“Desarrollo de algoritmos en modelos con Machine Learning”

Sprint 3

Bernardo Corona Domínguez

NAO ID: 1085

Ciudad de México, México

9 de julio de 2025

1. Backlog.

El desarrollo de un sistema predictivo en contextos clínicos implica no sólo el despliegue técnico del aprendizaje automático, sino también una comprensión fina de las necesidades que subyacen a cada uno de los roles involucrados. A partir de esta premisa, se delinean a continuación cinco relatos funcionales que, más allá de describir tareas, articulan intenciones, preocupaciones y objetivos desde distintas posiciones profesionales.

Historia 1

Desde la función analítica, el primer paso —ineludible por su valor diagnóstico— consiste en explorar con detenimiento el conjunto de datos disponible. Se trata, en esencia, de abrir la caja negra: identificar vacíos, detectar distorsiones, y comenzar a intuir patrones o irregularidades que podrían condicionar el desempeño futuro del modelo. Esta etapa fundacional permite no sólo mapear el terreno, sino también afinar la mirada crítica frente a los datos.

Historia 2

En la práctica de quien se especializa en *Machine Learning*, la fase de transformación y depuración de los datos adquiere un carácter meticuloso y técnico. No basta con limpiar errores evidentes: se trata de moldear el *dataset* hasta que sea un insumo robusto, libre de ruidos y adecuadamente escalado, donde cada variable represente con fidelidad lo que promete. Es aquí donde se definen, de forma silenciosa pero decisiva, los límites de lo que el modelo podrá (o no) alcanzar.

Historia 3

Para el científico de datos, el momento de modelar implica algo más que programar una función matemática. Al desarrollar una regresión logística con fines predictivos —en este caso, orientada a estimar el riesgo de enfermedades cardiovasculares— se pone en juego una hipótesis de causalidad estadística, una lógica de

inferencia, y una ética del error. El modelo no sólo clasifica: también sugiere decisiones futuras, y por ello su construcción debe cuidarse con rigor.

Historia 4

La mirada del investigador clínico introduce una exigencia adicional: no basta con que el modelo funcione bien en promedio; debe hacerlo también en los casos críticos. Por ello, optimizar su desempeño requiere aplicar técnicas que vayan más allá del ajuste superficial. Regularizar el modelo, abordar el desbalance de clases —tan frecuente en eventos de salud poblacional—, y validar sistemáticamente los resultados, se vuelve imprescindible para dotar de solidez empírica a la predicción.

Historia 5

Finalmente, desde la esfera de la comunicación institucional, surge la necesidad de traducir complejidad en claridad. Es decir, de convertir los hallazgos técnicos en insumos visuales, comprensibles y estratégicamente útiles para campañas de prevención o toma de decisiones públicas. No es menor esta tarea: en ella se juega la posibilidad de que el conocimiento generado tenga impacto más allá del código.

B) Tablas de registro.

Historia de usuario	Requerimientos específicos
HU-1: Exploración inicial	- Detección de vacíos y anomalías estadísticas. - Visualización exploratoria (distribuciones, dispersión, valores extremos). - Análisis univariado y bivariado. - Elaboración de representaciones gráficas (histogramas, diagramas).
HU-2: Preprocesamiento	- Depuración de registros incompletos o inconsistentes. - Codificación categórica (*one-hot* u ordinal según contexto). - Normalización de variables numéricas.

	- Separación de datos de entrenamiento y prueba cuidando evitar fugas de información (*data leakage*).
HU-3: Modelado base	- Definición de variables independientes y dependientes. - Entrenamiento preliminar con regresión logística. - Evaluación de métricas básicas: precisión, *recall*, F1, AUC-ROC. - Síntesis técnica de resultados obtenidos.
HU-4: Optimización	- Aplicación de regularización L1 y L2 para evitar sobreajuste. - Rebalanceo mediante técnicas como SMOTE, submuestreo o sobremuestreo. - Validación cruzada (*k-fold*) para evaluación robusta. - Registro sistemático de ajustes y efectos.
HU-5: Comunicación analítica	- Generación de visualizaciones interpretativas (curvas ROC, *precision-recall*). - Elaboración de informe integral en PDF. - Producción de videografía explicativa con narración del proceso y hallazgos. - Redacción de resumen ejecutivo orientado a públicos no técnicos.

Lista Priorizada.

Requerimiento	Sprint	Duración estimada	Producto esperado
Análisis exploratorio	Sprint 1	4 días	Informe descriptivo con visualizaciones
Transformación y limpieza	Sprint 1	3 días	*Dataset* normalizado y codificado
División de datos sin fugas	Sprint 1	1 día	*Dataset* separado y documentado
Modelado inicial	Sprint 2	3 días	Modelo base, notebook y métricas
Evaluación de desempeño	Sprint 2	2 días	Reporte técnico de rendimiento
Regularización	Sprint 3	2 días	Modelo optimizado con penalizaciones
Rebalanceo de clases	Sprint 3	2 días	*Dataset* balanceado y evaluación del impacto
Validación cruzada	Sprint 3	2 días	Informe de ajuste fino y validación
Presentación PDF final	Entrega final	2 días	Informe analítico completo y visualmente claro
Video explicativo	Entrega final	2 días	Videografía argumentada con duración aproximada de 10 minutos

Síntesis de requerimientos técnicos

Historia de Usuario	Objetivo operativo	Herramientas y métodos técnicos
HU-1	Comprender estructura y calidad del *dataset*	Python con Pandas, *Matplotlib* y *Seaborn* para análisis exploratorio
HU-2	Transformar el conjunto de datos en un insumo modelable	*Scikit-learn* para escalamiento y codificación; documentación del pipeline
HU-3	Construir un modelo inicial para predicción binaria	Uso de *LogisticRegression* y herramientas de evaluación estándar
HU-4	Maximizar la capacidad predictiva del modelo	Aplicación de regularización y SMOTE; validación con *GridSearchCV*
HU-5	Comunicar hallazgos de forma efectiva y visual	Curvas ROC/PR, PDF exportable, y videografía en Camtasia