Universidad Nacional Autónoma de México
Facultad de Ingeniería
División de Ingeniería Eléctrica
Ingeniería en Computación

Proyecto 2
“Arquitectura de Minería de Datos – Diabetes en México”

Integrantes:
- Castelan Ramos Carlos
- Bouchan Ramírez Abraham

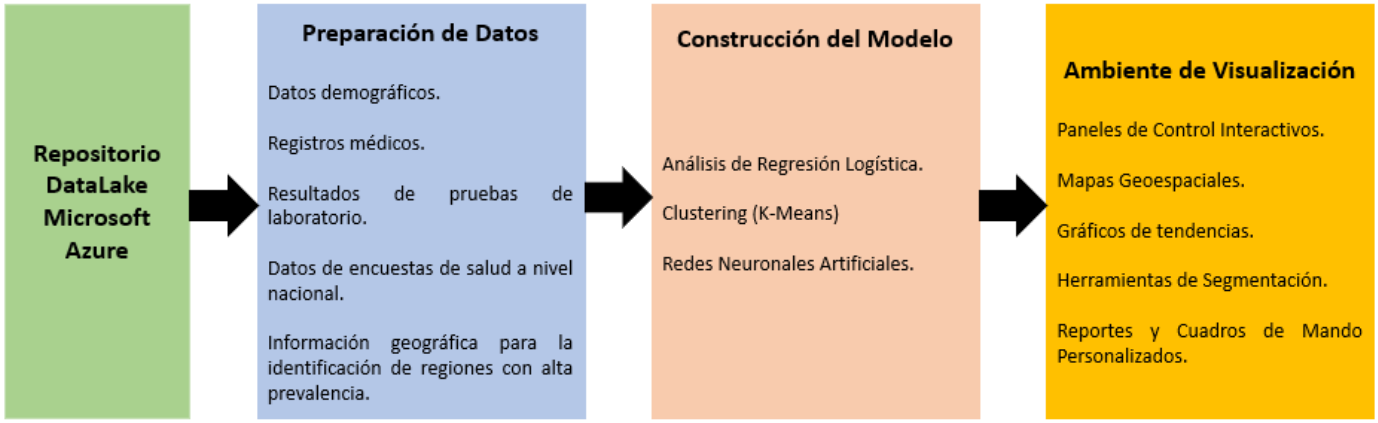
Materia: Minería de Datos
Grupo: 03
Semestre: 2024-1

Fecha de entrega: 14 de septiembre 2023

Proyecto 2 “Arquitectura de Minería de Datos – Diabetes en México”

Basado en un caso de Negocio elegido por cada grupo

- Describir el caso de Negocio y su propósito.
  - Industria: Salud Pública.
  - Tipo de Negocio: Desarrollo de una Campaña de Información sobre la Diabetes Utilizando Minería de Datos.
  - Cobertura: Nacional (con un enfoque en áreas de alta prevalencia de diabetes).
  - Propósito del negocio: La campaña busca abordar la falta de conciencia y educación sobre la diabetes, ya que la diabetes es una enfermedad en constante crecimiento en México, y la falta de información precisa y personalizada sobre la prevención y el manejo adecuado de la enfermedad contribuye a su propagación y al aumento de complicaciones.
- Crear Diagrama de la Arquitectura de Minería de Datos propuesta.



Arquitectura de Minería de Datos para Identificar datos relevantes para identificar causas de Diabetes en México.

  - Repositorio:
    - Data Lake.
  - Preparación de datos:
    - Datos demográficos de la población mexicana:
      La edad, el género y otros datos demográficos son fundamentales para identificar grupos de riesgo, ya que la diabetes puede afectar a diferentes segmentos de la población de manera diferente. Por ejemplo, la prevalencia de diabetes tiende a aumentar con la edad, y existen diferencias de género en la incidencia de la enfermedad. Estos datos ayudarán a personalizar las estrategias de prevención y tratamiento.
    - Registros médicos de pacientes con diabetes:
      Estos registros contienen información valiosa sobre la evolución de la enfermedad en pacientes ya diagnosticados. Los datos médicos, como los niveles de glucosa en sangre, los tratamientos previos y los resultados de las pruebas de laboratorio, son esenciales para comprender el manejo actual de la enfermedad y evaluar la eficacia de los tratamientos existentes.
    - Resultados de pruebas de laboratorio relacionadas con la diabetes:
      Los datos de laboratorio, como los niveles de glucosa en sangre, la hemoglobina A1c y otros marcadores, son indicativos del control de la diabetes y su impacto en la salud del paciente. Estos datos permiten evaluar la gravedad de la enfermedad y ajustar los enfoques de tratamiento.

- Datos de encuestas de salud a nivel nacional:
  Las encuestas de salud proporcionan información sobre los hábitos de vida, la dieta, la actividad física y la percepción de la diabetes entre la población. Estos datos son esenciales para comprender las causas subyacentes del crecimiento de la enfermedad y diseñar estrategias de prevención basadas en la educación y la concienciación pública.
- Información geográfica para la identificación de regiones con alta prevalencia de diabetes:
  La diabetes puede variar significativamente en su prevalencia según la ubicación geográfica. Identificar las regiones con tasas más altas de diabetes ayudará a dirigir los recursos de manera eficiente y priorizar intervenciones específicas en áreas críticas.

➢ Construcción de modelos:
- Análisis de Regresión Logística:
  Se utilizará para identificar factores de riesgo relacionados con la diabetes, como la edad, el género, el índice de masa corporal (IMC) y los hábitos de vida.
- Clustering (K-Means):
  Se aplicará para segmentar a la población en grupos con perfiles de riesgo similares, lo que ayudará a personalizar las estrategias de prevención y tratamiento.
- Redes Neuronales Artificiales (ANN):
  Se utilizarán para predecir la probabilidad de que una persona desarrolle diabetes en función de múltiples variables, incluidos los antecedentes familiares y los hábitos de vida.

➢ Ambiente de Visualización:
Se implementará un sistema de visualización de datos que permita a los usuarios, como ejecutivos de salud pública, supervisores y profesionales de la salud, interactuar y analizar los resultados de manera efectiva. Esto podría incluir:
- Paneles de Control Interactivos:
  Paneles personalizados que muestren métricas clave, tendencias de prevalencia de la diabetes y alertas sobre regiones críticas.
- Mapas Geoespaciales:
  Visualización geográfica de la prevalencia de la diabetes en diferentes regiones de México para identificar áreas de enfoque.
- Gráficos de Tendencias:
  Gráficos interactivos que representen la evolución de la diabetes a lo largo del tiempo y los resultados de intervenciones específicas.
- Herramientas de Segmentación:
  Permitir la segmentación de la población en grupos de riesgo y el análisis detallado de cada grupo.
- Reportes y Cuadros de Mando Personalizados:
  La capacidad de generar informes y cuadros de mando personalizados para satisfacer las necesidades individuales de los usuarios.

- Documentar los siguientes componentes de la arquitectura:
  ➢ Repositorios / Servidores de Datos participantes.
  - Repositorio elegido:

Data Lake (DL):

Un Data Lake podría ser una opción sólida si se espera trabajar con una amplia gama de datos, como registros médicos no estructurados o datos de encuestas de salud. Esto permitiría una mayor flexibilidad en el análisis y la inclusión de datos no convencionales.

- Ventajas:
  Los Data Lakes son ideales para almacenar datos en su formato original, incluyendo datos no estructurados y semiestructurados. Son altamente escalables y ofrecen flexibilidad para explorar y analizar una amplia variedad de datos.
- Desventajas:
  Pueden requerir más esfuerzo en la preparación y limpieza de datos, y la gestión de metadatos y la seguridad pueden ser desafíos.

Aplicación inmediata en el caso de negocio:

- Variedad de Datos:
  El análisis de la diabetes requiere la consideración de una amplia variedad de datos, incluyendo registros médicos, resultados de pruebas de laboratorio, datos demográficos, encuestas de salud y más. Los Data Lakes están diseñados para manejar datos en su formato original, lo que permite la inclusión de datos estructurados, semiestructurados y no estructurados sin necesidad de una estructura predefinida.
- Flexibilidad:
  Un Data Lake ofrece flexibilidad para explorar y analizar datos de manera ágil. Dado que el problema de la diabetes es multifacético y puede involucrar datos de diversas fuentes, la capacidad de adaptarse rápidamente a nuevas fuentes de datos y tipos de datos es esencial.
- Escalabilidad:
  La diabetes es un problema de salud que afecta a una gran población. Un Data Lake es altamente escalable y puede manejar grandes volúmenes de datos a medida que se recopilan más información con el tiempo.
- Análisis Avanzado:
  Un Data Lake permite la aplicación de algoritmos de Minería de Datos y Aprendizaje Automático (Machine Learning) sobre datos en su formato original. Esto es esencial para realizar análisis avanzados y descubrir patrones complejos relacionados con la diabetes.
- Integración de Datos de Múltiples Fuentes:
  Dado que se requieren datos de diversas fuentes, un Data Lake puede integrar fácilmente datos provenientes de registros médicos, encuestas de salud, laboratorios clínicos y otras fuentes relevantes.
- Preparación y Limpieza de Datos:
  Aunque la preparación y limpieza de datos puede requerir esfuerzo en un Data Lake, las herramientas y tecnologías modernas permiten abordar estos desafíos de manera efectiva.

- Servidor de datos elegido:

Microsoft Azure:

Azure ofrece una amplia gama de servicios, incluyendo Azure Data Lake Storage (la cual fue elegida)y Azure Synapse Analytics (anteriormente conocido como Azure SQL Data Warehouse). Azure

cuenta con capacidades sólidas de análisis y aprendizaje automático, así como una integración estrecha con herramientas de Microsoft como Power BI para la visualización de datos.
Además, la capacidad de implementación de otras herramientas como otros manejadores de BD y herramientas de creación de funciones Front End permiten una rápida adaptación e implementación del proyecto.

- Históricos y frecuencia:
  La consulta de datos se lleva a cabo en diferentes escalas y diferentes tiempos, los cuales se establecieron y categorizaron de la siguiente forma.
  - Datos Históricos a Largo Plazo:
    Para comprender las tendencias a lo largo del tiempo y las causas subyacentes del crecimiento de la diabetes en México, es importante contar con datos históricos a largo plazo.
    Frecuencia Histórica: Estos datos podrían abarcar varias décadas, proporcionando una visión de largo plazo de la prevalencia de la diabetes, cambios en la dieta, el estilo de vida y otros factores relevantes. Los datos anuales o incluso decenales pueden ser adecuados para este propósito.

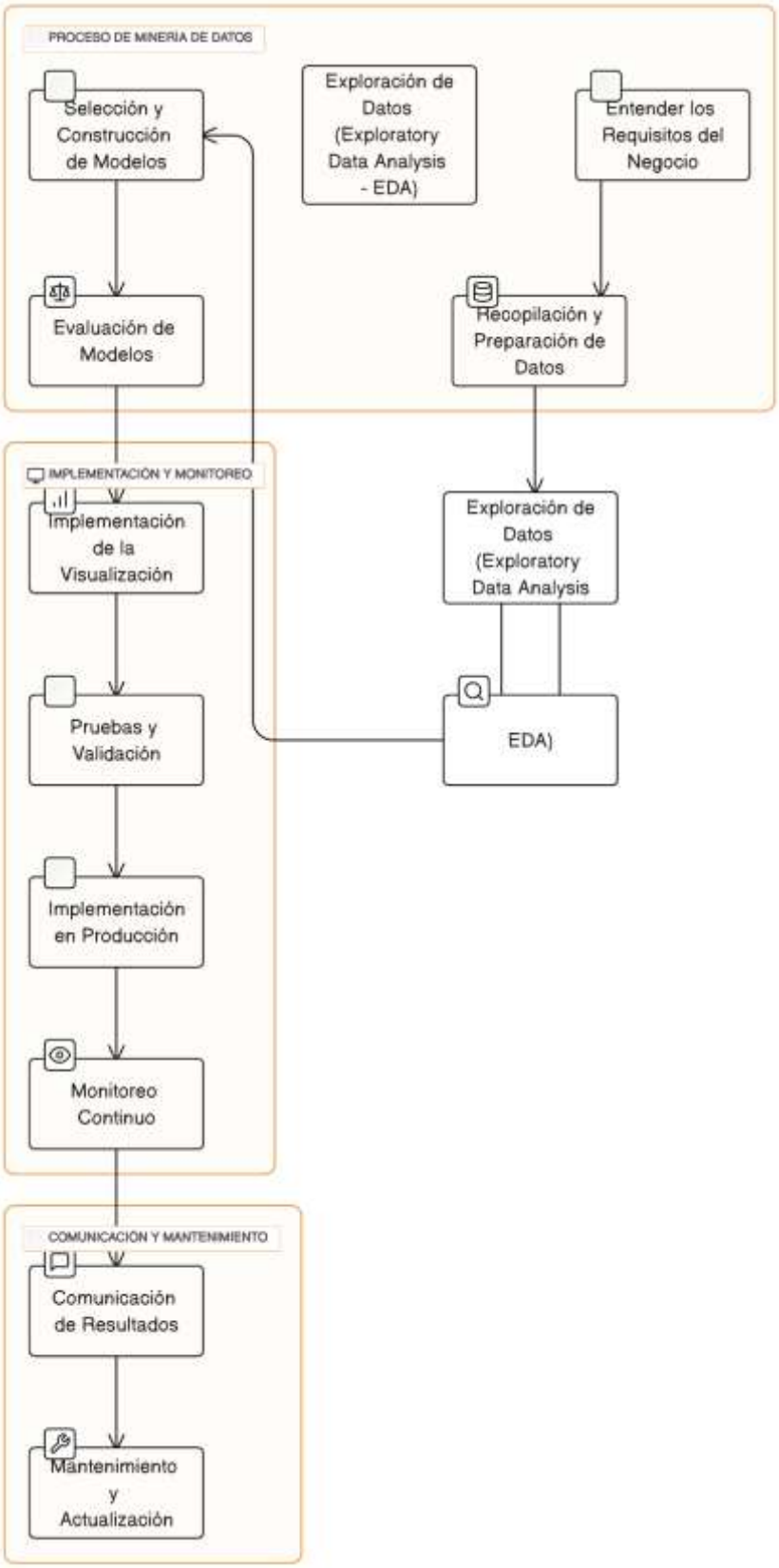
  - Datos Actuales y de Corto Plazo:
    Para tomar decisiones informadas y desarrollar estrategias de prevención y tratamiento actuales, es esencial contar con datos más recientes y de corto plazo.
    Frecuencia Histórica: Los datos actuales podrían tener una frecuencia mensual, trimestral o anual, dependiendo de la disponibilidad de fuentes de datos actualizadas y del ritmo de cambio en los indicadores clave de la diabetes, como la prevalencia y los factores de riesgo.

  - Datos de Encuestas y Estudios Puntuales:
    Para obtener información detallada sobre hábitos de vida, dieta, actividad física y percepciones sobre la diabetes entre la población, es necesario realizar encuestas y estudios puntuales.
    Frecuencia Histórica: Estos datos pueden tener una frecuencia variable y se pueden realizar de manera puntual o en intervalos regulares según las necesidades de investigación. Por ejemplo, una encuesta de salud nacional podría llevarse a cabo cada varios años.

  - Datos de Laboratorio Clínico y Registros Médicos:
    Para evaluar la gravedad y el control de la diabetes en pacientes, es esencial contar con datos de laboratorio y registros médicos.
    Frecuencia Histórica: Estos datos pueden tener una frecuencia más alta, como mensual o trimestral, para un seguimiento más detallado de la evolución de los pacientes con diabetes.

- Tipo de extracción y preparación de los datos:
  - Extracción de Datos de Registros Médicos:
    Se obtendrían datos de registros médicos de pacientes que han sido diagnosticados con diabetes. Estos registros pueden incluir información sobre diagnósticos, tratamientos, resultados de pruebas de laboratorio y seguimiento médico.
  - Extracción de Datos de Encuestas de Salud:
    Se recopilarían datos de encuestas de salud nacionales o regionales que contengan información sobre hábitos de vida, dieta, actividad física, conocimiento de la diabetes y percepciones sobre la salud.
  - Extracción de Datos de Laboratorios Clínicos:
    Se obtendrían datos de laboratorios clínicos que realicen pruebas relacionadas con la diabetes, como mediciones de glucosa en sangre, hemoglobina A1c y otros marcadores.
  - Extracción de Datos Demográficos:
    Se recopilarían datos demográficos de la población mexicana, incluyendo edad, género, ubicación geográfica y otros factores relevantes.
  - Extracción de Datos de Estudios de Investigación:
    Se podrían utilizar datos de estudios de investigación previos relacionados con la diabetes en México, si están disponibles. Estos estudios pueden proporcionar información valiosa sobre factores de riesgo y tendencias.
  - Extracción de Datos de Fuentes Abiertas y Gobierno:
    Se podrían utilizar datos de fuentes abiertas y gubernamentales que proporcionen información sobre la prevalencia de la diabetes, políticas de salud, acceso a servicios médicos y otros factores relacionados con la enfermedad.
  - Extracción de Datos de Registros de Seguros de Salud:
    Si es relevante, se podrían obtener datos de compañías de seguros de salud que tengan información sobre pacientes con diabetes, tratamientos y costos médicos asociados.

- Diagrama del Flujo de las actividades a realizar en el proceso de Minería de Datos, y documentar sus actividades



1. Entender los Requisitos del Negocio:

- Comprender en detalle los objetivos y las necesidades específicas de la Campaña de Información sobre la Diabetes, así como las preguntas de negocio que se deben responder mediante el análisis de datos.

2. Recopilación y Preparación de Datos:
   - Reunir y limpiar los datos necesarios de las fuentes identificadas, incluyendo datos demográficos, registros médicos, resultados de pruebas de laboratorio, datos de encuestas y geográficos.
   - Realizar la integración de datos de diferentes fuentes en el repositorio Data Lake.

3. Exploración de Datos (Exploratory Data Analysis - EDA):
   - Realizar un análisis exploratorio de los datos para comprender su distribución, calidad y relaciones entre variables.
   - Identificar posibles valores atípicos o datos faltantes que requieran atención.

4. Selección y Construcción de Modelos:
   - Seleccionar los algoritmos de minería de datos apropiados según los objetivos del proyecto, como regresión logística, clustering y redes neuronales artificiales.
   - Entrenar y validar los modelos utilizando los datos preparados.

5. Evaluación de Modelos:
   - Evaluar la calidad y el rendimiento de los modelos mediante métricas adecuadas, como precisión, recall, F1-score, etc.
   - Ajustar y optimizar los modelos según sea necesario.

6. Implementación de la Visualización:
   - Desarrollar el sistema de visualización de datos que incluye paneles de control interactivos, mapas geoespaciales, gráficos de tendencias y herramientas de segmentación.
   - Asegurarse de que los usuarios puedan interactuar y analizar los resultados de manera efectiva.

7. Pruebas y Validación:
   - Realizar pruebas exhaustivas del sistema de minería de datos y la visualización para garantizar su funcionamiento correcto y la precisión de los resultados.
   - Validar que los resultados obtenidos sean coherentes con los objetivos del negocio.

8. Implementación en Producción:
   - Llevar a cabo la implementación en producción del sistema, asegurando que esté listo para su uso en la Campaña de Información sobre la Diabetes.

9. Monitoreo Continuo:
   - Establecer un sistema de monitoreo continuo para supervisar el rendimiento de los modelos y la visualización en tiempo real.
   - Realizar ajustes y mejoras a medida que se recopila más información y se obtienen retroalimentaciones.

10. Comunicación de Resultados:
- Comunicar de manera efectiva los resultados del análisis de datos y las conclusiones al equipo de Salud Pública y otras partes interesadas.

- Proporcionar recomendaciones basadas en los hallazgos para mejorar la prevención y el manejo de la diabetes.

11. Mantenimiento y Actualización:

- Continuar manteniendo y actualizando el sistema de minería de datos y la visualización a medida que se requieran cambios en los datos o se identifiquen nuevas oportunidades.

- Documentar los modelos de minería de datos participantes y esquemas de monitoreo y validación

Modelos de Minería de Datos Participantes:

- Modelo de Regresión Logística:
  - Objetivo: Identificar factores de riesgo relacionados con la diabetes, como edad, género, IMC y hábitos de vida.
  - Variables de Entrada: Datos demográficos y hábitos de vida.
  - Métricas de Evaluación: Precisión, recall, F1-score.
  - Hiperparámetros: Coeficientes de regresión, umbral de clasificación.
- Modelo de Clustering (K-Means):
  - Objetivo: Segmentar a la población en grupos con perfiles de riesgo similares.
  - Variables de Entrada: Datos demográficos, resultados de pruebas de laboratorio, hábitos de vida.
  - Métricas de Evaluación: Coeficiente de silueta, inercia.
  - Hiperparámetros: Número de clusters.
- Modelo de Redes Neuronales Artificiales (ANN):
  - Objetivo: Predecir la probabilidad de desarrollar diabetes en función de múltiples variables, incluidos antecedentes familiares y hábitos de vida.
  - Variables de Entrada: Datos demográficos, antecedentes familiares, hábitos de vida.
  - Métricas de Evaluación: Precisión, pérdida logarítmica.
  - Hiperparámetros: Número de capas ocultas, número de neuronas, función de activación.

Esquemas de Monitoreo y Validación:

- Monitoreo Continuo de Modelos:
  - Establecer un sistema de monitoreo en tiempo real para los modelos de regresión logística y ANN.
  - Supervisar las métricas de rendimiento (precisión, recall, F1-score) y las tasas de falsos positivos y falsos negativos.
  - Definir umbrales de alarma para detectar degradación del rendimiento.
- Monitoreo de Cambios en los Datos:
  - Implementar una detección de cambios en los datos para identificar desviaciones significativas en los datos de entrada que puedan afectar los modelos.
  - Actualizar los modelos o reentrenarlos según sea necesario en respuesta a cambios significativos en los datos.
- Validación Periódica de Modelos:
  - Realizar validaciones periódicas de los modelos de clustering para garantizar que los grupos identificados sigan siendo relevantes y útiles.
  - Utilizar métricas como el coeficiente de silueta para evaluar la coherencia de los clusters.
- Validación con Datos de Prueba:

  - Utilizar conjuntos de datos de prueba separados para evaluar regularmente el rendimiento de los modelos de regresión logística y ANN.
  - Calcular métricas de rendimiento en los datos de prueba y compararlas con las métricas en los datos de entrenamiento.
- Actualización de Modelos:
  - Establecer un plan para actualizar los modelos de regresión logística y ANN en función de cambios en los datos o necesidades de mejora del rendimiento.
  - Documentar las versiones de modelos y los cambios realizados en cada actualización.
- Auditorías de Modelos:
  - Realizar auditorías periódicas de los modelos para asegurarse de que sigan siendo éticos y no sesgados en sus decisiones.
- Comunicación de Resultados:
  - Comunicar regularmente los resultados de monitoreo y validación al equipo de Salud Pública y otras partes interesadas para la toma de decisiones informadas.

- Documentar los beneficios que se obtendrían con la propuesta

La propuesta de desarrollar una Campaña de Información sobre la Diabetes utilizando Minería de Datos en la industria de la Salud Pública en México ofrece varios beneficios significativos para la salud pública y la sociedad en general. A continuación, se documentan los beneficios clave que se obtendrían con esta propuesta:

- Concientización y Educación Mejoradas:
  - Uno de los beneficios más destacados es la mejora en la concienciación y educación sobre la diabetes en la población mexicana. La campaña proporcionará información precisa y personalizada sobre la prevención y el manejo adecuado de la enfermedad, lo que contribuirá a una mayor conciencia pública sobre la diabetes y sus riesgos.
- Reducción de la Prevalencia de la Diabetes:
  - Al identificar factores de riesgo y grupos de población en mayor riesgo de desarrollar diabetes, la campaña permitirá la implementación de estrategias de prevención dirigidas. Esto puede ayudar a reducir la incidencia de la enfermedad a largo plazo.
- Mejor Control de la Diabetes:
  - La campaña también beneficiará a las personas que ya han sido diagnosticadas con diabetes al proporcionar información sobre el manejo adecuado de la enfermedad. Esto puede llevar a un mejor control de la diabetes, reduciendo las complicaciones a largo plazo y mejorando la calidad de vida de los pacientes.
- Optimización de Recursos de Salud:
  - Identificar las regiones con alta prevalencia de diabetes permitirá una asignación más eficiente de recursos de salud pública. Esto incluye la distribución de servicios de atención médica y programas de prevención en áreas críticas.
- Personalización de Estrategias de Salud Pública:
  - La minería de datos permitirá segmentar a la población en grupos con perfiles de riesgo similares. Esto facilitará la personalización de las estrategias de salud pública, lo que significa que se pueden adaptar enfoques específicos para diferentes grupos demográficos.
- Toma de Decisiones Informada:

  - La información basada en datos y los resultados de la minería de datos respaldarán la toma de decisiones informadas por parte de los profesionales de la salud pública y los responsables de políticas. Esto conducirá a políticas y programas más efectivos y basados en evidencia.
- Mejora de la Salud Pública en General:
  - Al abordar la diabetes, una enfermedad crónica de alto impacto, se contribuirá a una mejora general de la salud pública en México. Esto puede llevar a una población más saludable y a una reducción de los costos de atención médica relacionados con la diabetes y sus complicaciones.
- Reducción de la Carga Económica:
  - La diabetes impone una carga significativa en los sistemas de atención médica y la economía en términos de costos médicos y productividad perdida. La campaña puede contribuir a la reducción de esta carga económica a largo plazo.
- Generación de Conocimiento Continuo:
  - A través del monitoreo y la actualización constante de los modelos de minería de datos, se generará un conocimiento continuo sobre la diabetes y su evolución en México, lo que permitirá ajustar las estrategias a medida que cambien las circunstancias.

En resumen, la propuesta de desarrollar una Campaña de Información sobre la Diabetes utilizando Minería de Datos tiene el potencial de generar un impacto positivo significativo en la salud pública en México al aumentar la concienciación, mejorar la prevención y el manejo de la diabetes, y optimizar la asignación de recursos y políticas de salud pública. Además, promoverá la utilización ética de la minería de datos para el beneficio de la sociedad.