predicción del riesgo de diabetes a partir de indicadores de salud y estilo de vida

objetivo:

El objetivo principal de este proyecto es desarrollar un modelo de análisis y predicción del riesgo de diabetes usando indicadores de salud, hábitos y factores sociales, aplicando técnicas de ciencia de datos y aprendizaje automático. La idea es crear una herramienta confiable y fácil de usar en entornos clínicos y de salud pública que ayude a detectar temprano a personas con alto riesgo de desarrollar diabetes.

La diabetes, especialmente el tipo 2, ha crecido mucho en el mundo por cambios en los estilos de vida, aumento del sobrepeso y obesidad, además de factores genéticos y sociales. Pero la mayoría de los casos se pueden prevenir o retrasar si se detectan a tiempo los factores de riesgo. Un modelo predictivo con datos grandes tiene el potencial de convertirse en una herramienta importante para prevenir la enfermedad, tanto en la población general como en grupos específicos en riesgo.

Este proyecto quiere:

1. Identificar y analizar los factores de riesgo más importantes en la aparición de diabetes, como la obesidad, el índice de masa corporal (IMC), la hipertensión, el colesterol alto, la falta de actividad física, el consumo de tabaco, alcohol y los patrones alimenticios. Es importante distinguir entre factores clínicos que no se pueden cambiar (como la edad o la predisposición genética) y aquellos que sí se pueden modificar con educación y campañas de salud pública.
2. Crear modelos de predicción que sean confiables y puedan clasificar correctamente a las personas según su riesgo de tener diabetes. Para esto, se usarán algoritmos de aprendizaje automático, como regresión logística y árboles de decisión.
3. Estudiar cómo influyen variables sociodemográficas, como nivel de educación, ingresos y edad, en el riesgo de desarrollar diabetes. Aunque estos factores no están directamente relacionados con el metabolismo, afectan los hábitos de vida y el acceso a servicios de salud, por lo que son importantes para diseñar políticas públicas que incluyan a todos.
4. Generar gráficos y reportes fáciles de entender que ayuden a médicos e investigadores a entender mejor cómo se relacionan las variables. No solo queremos crear un modelo para predecir, sino también ofrecer información clara y útil para programas de prevención y campañas educativas para la población.

De manera general, el objetivo es cerrar la brecha entre el diagnóstico tradicional y el diagnóstico asistido por datos. Se busca demostrar cómo el análisis de grandes volúmenes de información, bien organizado y validado, puede mejorar la salud pública y la calidad de vida de muchas personas.

En la práctica, se espera que el modelo final aumente en al menos un 20% la detección temprana de personas en riesgo, en comparación con los métodos tradicionales basados solo en pruebas clínicas o cuestionarios de riesgo. Este valor ayudará a medir qué tan bien funciona el proyecto y servirá para implementarlo en sistemas de salud digitales.

En resumen, el objetivo no es solo predecir de forma individual, sino también producir conocimientos útiles, proponer intervenciones tempranas y promover una estrategia preventiva para una enfermedad que es uno de los mayores desafíos para la salud en todo el mundo.

Descripción del Problema

La diabetes se ha convertido en uno de los principales problemas de salud pública a nivel mundial. Es una enfermedad crónica en la que el cuerpo no regula bien los niveles de azúcar en la sangre. Actualmente, más de 420 millones de personas en todo el mundo tienen esta condición, según la Organización Mundial de la Salud (OMS). Lo preocupante no solo es cuántas personas la tienen, sino también que cada vez hay más casos. Esto se debe a factores como el aumento del sobrepeso y la obesidad, los cambios en los hábitos alimenticios, el sedentarismo y el envejecimiento de la población.

Uno de los grandes problemas con la diabetes es el diagnóstico tardío. Muchas personas no saben que tienen la enfermedad hasta que aparecen complicaciones graves, como problemas renales, ceguera, dolores nerviosos, amputaciones o problemas del corazón. Estas complicaciones no solo afectan mucho la calidad de vida, sino que también aumentan considerablemente los costos de atención médica en hospitales y sistemas de salud. Detectar la enfermedad temprano es muy importante para evitar que empeore y para reducir sus efectos. Sin embargo, las formas tradicionales, que solo usan pruebas clínicas, no siempre alcanzan a toda la población.

Por eso, es importante buscar nuevas formas de detectar y predecir la diabetes, usando datos y técnicas de análisis avanzado. La disponibilidad de grandes bases de datos con información sobre la salud, hábitos y características de las personas crea una oportunidad única para encontrar patrones de riesgo y detectar la enfermedad antes de que sea irreversible. El reto es que los datos deben ser limpiados, analizados y convertidos en información útil, que sirva para tomar decisiones en los sistemas de salud.

Otro aspecto importante es el papel de los factores sociales y económicos en el desarrollo de la enfermedad. El nivel de ingresos, la educación y el acceso a los servicios de salud influyen en los hábitos alimenticios, el ejercicio y la posibilidad de realizar chequeos médicos. Por ello, la diabetes no es solo un problema biológico, sino también un problema multidimensional, que involucra aspectos médicos, conductuales y sociales.

Este proyecto busca abordar esas dificultades creando un modelo predictivo que combine diferentes indicadores de salud y estilo de vida. La idea es identificar cuáles son los factores más importantes que influyen en la aparición de la diabetes. Con este análisis, se busca no

solo entender mejor los riesgos, sino también crear estrategias de prevención más efectivas y dirigidas.

Desde un punto de vista práctico, el problema principal es la falta de mecanismos fáciles de usar, accesibles y basados en datos confiables, para detectar la diabetes temprano a gran escala. La ciencia de datos y la mayor disponibilidad de información en los sistemas de salud pueden ayudar a cerrar esta brecha. Identificar a las personas en riesgo permite usar mejor los recursos, orientar campañas preventivas y, en última instancia, reducir el impacto social, económico y en la salud que causa la diabetes.

En resumen, la clave de este proyecto es la necesidad de prevenir y diagnosticar la diabetes temprano, usando herramientas basadas en datos.

Recursos disponibles

1. Tecnología y Herramientas

El análisis de los indicadores de salud relacionados con la diabetes requiere apoyarse en herramientas tecnológicas modernas que faciliten el manejo de grandes volúmenes de información, la aplicación de métodos estadísticos y la construcción de modelos predictivos. Para lograrlo, se utilizan diferentes recursos clave:

- Python: un lenguaje de programación ampliamente adoptado en el ámbito de la ciencia de datos, valorado por su flexibilidad y por la enorme variedad de librerías disponibles.
- Pandas y NumPy: dos librerías fundamentales para procesar y analizar datos. Permiten manipular grandes tablas de información y realizar cálculos estadísticos de manera rápida y eficiente.
- Matplotlib y Seaborn: herramientas enfocadas en la visualización de datos, muy útiles para crear gráficos descriptivos y exploratorios que facilitan la comprensión de los patrones y tendencias.
- Jupyter Notebook: entorno interactivo que integra código, visualizaciones y explicaciones teóricas en un solo documento, lo que favorece un flujo de trabajo más claro y colaborativo.
- Plataformas en la nube como Google Drive o Google colab, que permiten almacenar, compartir y gestionar los datos de forma práctica y segura.

2. Datos Disponibles

El principal recurso de este proyecto es el dataset de indicadores de salud y diabetes obtenido de Kaggle. Este conjunto de datos reúne miles de registros de personas con distintas características demográficas, conductuales y clínicas. Algunos de los campos más relevantes son:

- Diabetes_binary: variable objetivo que indica si la persona tiene o no diabetes (1 = Sí, 0 = No).
- HighBP: señala la presencia de hipertensión arterial.
- HighChol: refleja niveles altos de colesterol.
- CholCheck: indica si la persona se ha realizado un chequeo de colesterol recientemente.
- BMI: índice de masa corporal, indicador clave para evaluar sobrepeso u obesidad.
- Smoker: muestra si el individuo fuma o ha fumado alguna vez.
- Stroke: historial de accidentes cerebrovasculares.
- HeartDiseaseorAttack: identifica si la persona ha padecido alguna enfermedad cardiovascular.
- PhysActivity: frecuencia de actividad física.
- Fruits y Veggies: variables que describen los hábitos de consumo de frutas y verduras.
- Age y GenHealth: datos demográficos y de percepción de salud general.

Hipótesis iniciales

Hipótesis 1: Un índice de masa corporal (BMI) alto está muy relacionado con la probabilidad de tener diabetes.

El sobrepeso y la obesidad han sido señalados por varias investigaciones como los factores más importantes en el desarrollo de la diabetes tipo 2. El índice de masa corporal (BMI) es una medida común para clasificar a las personas en categorías de peso saludable, sobrepeso u obesidad. Un BMI alto indica que hay más grasa en el cuerpo, lo que suele estar vinculado con un estado inflamatorio constante y con la resistencia a la insulina, una causa principal de la diabetes.

Esta hipótesis sugiere que, en el conjunto de datos, las personas con BMI alto tendrán una proporción mucho mayor de casos de diabetes en comparación con quienes tienen un peso normal. Validar esto ayudará a confirmar el papel del BMI como una variable que predice esta enfermedad y también apoyará la idea de que la obesidad no es solo un factor de riesgo, sino un elemento clave en la causa de la enfermedad.

Hipótesis 2: Tener presión y colesterol altos aumenta mucho el riesgo de desarrollar diabetes.

Diversos estudios han encontrado que la diabetes a menudo aparece junto con otras enfermedades crónicas como la hipertensión y la dislipidemia. Estas condiciones forman

parte del síndrome metabólico, un conjunto de alteraciones que aumentan mucho las probabilidades de diabetes y problemas del corazón.

La hipertensión indica que hay un problema en la circulación y suele venir acompañada de resistencia a la insulina, mientras que el colesterol alto muestra que hay un problema en cómo el cuerpo procesa las grasas. Esta hipótesis plantea que las personas con hipertensión y/o colesterol alto tienen un riesgo mayor de desarrollar diabetes, incluso considerando otros factores como la edad o el peso.

Comprobar esto ayudará a determinar si estas condiciones actúan por separado o si juntas aumentan más el riesgo, lo cual sería importante para crear programas de detección temprana para pacientes con hipertensión o colesterol alto.

Hipótesis 3: Tener hábitos de vida poco saludables, como fumar, beber mucho alcohol y no hacer ejercicio, está vinculado con más casos de diabetes.

Esta hipótesis se centra en los hábitos que pueden modificarse y que, por esto, son claves para prevenir la enfermedad. Fumar afecta cómo el cuerpo procesa la glucosa y puede aumentar la resistencia a la insulina. Beber demasiado alcohol puede dañar el hígado y alterar la regulación de la glucosa. Además, la falta de ejercicio reduce la sensibilidad a la insulina y contribuye al aumento de peso, lo que también eleva el riesgo de diabetes.

Se supone que las personas que fuman y beben en exceso o no hacen actividad física tendrán más probabilidades de tener diabetes que quienes llevan un estilo de vida activo y saludable. Confirmar esto sería útil no solo para entender mejor el tema, sino también para diseñar campañas de salud pública que fomenten hábitos positivos y ayuden a prevenir la enfermedad.

Stakeholders clave

1. Pacientes y población en riesgo

Este grupo es el centro del proyecto, ya que son quienes cuidan su salud. Los resultados del análisis pueden ayudarles a entender sus factores de riesgo, cambiar hábitos y, en algunos casos, acceder temprano a programas de detección y tratamiento. La información les brinda la oportunidad de mejorar su calidad de vida, prevenir complicaciones y disminuir los efectos de la enfermedad.

2. Profesionales de la salud

Incluye médicos generales, endocrinólogos, nutriólogos, psicólogos y especialistas en medicina preventiva. Ellos usan la evidencia del análisis para mejorar sus diagnósticos, ajustar tratamientos y ofrecer recomendaciones personalizadas según los factores de riesgo más importantes. Con datos precisos, su trabajo para prevenir y tratar se vuelve más efectivo.

3. Instituciones de salud y sistemas hospitalarios

Los hospitales, clínicas y centros de atención primaria necesitan información clara para usar mejor los recursos, crear programas de prevención y planear campañas de educación en la comunidad. Un análisis confiable ayuda a priorizar a las personas en riesgo, prever la demanda de servicios y reducir gastos relacionados con complicaciones de la diabetes.

4. Autoridades de salud pública y gobiernos

A nivel general, las secretarías de salud y organismos gubernamentales diseñan políticas públicas y estrategias nacionales para prevenir y controlar la diabetes. Los resultados del proyecto brindan datos importantes para que tomen decisiones en asignación de presupuestos, campañas de sensibilización y regulación de factores de riesgo como el consumo de comida ultra procesada.

5. Investigadores y comunidad académica

Los académicos, epidemiólogos y científicos de datos usan los resultados del estudio para profundizar en nuevas investigaciones, comprobar hipótesis en diferentes contextos y mejorar los modelos predictivos. También ayudan a crear nuevas ideas y a difundir conocimientos en la sociedad.

6. Sector privado y aseguradoras de salud

Las empresas farmacéuticas, aseguradoras y proveedores de servicios médicos también forman parte de los grupos interesados. Ellos pueden usar los resultados para crear productos y servicios adaptados a la población, establecer primas de seguros basadas en riesgos y, en algunos casos, promover la prevención con programas de bienestar en las empresas.

Preguntas clave

1. ¿Cuál es la frecuencia de la diabetes en la muestra analizada?

   Conocer el porcentaje de personas diagnosticadas en el conjunto de datos ayudará a entender la magnitud del problema y a establecer una referencia inicial para el análisis.

2. ¿Qué características demográficas (edad, género, nivel socioeconómico) se relacionan con una mayor probabilidad de tener diabetes?

   La diabetes puede variar entre diferentes grupos de población, por eso es importante identificar si hay segmentos con más vulnerabilidad.

3. ¿Cómo se relaciona el Índice de Masa Corporal (IMC) con la presencia de diabetes?

El sobrepeso y la obesidad son riesgos conocidos. Esta pregunta servirá para ver si los datos muestran esa relación en la muestra.

4. ¿El nivel de actividad física afecta significativamente la probabilidad de tener diabetes?

   Ver si hacer ejercicio regularmente ayuda a prevenir la diabetes ayudará a confirmar la importancia de la actividad física.

5. ¿Qué efecto tienen los hábitos alimenticios, como consumir frutas y verduras, en la probabilidad de tener diabetes?

   La alimentación es uno de los factores que se pueden cambiar fácilmente. Esta pregunta mostrará cuánto influyen estos hábitos en la muestra.

6. ¿Existen otras condiciones de salud, como hipertensión o colesterol alto, que aumenten el riesgo de diabetes?

   Estudiar estas condiciones ayuda a entender cómo se relacionan con la diabetes y si empeoran la condición de los pacientes.

7. ¿La percepción general de la salud (GenHealth) se relaciona con tener diabetes?

   Explorar esta relación puede mostrar si la opinión que las personas tienen de su salud puede ser un indicador temprano de la enfermedad.

8. ¿Qué papel juega el fumar en el desarrollo de la diabetes en esta población?

   Aunque fumar se asocia más con problemas del corazón, también se busca entender si influye en la diabetes de forma indirecta.

9. ¿Qué grupos de la población están en mayor riesgo y deberían ser prioridad en campañas de prevención?

   Detectar estos grupos ayudará a hacer campañas y políticas más efectivas.

10. ¿Se puede crear un modelo confiable que prediga si una persona tiene diabetes según sus indicadores de salud?

    Esta pregunta resume el objetivo final: ver si los datos permiten hacer un sistema que ayude en decisiones médicas o en salud pública.

Fuentes de Datos Identificadas

1. Conjunto de Datos sobre Indicadores de Salud y Diabetes (Kaggle)

La principal fuente de datos es el Diabetes Health Indicators Dataset, que está disponible públicamente en Kaggle. Este conjunto de datos tiene 100,000 registros de personas y tiene variables relacionadas con aspectos demográficos, conductuales y clínicos. El dataset

es estructurado y está en formato CSV, lo que facilita su uso en herramientas de análisis como Python, R o SQL.

Dentro de este dataset, las variables más importantes incluyen:

- Diabetes_binary: variable destino, indica si la persona tiene o no diabetes.
- HighBP y HighChol: muestran si hay hipertensión o colesterol alto.
- BMI: índice de masa corporal, importante para detectar sobrepeso u obesidad.
- Smoker, PhysActivity, Fruits, Veggies: variables que muestran hábitos de vida.
- Age y GenHealth: factores demográficos y percepción de salud.
- HeartDiseaseorAttack y Stroke: historial de enfermedades cardiovasculares.

Este conjunto de datos es la base principal del análisis y permite tanto explorar como crear modelos predictivos.

2. Registros de Salud Pública (fuente secundaria de referencia)

Aunque no están en el dataset descargado, se tomarán como referencia informes oficiales de instituciones de salud pública como la Organización Mundial de la Salud (OMS). Estos informes ofrecen estadísticas confiables y actualizadas.

Justificación del proyecto

La diabetes mellitus es una de las enfermedades crónicas más importantes en el mundo por su alta prevalencia, aumento en casos y alto costo social y económico. Según la Organización Mundial de la Salud (OMS), el número de personas con esta enfermedad se ha cuadruplicado en las últimas décadas y se espera que siga creciendo, especialmente en países en desarrollo. Esto genera una gran carga para los sistemas de salud, ya que implica no solo atender a pacientes con diabetes, sino también tratar complicaciones como insuficiencia renal, ceguera, amputaciones y problemas cardiovasculares.

En este contexto, el proyecto de analizar los indicadores de salud relacionados con la diabetes es muy importante, ya que busca usar los datos masivos para encontrar factores de riesgo, patrones ocultos y relaciones entre variables que afectan la aparición de la enfermedad. Así, no solo se trata de describir el problema, sino de dar información útil que ayude en la prevención y en la toma de decisiones tanto en la clínica como en salud pública.

Desde el punto de vista de los pacientes y la sociedad, el proyecto se justifica porque permite identificar tempranamente a los grupos más vulnerables, facilitando medidas preventivas personalizadas y efectivas. Además, al analizar variables como la obesidad, fumar, los hábitos alimenticios o la actividad física, se pueden hacer recomendaciones específicas para fortalecer programas de salud y reducir la incidencia en el futuro.

Por otra parte, para las instituciones de salud, contar con un modelo predictivo a partir de datos reales ayuda a mejorar el uso de los recursos, priorizar a quienes tienen mayor riesgo

y diseñar mejores políticas públicas. También, para investigadores y científicos, estos proyectos representan una oportunidad para probar hipótesis y crear nuevos métodos para estudiar enfermedades crónicas no transmisibles.

¿Cuantos datos y de que tipo son?

Se cuenta con 100,000 datos de la base de kaggle con 31 columnas de información de los cuales 7 son cadenas de texto, 8 son tipo float y lod restantes 16 son datos numéricos enteros.