



OPERACIONES DE APRENDIZAJE AUTOMÁTICO

Fase 1 | Avance de Proyecto

Early stage diabetes risk prediction dataset

Profesor: Dr. Gerardo Rodríguez Hernández

Profesor Tutor: Mtro. Francisco Javier López Tiro

Equipo #01 | Integrantes del Equipo:

Daniel Acevedo Sainos - **A01795496**

Luis Alejandro Aguilar Diaz - **A01795362**

Marcos Eduardo García Ortiz - **A01276213**

Héctor Raúl Peraza Alavez - **A01795125**

Juan Manuel Rodríguez Mateos - **A01794890**

Andrea Monserrat Ruiz Gómez - **A01794631**

Índice

1. Análisis introductorio del problema en la actividad	3
2. Objetivo.....	3
3. Evaluación del problema	3
4. Roles o distribución de responsabilidades.....	3
5. ML Canvas	4
6. Técnicas y métodos aplicados	5
7. Resultados obtenidos.....	6
8. Conclusiones generales.....	7
9. Bibliografía.	7
10. Anexos.....	7

1. Análisis introductorio del problema en la actividad

El objetivo principal de este proyecto es desarrollar un modelo de Machine Learning capaz de predecir la presencia de diabetes en pacientes a partir de un conjunto de datos que contiene características clínicas. Para lograr este objetivo, se aplicarán técnicas de análisis exploratorio de datos (EDA), preprocesamiento de datos, selección de modelos, entrenamiento y evaluación, así como mejores prácticas de MLOps para asegurar la reproducibilidad, el despliegue eficiente y la monitorización del modelo.

2. Objetivo

A continuación, se detallan los objetivos específicos que guiarán el desarrollo de este proyecto:

- Desarrollar un modelo de Machine Learning con alta precisión para la predicción de diabetes.
- Identificar las variables clínicas que son más relevantes para la predicción de la diabetes.
- Automatizar el pipeline de Machine Learning utilizando MLOps.
- Implementar un sistema de monitorización del modelo para asegurar su rendimiento a lo largo del tiempo.

3. Evaluación del problema

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. La detección temprana y el manejo adecuado de la diabetes son cruciales para prevenir complicaciones graves. Este proyecto busca contribuir a la lucha contra la diabetes mediante el desarrollo de un modelo predictivo que pueda ayudar a los profesionales de la salud a identificar a los pacientes con mayor riesgo de desarrollar la enfermedad. El análisis de datos se centrará en la identificación de patrones y relaciones entre las variables clínicas y la presencia de diabetes.

4. Roles o distribución de responsabilidades

La correcta asignación de roles y responsabilidades es fundamental para asegurar el éxito, permite optimizar los recursos, mejorar la colaboración entre los miembros del equipo y

garantizar que todas las tareas sean completadas de manera eficiente. En esta sección, se detalla cómo se han organizado las responsabilidades, especificando los roles clave y las funciones asignadas a cada integrante, con el fin de asegurar una ejecución alineada con los objetivos establecidos.











- **Stakeholder:** Daniel Acevedo Sainos
 - Define los objetivos y necesidades del proyecto.
 - Proporciona retroalimentación y valida los resultados.
- **Ingeniero de Datos:** Andrea Monserrat Ruiz Gómez
 - Recopila, limpia y prepara los datos para el análisis.
 - Realiza el análisis exploratorio de datos (EDA).
- **Científico de Datos:** Juan Manuel Rodríguez Mateos
 - Selecciona, entrena y evalúa los modelos de Machine Learning.
 - Optimiza el rendimiento de los modelos.
- **Ingeniero de MLOps:** Luis Alejandro Aguilar
 - Implementa el pipeline de CI/CD.
 - Automatiza el entrenamiento y despliegue del modelo.
 - Monitoriza el rendimiento del modelo en producción.
- **Ingeniero de Software:** Marcos Eduardo García Ortiz
 - Desarrolla y mantiene el código del proyecto.
 - Asegura la calidad del código.
- **Project Manager:** Héctor Raúl Peraza Alavez
 - Gestiona el proyecto y coordina las tareas del equipo.
 - Documenta el proceso y los resultados.

5. ML Canvas

A continuación, se proporciona el ML Canvas, para facilitar la planificación y estructuración del proyecto, proporcionando una vista clara y comprensible de los componentes clave involucrados.

THE MACHINE LEARNING CANVAS

Designed for: Equipo 1 Designed by: Equipo 1 Date: 13/10/2024 Iteration: 1

PREDICTION TASK  <p>Tipo de tarea: Clasificación binaria</p> <p>Entidad de predicción: Todo paciente que se le realicen las pruebas para evaluar las diferentes condiciones médicas que evalúa el modelo.</p> <p>Posibles resultados: Detección temprana de la enfermedad diabetes con los resultados "Positivo" y "Negativo"</p> <p>Tiempo de espera antes de la observación: Los datos se recopilan por paciente se hará un análisis al momento de tomar los datos y se debería volver a realizar al momento de que se actualice algún parámetro.</p>	DECISIONS  <p>Valor de la predicción: Las predicciones generadas por el modelo ayudan a los médicos a identificar pacientes en riesgo de padecer diabetes en etapa temprana, para así tomar decisiones preventivas de manera oportuna</p> <p>Aplicación: El sistema puede integrarse como un componente extra en el sistema de registro de pacientes, tener vinculados a los médicos para recibir alertas sobre los casos positivos para poder realizar los estudios necesarios en la próxima revisión médica del paciente, de igual forma notificar al paciente para que acuda a su revisión médica debido a la detección de esta condición.</p>	VALUE PROPOSITION  <p>Usuario final</p> <p>Médicos generales, profesionales de la salud, nutriólogos: Apoyo en la toma de decisiones clínicas: El modelo proporciona a los médicos y especialistas una herramienta adicional e innovadora para la evaluación de los pacientes</p> <p>Pacientes: Brindando acceso a esta herramienta que evalúa el riesgo los pacientes pueden sentirse <u>mas</u> involucrados con su salud motivándolos a adoptar un estilo de vida saludable</p> <p>Objetivo: Identificar de forma rápida a los pacientes en riesgo de desarrollar la enfermedad de diabetes.</p>	DATA COLLECTION  <p>Estrategia para los datos de entrenamiento: El conjunto de datos inicial se basa en el obtenido de UC Irvine, Se recomienda añadir nuevos casos y realizar una actualización periódica</p> <p>Tasa de recolección: Dependerá de la disponibilidad de registros considerar un punto de control trimestral para reunir una cantidad de registros valiosa y considerar entrenar el modelo puramente con los datos experimentales. Monitoreando de manera anual en busca de nuevos patrones de comportamiento de los datos</p> <p>Holdout: Generar un conjunto puro de datos experimentales de nuevos pacientes</p> <p>Costos y restricciones: Seguir las normas medicas de uso de datos</p>	DATA SOURCES  <p>Obtención de más registros</p> <p>Extender y mejorar las variables de entrada realizando la adición de nuevos registros por parte de bases de datos medicas reconocidas para investigación como PUBMED, EMBASE y Cochrane Library</p> <p>APIs que provean información médica</p> <p>Intercambio de registros con hospitales y clínicas manteniendo la confidencialidad</p> <p>Extrapolación de datos clínicos a partir de estudios previos</p>
IMPACT SIMULATION  <p>Despliegue del modelo: Una vez realizado el entrenamiento y la validación el modelo se puede integrar en los sistemas de diagnóstico clínico</p> <p>Conjunto de prueba: Recabar un conjunto de registros de campo donde se tenga la confirmación de casos positivos y negativos</p> <p>Costo/ganancia de decisiones correctas e incorrectas (FP/FN): Al ser un modelo de detección temprana y de control tiene como finalidad avisar de un posible caso para su confirmación se tendrían que realizar otros estudios</p> <p>Restricción de equidad: Garantizar que el modelo no cree sesgos por lo que debe darse un data set de pacientes totalmente aleatorio</p>	MAKING PREDICTIONS  <p>Cuando predecir: Al momento de cargar la información de un nuevo paciente y al detectarse un cambio en alguno de los parámetros</p> <p>Tiempo para caracterización y post procesamiento: Dado que se trata de condiciones fisiológicas o condiciones que tardan en presentarse el tiempo de computación es mínimo y único</p> <p>Objetivo: Predicción binaria sin un paciente tiene riesgo de desarrollar diabetes</p>	<p>Beneficios del sistema de ML: Mejora la capacidad de diagnóstico sin necesidad de realizar pruebas invasivas, Aumentaría notablemente la precisión y rapidez en la toma de decisiones médicas evitando complicaciones de salud a largo plazo</p> <p>Workflow: Ingreso de datos del pacientes -> Preprocesamiento -> Predicción -> Generación de alerta médica -> Envío de alerta a médico y paciente -> Toma de decisiones por parte del médico</p>	BUILDG MODELS  <p>Número de modelos: 1 ó 2 que pueden ser optimizados para clasificación binaria mejorando la precisión</p> <p>Actualización: Actualizar el modelo cuando se tenga un número significativo de datos o se detecte un número muy alto de casos positivos verdaderos</p> <p>Tiempo para análisis: Se puede definir un análisis apegado al ciclo de actualización de datos clínicos</p>	FEATURES  <p>Entradas disponibles:</p> <p>Variables booleanas de síntomas: Poliuria, polidipsia, pérdida de peso repentina, debilidad, polifagia, candidiasis genital, visión borrosa, etc.</p> <p>Atributos adicionales: Edad, género y obesidad</p> <p>Salida esperada:</p> <p>Clase: La salida del modelo una predicción binaria "Diabetes en etapa temprana o no"</p>
MONITORING  <p>Métricas para medir el impacto:</p> <p>Precisión, sensibilidad y especificidad del modelo</p> <p>Impacto clínico: Reducción de diagnósticos tardíos, mejora en los tratamientos preventivos</p> <p>Económico: Costo ahorrado en cuidados del tratamiento</p>				

6. Técnicas y métodos aplicados

En este proyecto se emplearán diversas técnicas y métodos que permiten abordar los objetivos planteados de manera eficiente. La selección de estas estrategias ha sido cuidadosamente realizada para optimizar los resultados y garantizar un enfoque riguroso.

- **Análisis Exploratorio de Datos (EDA):** Se utilizarán técnicas de visualización de datos (histogramas, diagramas de dispersión, boxplots) y estadística descriptiva para

comprender las características del dataset y identificar patrones y relaciones entre las variables.

- **Preprocesamiento de Datos:** Se aplicarán técnicas de limpieza de datos (manejo de valores faltantes) y transformación de datos (escalado de variables, codificación de variables categóricas) para preparar los datos para el modelado.
- **Selección de Modelos:** Se evaluarán diferentes modelos de Machine Learning, como la Regresión Logística, Máquinas de Soporte Vectorial (SVM), Árboles de Decisión y Random Forest, para determinar cuál se ajusta mejor a los datos y al objetivo del proyecto.
- **Entrenamiento y Evaluación:** Se entrenarán los modelos utilizando técnicas de validación cruzada y se evaluará su rendimiento mediante métricas como la precisión, recall, F1-score, AUC-ROC.
- **MLOps:** Se implementará un pipeline de CI/CD utilizando herramientas como GitHub Actions para automatizar el entrenamiento, la evaluación y el despliegue del modelo. Se utilizarán técnicas de versionado de datos (DVC) y se monitorizará el rendimiento del modelo en producción.

A medida que el proyecto avanza, estas estrategias seguirán ajustándose para garantizar el éxito en cada etapa.

7. Resultados obtenidos.

Tras implementar las técnicas y métodos seleccionados, se obtuvieron resultados significativos en cuanto a la precisión y eficiencia del modelo de predicción de diabetes. Los principales hallazgos son los siguientes:

Se identificaron valores faltantes o cero. Para manejar estos valores faltantes, se aplicó una estrategia de imputación utilizando la media de cada columna. Posteriormente, se estandarizaron todas las variables numéricas utilizando StandardScaler para asegurar que tuvieran la misma escala y evitar que alguna variable dominara el proceso de modelado.

Para la predicción de la diabetes, se entrenaron tres modelos de Machine Learning: Regresión Logística, Máquinas de Soporte Vectorial (SVM) y Árbol de Decisión. El rendimiento de cada modelo se evaluó utilizando métricas como la precisión, recall, F1-score y AUC-ROC. Se utilizó MLflow para el seguimiento de los experimentos y el registro de las métricas, lo que facilitó la comparación del rendimiento de los diferentes modelos.

Variables más relevantes: A través de técnicas de importancia de características, se identificó que el dataset contiene 16 variables como edad, género etc. y síntomas como poliuria, polidipsia, pérdida de peso repentina, debilidad, polifagia, candidiasis genital y trastornos visuales. Estas variables resultaron ser determinantes para predecir la presencia de diabetes.

8. Conclusiones generales.

En conclusión, el desarrollo de este modelo de predicción de diabetes ha demostrado ser eficaz para identificar a pacientes en riesgo a partir de datos clínicos básicos. La combinación de un enfoque basado en Machine Learning con prácticas robustas de MLOps ha permitido no solo crear un modelo preciso, sino también asegurar su despliegue y mantenimiento a largo plazo.

Además, la identificación de las variables más relevantes puede proporcionar información valiosa a los profesionales de la salud para la toma de decisiones informadas, mejorando la detección temprana de la diabetes y contribuyendo así a la prevención de complicaciones graves.

Este proyecto sienta las bases para futuros desarrollos y mejoras, como la integración de datos adicionales o la optimización de los algoritmos, lo que podría aumentar aún más la precisión y utilidad clínica del modelo.

9. Bibliografía.

UCI Machine Learning Repository. (s. f.).

<https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset>.

10. Anexos.

- Se comparte la url del repositorio de GitHub.

https://github.com/LAguilar35/diabetes_mlops

- Se comparte la url del video.

<https://youtu.be/iJoZs2tGbD0>