «Universidad Católica de Colombia,» Universidad Católica de Colombia, 14 Noviembre 2017. [En línea]. Available: <https://repository.ucatolica.edu.co/bitstream/10983/15329/1/Trabajo%20de%20grado.pdf>. [Último acceso: 9 Noviembre 2021].
- [14] Data Science Process Alliance, «CRISP-DM,» Data Science Process Alliance, [En línea]. Available: <https://www.datascience-pm.com/crisp-dm-2/>. [Último acceso: 09 Noviembre 2021].

8. Alumnos y directores

David Lopez Hernandez. - Alumno de la carrera de Ingeniería en Sistemas computacionales en ESCOM, Especialidad Sistemas, Boleta:2018631531, Tel. 55-50512742, email dlopezh1702@alumno.ipn.mx

Firma: _____

Alejandro Escamilla Sánchez. - Alumno de la carrera de Ingeniería en Sistemas computacionales en ESCOM, Especialidad Sistemas, Boleta: 2015130337, Tel. 5534333175, email aescamillas1400@alumno.ipn.mx

Firma: _____

Nadia Gabriela Escobedo Domínguez. - Alumno de la carrera de Ingeniería en Sistemas computacionales en ESCOM, Especialidad Sistemas, Boleta:2014080391, Tel. 55-74366113, email nescobedod1300@alumno.ipn.mx

Firma: _____


Dr. Roberto Zagal Flores. Es egresado de la Ingeniería en Sistemas Computacionales de la Escuela Superior de Cómputo del IPN, culminó sus estudios de Maestría en Ciencias de la Computación en el Centro de Investigación en Computación del IPN (No. Cedula 11050111). Tiene un doctorado en tecnología avanzada de la Sección de Estudios de Posgrado de la UPIITA IPN. Actualmente es profesor de la Escuela Superior de Cómputo y sus áreas de interés son Data Mining, Spatial Data Mining, GIS, Web Semántica, Data Integration, IoT y Arquitecturas de Sistemas de Información. Ha trabajado en proyectos de tecnología en la iniciativa privada y en el sector público. Email: zagalmmx@hotmail.com, Tel. 57296000, Ext. 52032.

Firma: _____

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Firma David Lopez Hernandez

Protocolo de trabajo de terminal

 **David Lopez Hernandez** <dlopezh1702@alumno.ipn.mx>
02:18 a. m.


Para: David Lopez Hernandez Cc: Nadia Gabriela Escobedo Dominguez; Alejandro Escamilla Sanchez

Yo David Lopez Hernandez estudiante con numero de boleta 2018631531 de la carrera Ingeniería en Sistemas Computacionales estoy de acuerdo en trabajar en el protocolo de trabajo terminal.

"Prototipo para analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México".

Firma Alejandro Escamilla

Firma de participación para el TT "Prototipo para analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México"


 **Alejandro Escamilla Sanchez** <aescamillas1400@alumno.ipn.mx>
02:00 a. m.

Para: David Lopez Hernandez Cc: Nadia Gabriela Escobedo Dominguez

Yo Alejandro Escamilla Sánchez estudiante con numero de boleta 2015130337 de la carrera Ingeniería en Sistemas Computacionales estoy de acuerdo en trabajar en el protocolo de trabajo terminal.

"Prototipo para analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México".

Firma Nadia Gabriela Escobedo Domínguez

 **Nadia Gabriela Escobedo Dominguez** <nescobedod1300@alumno.ipn.mx>
02:21 a. m.


Para: David Lopez Hernandez

Yo Nadia Gabriela Escobedo Domínguez estudiante con numero de boleta 2014080391 de la carrera Ingeniería en Sistemas Computacionales estoy de acuerdo en trabajar en el protocolo de trabajo terminal.

"Prototipo para analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México".

Firma Director:

Re: Dudas respecto a el TT - Análisis socio-espacio temporal de datasets de enfermedades respiratorias

 **Roberto Zagal** <zagalmmx@gmail.com>
12:03 a. m.

Para: David Lopez Hernandez Cc: Alejandro Escamilla Sanchez; Nadia Gabriela Escobedo Dominguez

Estoy de acuerdo con formar parte del

"Prototipo para la analizar el comportamiento e impacto de enfermedades respiratorias en el Valle de México ".

Muchas gracias