



“Desarrollo de algoritmos en modelos con Machine Learning”

Sprint 1

Bernardo Corona Domínguez

NAO ID: 1085

Ciudad de México, México

18 de junio de 2025

## 1. Backlog.

### A) Historias de usuario.

A continuación, las 5 (cinco) historias de usuario:

Historia 1:

Como analista de datos, necesito realizar un análisis exploratorio inicial, para entender la estructura y calidad de los datos disponibles.

Historia 2:

Como especialista en Machine Learning, necesito transformar y limpiar los datos, para garantizar precisión y eficiencia en el modelo predictivo.

Historia 3:

Como científico de datos, requiero desarrollar un modelo inicial de regresión logística, para predecir la incidencia de enfermedades cardiovasculares.

Historia 4:

Como investigador clínico, necesito evaluar y optimizar el modelo mediante técnicas avanzadas (regularización y manejo de datos desbalanceados), para mejorar la precisión de las predicciones.

Historia 5:

Como directora de comunicación social, quiero presentar resultados analíticos claros y visuales, para respaldar campañas preventivas efectivas.

### B) Tablas de registro.

Historia de usuario	Requerimientos específicos
HU-1: Análisis exploratorio inicial	- Identificación de valores faltantes.

	<ul style="list-style-type: none"> <li>- Detección y visualización de datos atípicos.</li> <li>- Análisis estadístico descriptivo básico.</li> <li>- Gráficas exploratorias iniciales (histogramas, boxplots).</li> </ul>
<b>HU-2: Preparación y transformación de datos</b>	<ul style="list-style-type: none"> <li>- Limpieza de datos nulos o erróneos.</li> <li>- Codificación de variables categóricas.</li> <li>- Normalización o estandarización de variables numéricas.</li> <li>- División adecuada de datos para evitar data leaking.</li> </ul>
<b>HU-3: Modelo inicial de regresión logística</b>	<ul style="list-style-type: none"> <li>- Selección y definición de variables predictoras.</li> <li>- Entrenamiento inicial del modelo con regresión logística.</li> <li>- Evaluación con métricas de desempeño iniciales (precisión, recall, F1-score, ROC-AUC).</li> <li>- Reporte preliminar de resultados.</li> </ul>
<b>HU-4: Optimización del modelo mediante técnicas avanzadas</b>	<ul style="list-style-type: none"> <li>- Aplicación de técnicas de regularización (L1, L2).</li> <li>- Manejo avanzado de datos desbalanceados (SMOTE, undersampling, oversampling).</li> <li>- Validación cruzada para evaluación robusta del modelo.</li> <li>- Documentación detallada de las mejoras logradas.</li> </ul>
<b>HU-5: Reporte y presentación de resultados</b>	<ul style="list-style-type: none"> <li>- Generación de gráficos avanzados (ROC, Curvas Precision-Recall).</li> <li>- Creación de reporte analítico detallado en formato PDF.</li> <li>- Desarrollo de videografía explicativa en formato MP4.</li> <li>- Resumen ejecutivo para comunicación clara de hallazgos.</li> </ul>

## Lista Priorizada.

Requerimientos	Etapas	Estimación de Tiempo	Entregables
<b>Análisis exploratorio de datos</b>	Sprint 1	4 días	Reporte de hallazgos con visualizaciones y estadísticas descriptivas
<b>Transformación y limpieza de datos</b>	Sprint 1	3 días	Dataset limpio, normalizado y codificado
<b>División de datos sin generar data leaking</b>	Sprint 1	1 día	Dataset separado en training y test con documentación de proceso
<b>Entrenamiento de modelo inicial (regresión logística)</b>	Sprint 2	3 días	Modelo entrenado, notebook de implementación y métricas iniciales
<b>Evaluación del modelo (precisión, recall, F1, ROC-AUC)</b>	Sprint 2	2 días	Reporte de desempeño del modelo
<b>Aplicación de técnicas de regularización (L1, L2)</b>	Sprint 3	2 días	Versión optimizada del modelo
<b>Manejo del desbalanceo de datos (SMOTE u otra técnica adecuada)</b>	Sprint 3	2 días	Dataset balanceado y validación del impacto en el modelo
<b>Validación cruzada y ajuste fino de hiperparámetros</b>	Sprint 3	2 días	Reporte con resultados de validación cruzada
<b>Elaboración de presentación visual y analítica en PDF</b>	Entrega final	2 días	Archivo PDF con análisis, gráficas, interpretación y recomendaciones
<b>Desarrollo de videografía de presentación (10 minutos aprox.)</b>	Entrega final	2 días	Video MP4 explicativo sobre el proyecto y sus resultados

## Lista de requerimientos

Historias de Usuario	Requerimientos Específicos	Requerimientos Técnicos
<b>HU-1: Análisis exploratorio inicial</b>	Identificar valores atípicos, comprender la distribución de variables y establecer hipótesis iniciales.	Implementación en Python (Pandas, Matplotlib, Seaborn). Scripts para análisis univariado y bivariado.
<b>HU-2: Preparación y transformación de datos</b>	Limpiar datos faltantes o inconsistentes, transformar variables, codificar categorías, escalar numéricos.	Uso de Scikit-learn (StandardScaler, OneHotEncoder, etc.). Documentación del pipeline de preprocesamiento.
<b>HU-3: Modelo inicial de regresión logística</b>	Entrenar un modelo predictivo básico, evaluar métricas como precisión, recall y F1-score.	Implementación de LogisticRegression de Scikit-learn. Uso de métricas con classification_report, confusion_matrix.
<b>HU-4: Optimización del modelo</b>	Mejorar el desempeño usando regularización (L1, L2) y técnicas de rebalanceo de clases (SMOTE, etc.).	Uso de LogisticRegression(penalty='l1/l2'), integración de SMOTE con imblearn. Validación cruzada (GridSearchCV).
<b>HU-5: Reporte y presentación de resultados</b>	Generar informes gráficos y escritos que comuniquen los hallazgos y desempeño del modelo.	Generación de gráficos ROC, PR curve. Exportación de resultados en PDF. Creación de videografía en MP4 (via OBS o Canva).