## Prevención de la diabetes

Semana 5

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Grupo 22**

**Integrantes:**

Angie Camila Valdes Fuentes

Diana Camila Cordoba Arenas

Robert Kenzo Medina Monsalve

Julian David Hernandez Correcha

## Despliegue de soluciones analíticas

UNIVERSIDAD DE LOS ANDES

2024

Prevención de diabetes 

**Contexto del problema:**

La diabetes es una enfermedad crónica de alta prevalencia que representa una carga significativa para los sistemas de salud y las economías a nivel global. La falta de detección y manejo temprano conduce a complicaciones graves y reduce la calidad de vida de los afectados. Actualmente, el desarrollo de modelos predictivos permite identificar de forma anticipada a individuos con alto riesgo de desarrollar diabetes, lo cual es esencial para implementar intervenciones preventivas efectivas. En un contexto organizacional, estas intervenciones pueden enfocarse en programas de educación en nutrición, promoción de actividad física y seguimiento médico personalizado, reduciendo la incidencia de la diabetes y sus complicaciones a largo plazo.

**Pregunta de negocio y alcance del proyecto:**

Este proyecto tiene como objetivo desarrollar un modelo predictivo que utilice datos de salud, hábitos de vida y factores socioeconómicos para anticipar el riesgo de diabetes en una población determinada. La pregunta de negocio central es: *¿Cómo pueden los modelos predictivos basados en indicadores de salud, hábitos de vida y factores socioeconómicos contribuir a la prevención de la diabetes mediante la predicción precisa del riesgo de la enfermedad?* A través de este modelo, las organizaciones de salud podrán tomar decisiones informadas, asignar recursos de manera eficiente y dirigir programas de intervención personalizados.

**Descripción del conjunto de datos a emplear:**

La base de datos incluye una variable objetivo y varios factores de riesgo asociados. A continuación, se presentan los hallazgos clave:

* Variable objetivo:
  + Diabetes\_binary: Diagnóstico de diabetes (1 = sí, 0 = no). De la muestra total, 35,346 personas padecen diabetes, lo cual representa un 13% de los participantes, mientras que el 87% no presenta diabetes. Este desbalance en los datos sugiere que sería adecuado aplicar técnicas de balanceo en un modelo predictivo para mejorar la precisión en la predicción de casos de diabetes.
* Variables binarias: Las variables como presión arterial alta (HighBP), colesterol alto (HighChol), y consumo de frutas y vegetales (Fruits, Veggies) indican la presencia (1) o ausencia (0) de estas condiciones. Una mayoría de los participantes reporta acceso a atención médica (AnyHealthcare) y actividad física (PhysActivity), mientras que una minoría reporta consumo elevado de alcohol (HvyAlcoholConsump) o haber tenido un derrame cerebral (Stroke).
* Índice de Masa Corporal (BMI): Con una media de 28.38, sugiere una población con tendencia al sobrepeso. La distribución está sesgada hacia la derecha, indicando que una parte significativa tiene sobrepeso u obesidad.
* Salud General (GenHlth): Valorada de 1 a 5, con una media de 2.51, sugiere una percepción de salud general relativamente buena.
* Días de Mala Salud (MentHlth, PhysHlth): La mayoría reporta cero días de mala salud mental y física en el último mes, aunque una minoría presenta una carga significativa de enfermedad.
* Variables Demográficas: Las distribuciones sugieren una variada composición de edad, educación e ingresos. La edad media y avanzada predomina, y niveles educativos más altos se asocian a mayores ingresos.

Correlaciones destacadas con Diabetes\_binary:

* GenHlth (0.29) y HighBP (0.26) tienen la mayor correlación positiva con el diagnóstico de diabetes, sugiriendo que una peor salud general y la presencia de hipertensión aumentan la probabilidad de diabetes.
* DiffWalk (0.22) y BMI (0.22) también están asociados positivamente, indicando que las dificultades para caminar y un mayor índice de masa corporal están vinculados a un mayor riesgo de diabetes.
* HighChol (0.20) muestra una correlación menor pero significativa, reflejando el riesgo cardiovascular.

Correlaciones entre variables de salud y demográficas:

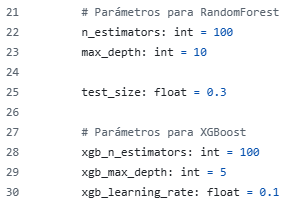
* GenHlth y DiffWalk (0.46) y PhysHlth y GenHlth (0.52), resaltando que una peor salud física y problemas de movilidad están asociados a una mala salud general.
* Income y Education (0.45) sugieren que mayores niveles educativos están asociados con mayores ingresos.

Agrupación de Variables: Para facilitar el análisis, se agruparon las categorías de edad e ingreso en rangos más amplios, mejorando la interpretación en futuros modelos.

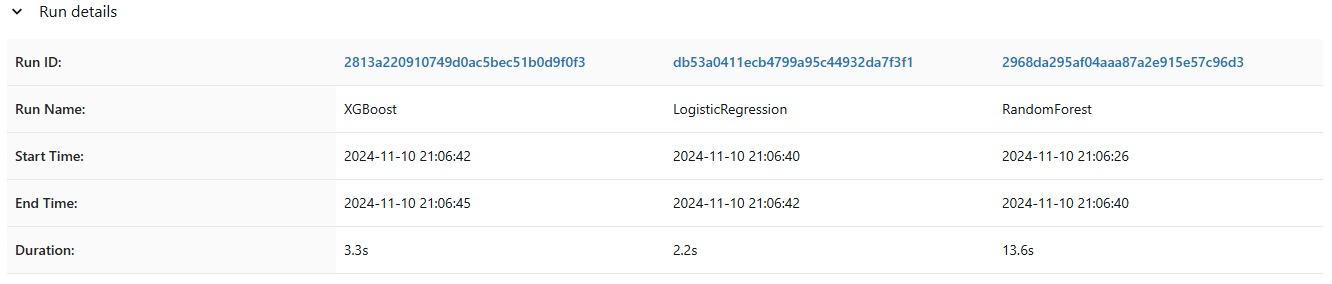
**Modelos desarrollados y su evaluación:**

Se evaluaron tres modelos de clasificación supervisada (XGBoost, Regresión Logística y RandomForest) aplicados a un conjunto de datos cuya división para entrenamiento y prueba fue del 70% y 30%, respectivamente. El objetivo es identificar el modelo con mejor rendimiento basado en métricas de precisión, recall, área bajo la curva ROC (AUC) y tiempo de ejecución.

A continuación, se detallan los principales hiperparámetros ajustados para cada modelo y como se realizó el porcentaje de base que se usó para test:

****

También se midieron los tiempos de inicio y finalización para cada modelo, así como su duración total de ejecución:



Las métricas de evaluación utilizadas para comparar los modelos fueron precisión, AUC, precisión y recall. Los resultados se presentan en la siguiente imagen:

**Logistic Regression** fue el modelo más rápido, completando su ejecución en 2.2 segundos, seguido por **XGBoost** con 3.3 segundos.

**RandomForest** tuvo el tiempo de ejecución más largo, tardando 13.6 segundos, lo cual podría influir en su viabilidad en casos donde el tiempo de respuesta es crítico.

Para comparar y evaluar la calidad de los modelos se usaron las métricas precisión, AUC, precisión y recall. Los resultados se presentan en la siguiente tabla:



**Exactitud (Accuracy)**: XGBoost presenta el mayor valor de exactitud con 0.868, seguido de Logistic Regression y RandomForest con 0.866, lo que indica un rendimiento muy similar entre los tres modelos en términos de clasificación global.

**Área Bajo la Curva (AUC)**: XGBoost también se destaca en esta métrica con un valor de 0.832, lo cual sugiere una buena capacidad para distinguir entre las clases. RandomForest y Logistic Regression tienen valores de 0.828 y 0.825, respectivamente.

**Precisión (Precision)**: RandomForest tiene el mejor valor de precisión (0.605), seguido por XGBoost (0.573) y Logistic Regression (0.545). Esto indica que RandomForest tiene menos falsos positivos, pero no necesariamente es el mejor modelo en general.

**Recall**: Logistic Regression tiene un recall levemente mayor (0.161) en comparación con XGBoost (0.157) y RandomForest (0.085), lo que indica una mayor capacidad para capturar correctamente las instancias positivas.

**Observaciones y conclusiones sobre los modelos**

**Logistic Regression** podría ser una opción viable cuando se requiere un tiempo de ejecución muy bajo y el recall es un factor importante, aunque a expensas de una precisión ligeramente menor.

**RandomForest** mostró una alta precisión, lo cual es útil en casos donde es importante reducir los falsos positivos, pero su bajo recall y mayor tiempo de ejecución lo hacen menos atractivo en comparación con los otros modelos.

**XGBoost** es el **modelo escogido** debido a su buen balance entre exactitud, AUC y tiempo de ejecución. Su alto valor de AUC sugiere una buena capacidad para distinguir entre las clases, y su tiempo de ejecución es razonable en comparación con RandomForest. Sin embargo también se identificó que se puede realizar un modelo que balancee un poco la población y corre nuevamente en este caso un Xgboost y comparar resultados.

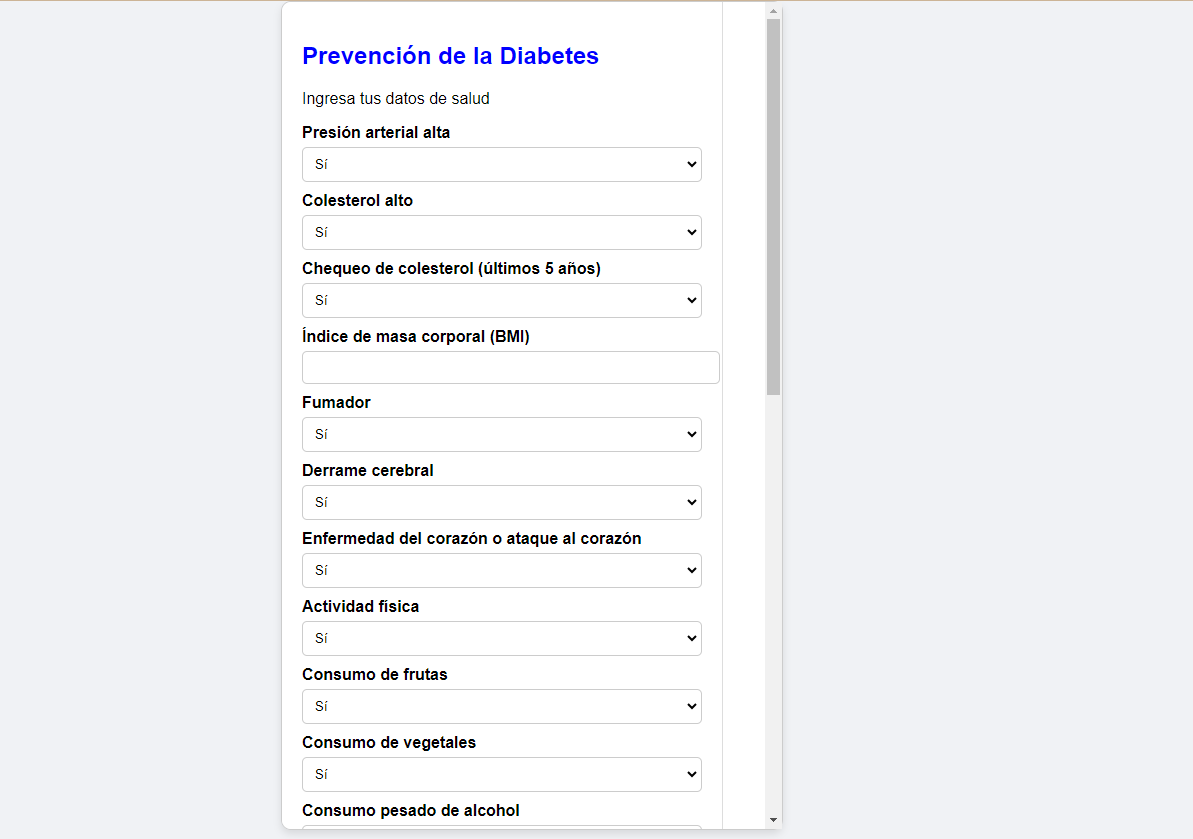
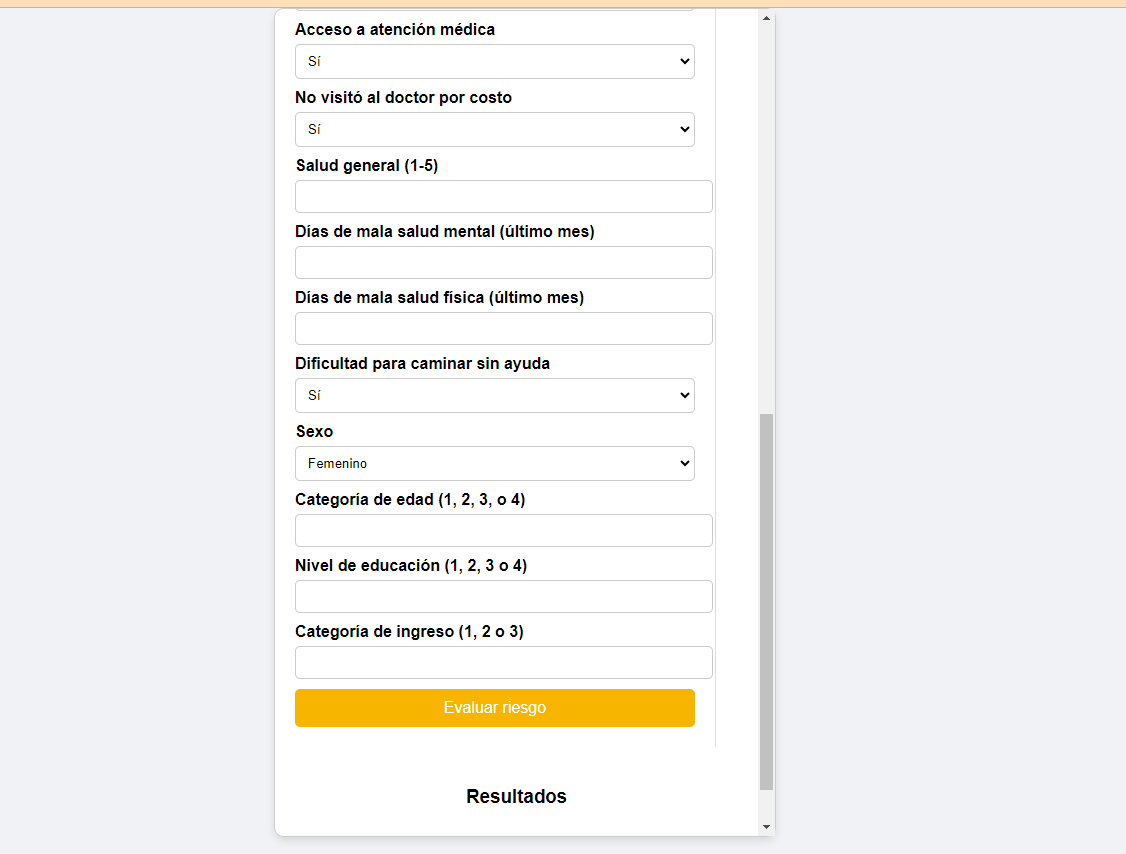
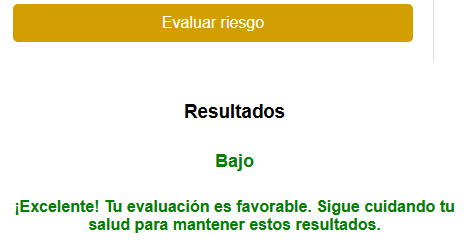
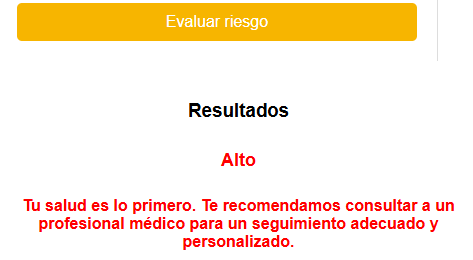
**Descripción del tablero desarrollado y la funcionalidad que éste ofrece**

El Tablero de Prevención de la Diabetes es una herramienta interactiva diseñada para evaluar el riesgo de desarrollar diabetes en función de varios factores de salud y estilo de vida. El tablero permite a los usuarios ingresar datos específicos sobre su estado de salud actual y sus hábitos, con el objetivo de proporcionarles una estimación del riesgo que tienen de desarrollar diabetes.

Como se observa en las imágenes el tablero incluye una sección en la que el usuario puede ingresar sus datos personales y de salud, cada entrada tiene un menú desplegable o un campo numérico adecuado para capturar la información de manera fácil y estructurada. Una vez ingresados todos los datos, el usuario puede hacer clic en el botón "Evaluar riesgo". Esto activa una función que evalúa el riesgo de padecer diabetes según los datos ingresados a través de una API que emplea el modelo seleccionado.

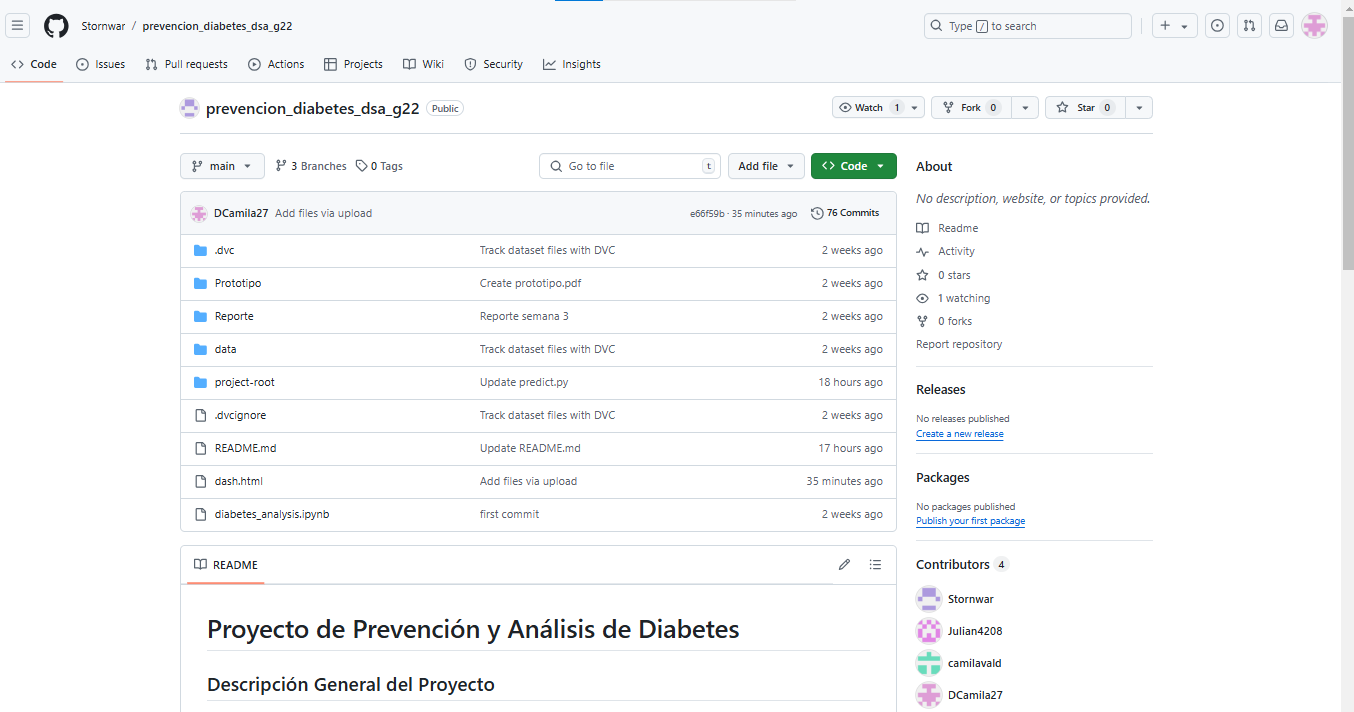
La sección de resultados muestra al usuario su nivel de riesgo, clasificado como **"Alto"** o **"Bajo"**:

* Si el riesgo es **alto**, se muestra un mensaje de advertencia en color rojo que recomienda al usuario consultar a un profesional médico.
* Si el riesgo es **bajo**, el tablero muestra un mensaje positivo en color verde, alentando al usuario a mantener sus buenos hábitos de salud.

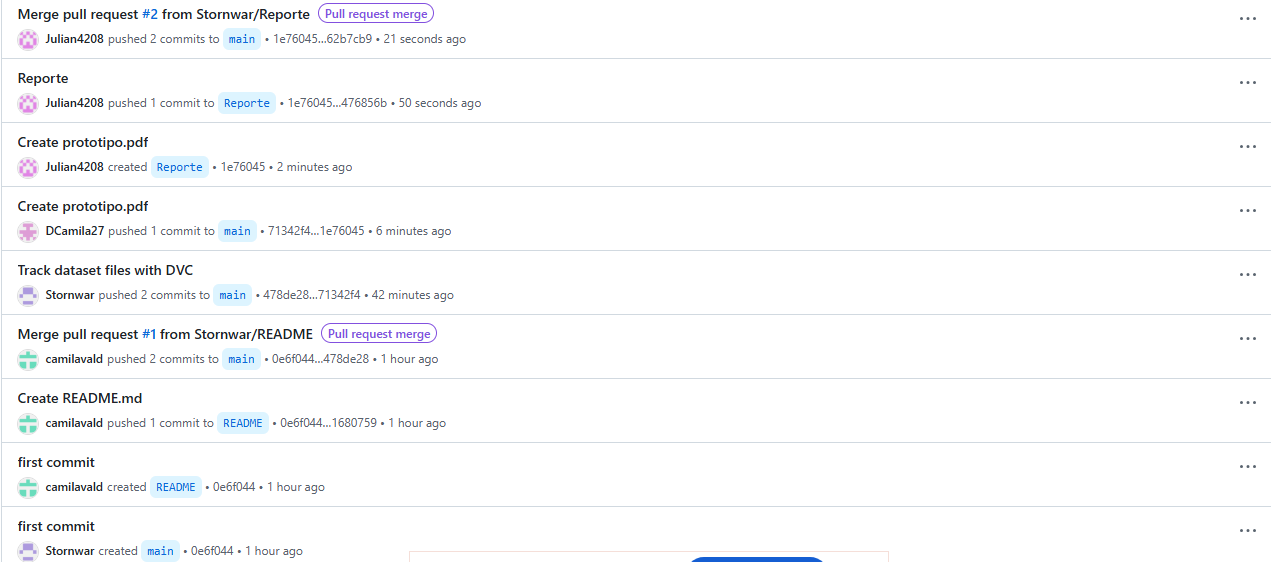
**Repositorio Git**

**Link:** <https://github.com/Stornwar/prevencion_diabetes_dsa_g22.git>



## 

**Reporte de trabajo en equipo**



### Robert Medina

1. **Configuración del Repositorio:** Organicé el repositorio del proyecto, definiendo una estructura ordenada y configurando el entorno de desarrollo para facilitar el trabajo en equipo.
2. **Ajustes en la Instancia:** Implementé y optimicé la instancia en la nube, asegurando que todas las dependencias necesarias estuvieran correctamente instaladas.
3. **Ejecución del Modelo:** Ejecuté el modelo de predicción de diabetes, realizando ajustes de hiperparámetros y validaciones para mejorar su rendimiento.
4. **Documentación:** Documenté el proceso y el uso de scripts, asegurando que el equipo tenga una guía clara para futuras referencias y reproducibilidad del proyecto.
5. Creación del tablero en HTML