

**Predicción de Riesgos financieros y
Estimaciones Presupuestarias en
Construcción inmobiliaria con
Modelos de Aprendizaje automático**

Autor:

Ing. Danilo Simón Reitano Andrades

Director:

A definir (A definir)

Índice

1. Descripción técnica-conceptual del proyecto a realizar	5
2. Identificación y análisis de los interesados	8
3. Propósito del proyecto	9
4. Alcance del proyecto	9
5. Supuestos del proyecto	10
6. Product Backlog	10
7. Criterios de aceptación de historias de usuario	12
8. Fases de CRISP-DM	14
9. Desglose del trabajo en tareas	16
10. Diagrama de Gantt	23
11. Planificación de Sprints	23
12. Normativa y cumplimiento de datos (gobernanza)	25
13. Gestión de riesgos	25
14. Sprint Review	27
15. Sprint Retrospective	27

Registros de cambios

Revisión	Detalles de los cambios realizados	Fecha
0	Creación del documento	29 de abril de 2025
1	Se completa hasta el punto 5 inclusive	12 de mayo de 2025
2	Se completa hasta el punto 9 inclusive	19 de mayo de 2025

Acta de constitución del proyecto

Mendoza, 29 de abril de 2025

Por medio de la presente se acuerda con el Ing. Danilo Simón Reitano Andrades que su Trabajo Final de la Carrera de Especialización en Inteligencia Artificial se titulará “Predicción de Riesgos financieros y Estimaciones Presupuestarias en Construcción Inmobiliaria con Modelos de Aprendizaje automático” y consistirá en el desarrollo un modelo de predicción de préstamos y presupuestos de construcción. El trabajo tendrá un presupuesto preliminar estimado de 600 horas y un costo estimado de \$ 1500, con fecha de inicio el 29 de abril de 2025 y fecha de presentación pública al definir noviembre de 2024.

Se adjunta a esta acta la planificación inicial.

Dr. Ing. Ariel Lutenberg
Director posgrado FIUBA

Martin Brambati
Built Technologies

A definir
Director del Trabajo Final

1. Descripción técnica-conceptual del proyecto a realizar

Introducción, contexto y propuesta de solución

El presente trabajo surge a partir de una necesidad concreta de Built Technologies en la empresa con sede en Nashville, Tennessee, Estados Unidos. Dicha necesidad de responder preguntas sobre los presupuestos de sus clientes fue identificada por el autor del documento, Danilo Reitano. Dicha empresa desarrolla una plataforma SaaS llamada gestión de préstamos para la construcción. Trabaja con múltiples entidades bancarias y financieras que organizan financiamiento a personas que planean construir sus casas y desarrolladores y contratistas para la ejecución de proyectos residenciales y comerciales. La plataforma permite monitorear el uso del préstamo, la ejecución del presupuesto y los avances del proyecto de forma integrada.

Uno de los principales desafíos operativos que enfrenta la empresa, y en general el sector de préstamos para construcción, es la dificultad para validar las etapas tempranas si el importe del préstamo solicitado por el cliente será suficiente para cubrir los costos del proyecto. Actualmente, esta validación se realiza mediante revisión manual de los presupuestos enviados, que suelen estar desagregados en múltiples partidas (ej.: cimentación, materiales, permisos, mano de obra, pisos, techos, terminaciones, etc.). Este proceso es altamente dependiente del criterio del analista y de sus experiencias pasadas, sin duda herramientas sistemáticas que permitan predecir desvíos, subestimaciones o riesgos de insuficiencia presupuestaria. A menudo, estas deficiencias, recién se manifiestan cuando el proyecto ya está en ejecución, mientras se generan sobrecostos, atrasos, renegociaciones contractuales o incluso abandono de obra.

En este contexto, la propuesta de esta tesis es el desarrollo de un sistema predictivo basado en modelos de aprendizaje automático que permita asistir de manera inteligente a los analistas de crédito en dos aspectos clave:

1. Estimar, a partir de la información disponible al momento de analizar el préstamo, si el importe solicitado será suficiente para cubrir el presupuesto completo del proyecto.
2. Predecir los costos esperados para cada una de las partidas presupuestarias del proyecto (por ejemplo: excavación, estructura, instalaciones eléctricas, techo, decorados). Se deberá tener en cuenta características del proyecto como su ubicación, tipo, proveedor, constructor, superficie y otros factores históricos.

Este doble enfoque con distintas capas de análisis permitirá no solo detectar situaciones de riesgo financiero de manera temprana, sino también ofrecer recomendaciones concretas sobre los costos esperados. Todo esto en función de los datos históricos, ajustados al contexto de cada proyecto.

Contexto y condiciones particulares del proyecto

Contexto y condiciones particulares del proyecto

Este proyecto se desarrollará en estrecha colaboración con el equipo de datos de Built Technologies. Se cuenta con acceso autorizado a un dataset anonimizado que incluye información detallada de más de 10 años de historia de proyectos, con datos por rubro presupuestario, tipo de préstamo, ubicación, resultados de ejecución y características del contratista y del proveedor. Por cuestiones de privacidad y cumplimiento normativo (SOC 2 y GDPR), los datos no contienen información sensible de clientes y el trabajo se limita a procesar información estructurada, descartando el uso de documentos escaneados o imágenes.

Foto: Contiene información sensible de clientes y elaboración se limita exclusivamente a información estructurada, sin el uso de documentos escaneados o imágenes.

En términos generales, existen diversas aplicaciones de modelos de machine learning en el ámbito financiero, particularmente en la evaluación de crédito, detección de fraude y puntuación de Estado del arte y diferenciación de la solución. Sin embargo, el uso de aprendizaje automático para prever costos de construcción y validar la suficiencia de préstamos basándose en presupuestos históricos es aún una línea de investigación y desarrollo incipiente. En términos generales, existen diversas aplicaciones de modelos de machine learning en el ámbito financiero, particularmente en la evaluación de crédito, detección de fraude y puntuación de Clientes. Sin embargo, el uso de aprendizaje automático para prever costos de construcción y validar la suficiencia de préstamos basándose en presupuestos históricos es aún una línea de investigación y desarrollo incipiente. La mayoría de las soluciones actuales en el sector se apoyan en heurísticas basadas en precios imitarios, bases de datos estáticas por región que benchmarking manual entre proyectos. Esto tiene limitaciones evidentes: no contempla el contexto completo del proyecto ni aprende de los patrones reales de ejecución observados en los últimos años. Además, a medida que los La mayoría de las soluciones actuales en el sector se apoyan en heurísticas basadas en precios imitarios, bases de datos estáticas por región que benchmarking manual entre proyectos. Esto tiene limitaciones evidentes: no contempla el contexto completo del proyecto ni aprende de Los patrones reales de ejecución observados en los últimos años. Además, a medida que los proyectos modernos incrementaron en escala y complejidad, los métodos convencionales resultan insuficientes para capturar todas las variables que afectan los costos. Estas regresiones multivariadas han sido un enfoque básico para estimar costos de construcción. Estos modelos asumen relaciones lineales entre los factores (como tamaño de la obra, calidad de los materiales, etc.) y el costo total. En escenarios relativamente simples, las regresiones pueden brindar estimaciones razonables: típicamente logran una precisión del orden de 75 %-80 %. No obstante, en la mayoría de los proyectos existen relaciones no lineales y dependencias complejas entre variables (economías de escala, influencias del mercado, interacciones entre diseño y método constructivo) que limitan la capacidad predictiva. También se ha explorado la utilización de otros tipos de modelos como Random Forest y XGBoost. Estos enfoques han mostrado mejoras significativas en precisión frente a métodos tradicionales. Por ejemplo, un estudio recopiló datos de 95 proyectos de edificios y implementó un modelo de Random Forest para predecir riesgos de sobrecostos.^[3] Otro ejemplo relevante es También se ha explorado la utilización de otros tipos de modelos como Random Forest y XGBoost. Estos enfoques han mostrado mejoras significativas en precisión frente a métodos tradicionales. Por ejemplo, un estudio recopiló datos de 95 proyectos de edificios y implementó un modelo de Random Forest para predecir riesgos de sobrecostos. Otro ejemplo relevante es un trabajo que utilizó XGBoost para seleccionar las variables más influyentes y estimar el costo de proyectos de edificación. La precisión del 85 %-90 %, superior a la obtenida por modelos más simples. Esto se traduce en menores errores de predicción para proyectos complejos, aunque a costa de una mayor demanda En otros casos, se ha explorado la posibilidad de utilizar modelos de deep learning para estimación de costos en el desarrollo de proyectos. Algunas investigaciones han resultado en precisiones del 85 %-90 %, superior a la obtenida por modelos más simples. Esto se traduce en menores errores de predicción para proyectos complejos, aunque a costa de una mayor demanda de datos y capacidad computacional. Así, los resultados de dichos proyectos, se puede apreciar que las variables de mayor influencia son la escala del proyecto (metros cuadrados construidos, número de plantas), funcionalidad (edificio residencial, comercial, salud, etc.), así como especificaciones técnicas principales (tipo de climatización, sistema estructural, acaudos exteriores e interiores, presencia de elementos especiales como ascensores, etc.), ubicación geográfica, año o época de construcción y tipo de contrato, entre otros. La solución propuesta se diferencia al usar dos tipos de aprendizaje supervisado entrenados sobre datos reales de ejecución y financiamiento, ajustando las predicciones al comportamiento histórico.

La solución propuesta se diferencia en que:

- Integra múltiples variables categóricas y numéricas, como ubicación geográfica, superficie construida, tipo de contratista y composición del presupuesto, ofreciendo una predicción contextualizada.
- Utiliza modelos de aprendizaje supervisado entrenados sobre datos reales de ejecución y financiamiento, con ajuste de las predicciones al comportamiento histórico.
- Utiliza un volumen de datos mayor a los proyectos realizados hasta el momento.

- Integra múltiples variables categóricas y numéricas (Sobre ubicación geográfica, superficie construida, tipo de construcción y tipo de composición del presupuesto), facilitando una predicción contextualizada a los analistas humanos.
- Utiliza un volumen grande de datos mayores que los proyectos ya realizados hasta el momento, regresiva multirrubro (estimación de costos por partida).
- Incorpora técnicas de explicabilidad como SHAP (Shapley Additive Explanations) para interpretar por qué el modelo predice insuficiencia o sobrecosto, con adopción por parte de los analistas humanos.

Propuesta de valor e impacto esperado

- Ofrece tanto una salida binaria (¿es suficiente el préstamo?) como una salida regresiva multirrubro (estimación de costos por partida). La implementación de este sistema generaría múltiples beneficios para Built Technologies y sus socios financieros:

Propuesta de valor e impacto esperado

■ Reducir el porcentaje de proyectos con préstamos insuficientes, disminuyendo renegociaciones y sobrecostos.

La implementación de este sistema generará múltiples beneficios para Built Technologies y sus socios financieros:

- Mejorar la eficiencia del análisis crediticio, apoyando a los analistas con una herramienta predictiva basada en datos.

- Aumentar la satisfacción del cliente final, evitando interrupciones en la ejecución del proyecto por errores de estimación.
- Reducir el porcentaje de proyectos con préstamos insuficientes y así lograr evitar renegociaciones y sobrecostos.

- Detectar patrones de subestimación crónica en ciertos rubros o regiones, lo que podría informar futuras políticas de origenación de préstamos.

El sistema implementará la satisfacción del cliente final, sin interrupciones en la ejecución del proyecto tabulado por errores de estimación, integrado en el flujo de trabajo de análisis crediticio mediante una API o módulo dentro de la plataforma existente.

- Detectar patrones de subestimación crónica en ciertos rubros o regiones, lo que podría informar futuras políticas de origenación de préstamos.

Descripción funcional de la solución

El sistema se implementará inicialmente como un prototipo funcional con salida en formato tabular y visual, para luego ser integrado en el flujo de trabajo de análisis crediticio mediante una API o módulo dentro de la plataforma existente.

- Módulo de extracción de datos: Identificación, scrapping y unificación de datos.

Descripción funcional de la solución: Limpieza, transformación y validación de los datos históricos estructurados.

La solución se compone de los siguientes bloques funcionales:

- **Módulo de entrenamiento:** Entrenamiento de dos modelos: uno de clasificación binaria (suficiencia del préstamo) y otro de regresión multirrubro (predicción de costos por partida).
- **Módulo de extracción de datos:** identificación, scrapping y unificación de datos.
- **Módulo de inferencia:** Dado un nuevo proyecto con sus características, el sistema entrega predicciones de suficiencia y una tabla con los valores esperados por rubro presupuestario.
- **Módulo de procesamiento de datos:** limpieza, transformación y validación de los datos históricos estructurados.
- **Módulo de entrenamiento:** entrenamiento de dos modelos: uno de clasificación binaria (suficiencia del préstamo) y otro de regresión multirrubro (predicción de costos por partida).
- **Módulo de visualización:** Presenta los resultados a través de gráficos, tablas y explicaciones interpretables, facilitando el uso por parte del equipo financiero.

En la figura se presenta el diagrama de flujo de trabajo, que muestra la secuencia de pasos: extracción de datos, procesamiento de datos, inferencia, visualización y finalmente la presentación de resultados.



- Orientador: a definir.

- Cliente: Martin Brambati tiene muchos años de experiencia en el puesto de manager y director de ingeniería. Su experiencia puede ser muy útil para el desarrollo del proyecto.
- Colaboradores: Thomas Schlegel tiene más de 8 años de experiencia a cargo del área de datos en Built Technologies, lo cual será de mucha ayuda a la hora de entender el dataset.
- Opositores: Procore Technologies proporciona una plataforma unificada de gestión financiera de proyectos de construcción. La empresa ya ha incorporado herramientas de inteligencia artificial, por lo que tendría intenciones de que este proyecto no llegue a concluirse.

3. Propósito del proyecto

Desarrollar una solución basada en modelos de aprendizaje automático que permita asistir a los analistas financieros de Built Technologies en la evaluación de la suficiencia de préstamos otorgados para proyectos de construcción, así como en la estimación detallada de los costos asociados a cada una de las partidas presupuestarias. La solución busca anticipar riesgos de subfinanciamiento y sobrecostos mediante el análisis de datos históricos anonimizados, promoviendo decisiones más informadas, eficientes y escalables dentro del flujo de originación de créditos.

- Orientador: a definir.

- **Alcance del proyecto:** Martin Brambati posee una amplia trayectoria como mánager y director de ingeniería. Su experiencia será de gran valor para el desarrollo del proyecto.

El presente proyecto incluye:

Role	Nombre y Apellido del participante	Organización	Puesto	a la evaluación de
Cliente/satisfactoria	Martín Brany batíumación	Built Technologies	Engineering Manager	
Responsable	Ing. Danilo Simón Reitano Andrades	FIUBA	Alumno	
Colaboradores	Thomas Schlegel	Built Technologies	Distinguished Engineer	
Orientador	A definir	A definir	Director del Trabajo Final	
Opositores	-	Procore Technologies	-	
Usuario final	Tratamiento de valores faltantes	Clients de Built	-	

- Codificación de variables categóricas.
- Colaboradores: Thomas Schlegel cuenta con más de 8 años de experiencia a cargo del área de datos en Built Technologies, lo que será de gran ayuda para la comprensión (para conjunto de datos) de clasificación).
- Opositores: Procore Technologies proporciona una plataforma unificada de gestión financiera de proyectos de construcción. La empresa incorporó anteriormente herramientas de inteligencia artificial, por la que tendría intenciones de que este proyecto no llegue a concluirse.
 - Un modelo de regresión multivariada para estimar los costos esperados por categoría presupuestaria (por ejemplo: cimentación, electricidad, techo, etc.).

3. Propósito del proyecto

- Evaluación de los modelos desarrollados mediante métricas adecuadas (F1-score, AUC-ROC, MAE, RMSE).

Desarrollar una solución basada en modelos de aprendizaje automático que permita asistir a los analistas financieros de Built Technologies en la evaluación de la suficiencia de préstamos otorgados para proyectos de construcción. Dicha solución deberá orientar en la estimación detallada de los costos asociados a cada una de las partidas presupuestarias. La solución busca anticipar riesgos de subfinanciamiento y sobrecostos mediante el análisis de datos históricos anonimizados. Esto produce decisiones más informadas, eficientes y escalables dentro del flujo de generación de créditos.

El presente proyecto no incluye:

4. Alcance del proyecto

- El desarrollo de una interfaz gráfica de usuario.

El presente proyecto incluye: los modelos desarrollados en los sistemas de producción de Built Technologies.

- La automatización completa del pipeline en un entorno de producción: la evaluación de suficiencia de préstamos y estimación presupuestaria en proyectos de construcción.
- La toma de decisiones finales sobre políticas de crédito, las cuales quedarán en manos del equipo financiero de la empresa.
- Exploración, limpieza y transformación del conjunto de datos históricos anonimizados proporcionado por Built Technologies.
- El análisis de documentos no estructurados.
- Desarrollo de un pipeline de preprocesamiento de datos, que incluirá:

5. Supuestos del proyecto

- Tratamiento de valores faltantes.

- Codificación de variables categóricas.

Para el desarrollo del presente proyecto se supone que:

- Normalización y estandarización de variables numéricas.
- Balanceo de clases en caso de desbalance significativo en la variable objetivo (para el modelo de clasificación).
- El dataset estructurado y anonimizado proporcionado por Built Technologies estará disponible desde el inicio del proyecto, y contará con la calidad y cantidad suficientes para el entrenamiento de modelos de aprendizaje automático.

- No se dispone de un modelo de clasificación binaria para predecir si el préstamo solicitado es suficiente para cubrir el presupuesto total.
- El alcance del proyecto se mantendrá centrado en el desarrollo de modelos predictivos
 - Un modelo de regresión multivariada para estimar los costos esperados por categoría (clasificación y regresión), sin requerir su integración en producción o implementación de presupuestaria (por ejemplo: cimentación, electricidad, techado, etc.).
- Evaluación de los modelos desarrollados mediante métricas adecuadas (*F1-score*, *AUC-ROC*, *MAE*, *RMSE*).
Se dispondrá de al menos 8 horas semanales para dedicar al proyecto a lo largo del año calendario 2025, equilibrando responsabilidades laborales con los tiempos del posgrado.
- Generación de explicaciones interpretables de las predicciones con técnicas como *SHAP* o *Feature Importance*. Se contará con acceso continuo a recursos computacionales adecuados (principalmente Jupyter notebooks, Python y bibliotecas de ML como scikit-learn, XGBoost, pandas y *SHAP*) sin necesidad de infraestructura especializada ni servicios cloud pagos adicionales.
- Documentación del proceso completo y presentación de los resultados en un formato replicable y académico.
- Se contará con la disponibilidad del director del trabajo final para realizar revisiones metodológicas periódicas y seguimiento académico del progreso del proyecto.

El presente proyecto no incluye:

6. Product Backlog

- El desarrollo de una interfaz gráfica de usuario.

El *Product Backlog* debe organizarse en *epics* fundamentales del proyecto. Cada *built technology* debe contener al menos dos historias de usuario que describan funcionalidades clave.

El *Product Backlog* debe permitir interpretar cómo será el proyecto y su funcionalidad. Se deben indicar claramente las prioridades entre las historias de usuario y si hay alguna opcional.
La automatización completa del pipeline en un entorno de producción.

Las historias de usuario deben ser breves, claras y medibles, expresando el rol, la necesidad y el propósito de cada funcionalidad. También deben tener una prioridad definida para facilitar la planificación de los sprints.

Cada historia de usuario debe incluir una ponderación en *Story Points*, un número entero que representa el tamaño relativo de la historia. El criterio para calcular los Story Points debe indicarse explícitamente.

Para el desarrollo del presente proyecto se supone que:

Las historias deben seguir el formato: “*Como [rol], quiero [tal cosa] para [tal otra cosa]*”.

Las épicas deben estructurarse de la siguiente forma:
■ El dataset estructurado y anonimizado proporcionado por Built Technologies estará disponible desde el inicio del proyecto, y contará con la calidad y cantidad suficientes para el entrenamiento de modelos de aprendizaje automático.

- **Épica 1**
 - No se requerirá solicitar acceso a datos sensibles o confidenciales.
 - HUI
 - El alcance del proyecto se mantendrá centrado en el desarrollo de modelos predictivos (clasificación y regresión), sin requerir su integración en producción o implementación de interfaces visuales para usuarios finales.

- Se dispondrá de una dedicación mínima de 8 horas semanales para el proyecto a lo largo del año calendario 2025, con el propósito de compatibilizar las responsabilidades laborales con los tiempos del posgrado.

■ Épica 3

- Se contará con acceso continuo a recursos computacionales adecuados (principalmente Jupyter notebooks, Python y bibliotecas de ML como scikit-learn, XGBoost, pandas y *SHAP*) sin necesidad de infraestructura especializada ni servicios cloud pagos adicionales.

■ Épica 4

- Se contará con la disponibilidad del director del trabajo final para realizar revisiones metodológicas periódicas y seguimiento académico del progreso del proyecto.

6. Product Backlog

- HU8

El criterio utilizado para asignar los *Story Points* se basa en una escala relativa de complejidad, esfuerzo y riesgo:

Reglas para definir historias de usuario:

- 1 punto: tarea simple, conocida, sin incertidumbre técnica.
- Ser concisas y claras.
- 2–3 puntos: tarea con un nivel medio de procesamiento o exploración de datos.
■ Expresarlas en términos cuantificables y medibles.
- 5 puntos: tarea técnica de mediana complejidad o con validaciones múltiples.
■ No dejar margen para interpretaciones ambiguas.
- 8+ puntos: tarea compleja o con alto nivel de incertidumbre en datos o rendimiento del modelo.
■ Indicar claramente su prioridad y si son opcionales.
- Considerar regulaciones y normas vigentes.

A continuación, se detallan las épicas y sus respectivas historias de usuario:

7. Criterios de aceptación de historias de usuario

■ Épica 1: planificación y organización del proyecto

- **HU1:** Como responsable del proyecto, quiero definir un cronograma tentativo con fases y sprints para distribuir el trabajo de manera equilibrada.
Los criterios de aceptación deben establecerse para cada historia de usuario, asegurando que se cumplan las condiciones necesarias para que la funcionalidad sea validada correctamente.

Prioridad: Alta — Story Points: 3

Cada historia debe tener criterios medibles, específicos y verificables. Deben permitir validar que se cumplen con las necesidades del usuario encontrados.

Prioridad: Media — Story Points: 2

Se estructuran de forma análoga a las épicas del backlog:

■ Épica 2: relevamiento y análisis de datos históricos

- **Épica 1:** Como analista de datos, quiero explorar y limpiar el dataset histórico para asegurar que los datos sean utilizables para modelado.

• Prioridad: Alta — Story Points: 3

- **Épica 2:** Como ingeniero de datos, quiero identificar las variables más relevantes para el análisis, clasificándolas según tipo y calidad.

• Prioridad: Alta — Story Points: 3

■ Épica 3: preprocessamiento y preparación de datos

- Criterios de aceptación HU4

- **HU5:** Como científico de datos, quiero balancear las clases de la variable objetivo para asegurar un entrenamiento adecuado del modelo de clasificación.

Prioridad: Media — Story Points: 5

- Criterios de aceptación HU5

- **HU6:** Como científico de datos, quiero normalizar y codificar las variables categóricas y numéricas para que puedan ser interpretadas por los modelos.

Prioridad: Alta — Story Points: 3

■ Épica 4

■ Épica 4: entrenamiento y validación de modelos

- Criterios de aceptación HU7

- **HU7:** Como desarrollador de modelos, quiero entrenar un modelo de clasificación para predecir si los presupuestos cumplen con su objetivo, y evaluar su rendimiento con métricas como *F1-score* y *AUC*.

Reglas para la Épica 4: Prioridad: Alta — Story Points: 5

- **HU8:** Como desarrollador de modelos, quiero entrenar un modelo de regresión para estimar los costos por partida presupuestaria, y analizar su precisión mediante *MAE* y *RMSE*.

• Especificación: Prioridad: Alta — Story Points: 5 completada.

- Épica 5: interpretación, documentación y validación

- No. **HU9:** Como analista, quiero aplicar técnicas de interpretabilidad (*SHAP, Feature Importance*) para entender las variables que más influyen en las predicciones.

■ Probables de testear funcional o técnicamente.
Prioridad: Media — Story Points: 3

- Mínimo. **HU10:** Como responsable del proyecto, quiero documentar el proceso, los resultados y sus limitaciones para facilitar su presentación y revisión académica.

Prioridad: Alta — Story Points: 2

8. Fases de CRISP-DM

- Épica 6: sistematización y documentación técnica

1. **HU11:** Como autor del modelo, quiero documentar todas las decisiones de preprocesamiento y justificación de selección de variables para asegurar trazabilidad.

Prioridad: Alta — Story Points: 3

2. **HU12:** Como desarrolladora, quiero versionar el código y registrar los experimentos de modelado para facilitar replicabilidad futura.

Prioridad: Media — Story Points: 3

3. **Evaluación del modelo:** métricas de rendimiento.

4. **Despliegue del modelo (opcional):** tipo de despliegue y herramientas.
HU13: Como alumno de posgrado, quiero redactar la memoria escrita del trabajo final, con antecedentes, metodología, resultados y conclusiones.

Prioridad: Alta — Story Points: 8

9. Desglose del trabajo en tareas

- **HU14:** Como autor, quiero adaptar el documento a los lineamientos formales de la universidad para asegurar su correcta presentación.

A partir de **Prioridad: Alta** y **8bfStory Points: 3** se detallan las tareas, técnicas y medibles:

- Épica 8: preparación de la defensa oral

- Duración estimada: entre 2 y 8 h. Evitar tareas genéricas.

- **HU15:** Como expositor, quiero preparar una presentación clara y visual para explicar

■ Si una tarea excede 8 h, dividirla en partes y explicar la problemática, la solución y los resultados a un jurado.
Prioridad: Alta — Story Points: 5

- **HU16:** Como expositor, quiero ensayar la defensa oral para responder preguntas técnicas y asegurarme de cumplir con los tiempos estipulados.

Historia de usuario	Tarea Técnica	Estimación	Prioridad
HU1	Tarea 1 HU1	6 h	Alta
HU1	Tarea 2 HU1	8 h	Alta
HU2	Tarea 1 HU2	5 h	Media
HU2	Tarea 2 HU2	6 h	Alta
...

- Épica 1: planificación y organización del proyecto

Criterios para estimar tiempos:

- Criterios de aceptación HU1

- Considerar la complejidad técnica y gravedad de incertidumbre de cada tarea.
○ Se ha elaborado un cronograma tentativo con fases alineadas al backlog.
- El cronograma incluye tiempos estimados por tarea y asignación tentativa por sprint.

- Evitar subestimar el esfuerzo requerido. Si una tarea supera las 8 h, dividirla en subtareas.

- Basar la estimación en la experiencia propia o en referencias de tareas similares.

- Criterios de aceptación HU2

Sobre la priorización: Se han identificado los ajustes realizados al backlog o tareas planificadas.

- Los cambios se encuentran justificados en base a imprevistos o progresos.
- Asignar el backlog actualizado (ha sido revisado 3 veces) a los menos relevantes de acuerdo a los hitos importantes de la tarea y su impacto en los entregables.

- **Épica 2: relevamiento y análisis de datos históricos** Estación de las HU o que sean necesarias para desbloquear otras.
 - **Criterios de aceptación HU3**

- Incluir tareas optionales solo si están bien justificadas
 - El conjunto de datos ha sido cargado correctamente en el entorno de trabajo.
 - Se identificaron y eliminaron valores atípicos y registros duplicados.

Recomendaciones generales: Se generó un informe exploratorio con estadísticas descriptivas y visualizaciones básicas.

- Incluir **Criterios de aceptación HU4** de usuario.
 - Se ha identificado el conjunto de variables relevantes para el modelo.
 - El total estimado debe ser coherente con la planificación global de unas 600 horas (la cantidad de horas sugeridas para un trabajo de posgrado).
 - Las variables han sido clasificadas como numéricas o categóricas.
 - Se documentó la justificación técnica de la selección de cada variable.
- Este desglose se utilizará en las secciones siguientes para armar el diagrama de Gantt.
- **Épica 3: preprocesamiento y preparación de datos** (sección 10) y la planificación de sprint (sección 11).
 - **Criterios de aceptación HU5** Recordar que la calidad del desglose impacta directamente en la calidad de la planificación del proyecto. Se analizó la distribución de la variable objetivo y se detectó desbalance significativo (si aplica).
 - Se aplicó una técnica de balanceo (*undersampling*, *oversampling* o *SMOTE*).
 - **Criterios de aceptación HU6** Se verificó que el modelo no pierda precisión tras aplicar balanceo.

10. Diagrama de Gantt

○ Se verificó que el modelo no pierda precisión tras aplicar balanceo.

- **Criterios de aceptación HU6**

El diagrama de Gantt debe representar de forma visual y cronológica todas las tareas del proyecto, abarcando aproximadamente 600 horas totales, de las cuales entre 480 y 500 deben destinarse a tareas técnicas (desarrollo, pruebas, implementación) y entre 100 y 120 a tareas no técnicas (planificación, documentación, escritura de memoria y preparación de la defensa).

- Todas las variables categóricas han sido codificadas mediante *one-hot encoding* o similar.
- Las variables numéricas fueron normalizadas o estandarizadas.
- El conjunto de datos final está listo para ser consumido por los modelos.

Consignas y recomendaciones:

- **Epica 4: entrenamiento y validación de modelos**

- **Criterios de aceptación HU7** Incluir tanto tareas técnicas derivadas de las HU como tareas no técnicas generales del proyecto. El modelo de clasificación fue entrenado con un conjunto dividido en entrenamiento, validación y prueba.
 - El eje vertical debe listar las tareas y el eje horizontal representar el tiempo en semanas o fechas.
 - Se reportan métricas *F1-score* y *AUC* en el conjunto de prueba.
 - Se alcanza un *F1-score* mayor al valor base definido como mínimo aceptable.
- Utilizar **Criterios de aceptación HU8** para distinguir tareas técnicas y no técnicas.
 - Se ha entrenado un modelo de regresión con variables de entrada preprocesadas.
 - Se calculan métricas *MAE*, *RMSE* y *R²* sobre el conjunto de prueba.
- Iniciar con la planificación del proyecto (coincidente con el inicio de Gestión de Proyectos) y finalizar con la defensa, próxima a la fecha de cierre del trabajo.
- **Épica 5: interpretación, documentación y validación** Configurar el software para mostrar los códigos del desglose de tareas y los nombres junto a cada **Criterios de aceptación HU9**
 - Se aplicó una técnica de explicabilidad (ej. *SHAP*) sobre los modelos entrenados.
 - Asegurarse de que la fecha final coincida con la del Acta Constitutiva.
 - Se identificaron las variables más influyentes en las predicciones.
- Evitar tareas genéricas o ambiguas y asegurar una secuencia lógica y realista. Los resultados de interpretabilidad se incluyen en un informe gráfico y textual.
- Las **Criterios de aceptación HU10** fechas pueden ser aproximadas, ajustar el ancho del diagrama según el texto y el parámetro *width*. Se generó un documento técnico con los pasos realizados, decisiones tomadas y valor y, quizás, los nombres de las tareas.
 - El documento incluye gráficas de métricas y análisis de errores.

Herramientas sugeridas:

- **Epica 6: sistematización y documentación técnica**

■ Plan Criterios de aceptación HU11

- <https://blog.trello.com/es/diagrama-de-gantt-de-un-proyecto> Las decisiones de preprocesamiento fueron registradas en un documento o notebook.

■ Creately (colaborativa online)

- <https://creately.com/diagram/example/1e63p3m1/LaTeX> Las transformaciones realizadas son reproducibles en otros entornos.

- El documento técnico describe el flujo completo de tratamiento de datos.

■ LaTeX con pgfgantt:

- <http://ctan.dcc.uchile.cl/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf>

- Se utilizó control de versiones (Git) para registrar avances del proyecto.

- Los experimentos de modelado se guardaron en forma estructurada (scripts, parámetros).

Incluir una imagen legible del diagrama de Gantt. Si es muy ancho, presentar primero la tabla y luego el gráfico de barras.

- La replicabilidad fue validada con el pipeline en una sesión independiente.

■ Épica 7: Redacción del trabajo final

11. Planificación de Sprints

• Criterios de aceptación HU13

Organizar las tareas técnicas del proyecto en sprints de trabajo que permitan distribuir de forma equilibrada la carga horaria total, estimada en 600 horas.

- Se incluyeron tablas, gráficos y citas académicas relevantes.

Consigna: ○ La versión final fue revisada y corregida en base a retroalimentación del orientador.

■ Completar Criterios de aceptación HU14 con HU y tareas técnicas correspondientes.

■ Incluir estimación en horas para cada tarea.

- Se verificó cumplimiento de normas de estilo, ortografía y citas bibliográficas.

■ Indicar responsable y porcentaje de avance estimado o completado.

- El archivo final está listo para ser presentado ante la coordinación académica.

■ Épica 8: Preparación de la defensa oral, documentación, redacción de memoria y preparación de defensa.

• Criterios de aceptación HU15

Conceptos clave: ○ Se elaboró una presentación visual con estructura clara (problema, solución, resultados).

- Se incorporaron visualizaciones clave del modelo y sus métricas.

■ Una épica es una unidad funcional amplia; una historia de usuario es una funcionalidad concreta; un sprint es una unidad de tiempo donde se ejecutan tareas.

• Criterios de aceptación HU16

■ Las tareas son el nivel más desagregado: permiten estimar tiempos, asignar responsables y monitorear progreso.

- Se realizaron al menos dos ensayos de defensa oral (simulados).

- Se practicaron respuestas a posibles preguntas del jurado técnico.

- Se ajustó el discurso para cumplir con el tiempo estipulado sin omitir secciones importantes.

Duración sugerida:

■ Para un proyecto de 600 h, se recomienda planificar entre 10 y 12 sprints de aproximadamente 2 semanas cada uno.

8. Fases de CRISP-DM

■ Asignar entre 45 y 50 horas efectivas por sprint a tareas técnicas.

1. Comprensión del negocio: El objetivo principal del proyecto es asistir a analistas

■ Recopilar 100 requerimientos mediante modelos predictivos que permitan: reuniones, defensa).

- Estimar si el préstamo otorgado será suficiente para cubrir el presupuesto total de un proyecto de construcción.

Importante:

- Predecir el costo esperado por categoría presupuestaria (materiales, mano de obra, cimentación, etc.).

- En proyectos individuales, el responsable suele ser el propio autor.

- El valor agregado de incorporar IA para el antecipación riesgos de subfinanciamiento y sobrecostos desde etapas tempranas del proceso crediticio. **Métricas de éxito:** precisión del modelo (*F1-score* y *AUC* para clasificación, *MAE* y *R²* para regresión), interpretabilidad y aplicabilidad práctica para el analista.

- 2. Comprendimiento de los datos:** El dataset proviene de registros históricos anonimizados de Built Technologies, con más de 10 años de datos sobre proyectos financiados.

- Tener en cuenta aproximaciones tipo Fibonacci.
 - **Tipo:** datos estructurados tabulares, con variables numéricas y categóricas.
 - **Origen:** base interna de proyectos y préstamos otorgados.
 - **Cantidad:** en el orden de los millones de registros.

Sprint	H1 o fase	Tarea	Horas SP	Responsable	% Completado
Sprint 0	■ Calidad: excelente, pero con presencia de valores faltantes, codificación inconsistente de categorías y posibles outliers.	Definir alcance y cronograma	10 h	Alumno	100 %
Sprint 1	Preparación de los datos:	Esta etapa incluye limpieza y transformación del dataset para su uso por los modelos.			
Sprint 0	■ Planificación	Ajuste de entre-registros con errores evidentes o duplicados.	6 h	Alumno	25 %
Sprint 1	■ H1	Tarea 1 H1	6 h / 3 SP	Alumno	0 %
Sprint 1	■ H2	Tarea 2 H2	6 h / 3 SP	Alumno	0 %
Sprint 2	H1	Normalización y Estandarización de variables numéricas.	7 h / 3 SP	Alumno	0 %
...		Balanceo de clases para la variable objetivo en el modelo de clasificación.			...
Sprint 5	Escríptura	Redacción memo- rial	50 h / 34 SP	Alumno	0 %
		Selección de variables relevantes mediante análisis exploratorio y técnicas automáticas (<i>feature importance</i> , selección recursiva).			
Sprint 6	Defensa	Preparación	20 h / 13 SP	Alumno	0 %

- 4. Modelado:** Se abordan dos problemas distintos:

- Recomendaciones:** ■ **Probabilidad de completitud:** predecir si el préstamo será suficiente para cubrir los costos.

- **Regresión multivariada:** estimar el costo por partida presupuestaria.
- Verificar que la carga horaria por sprint sea equilibrada.
- Los algoritmos candidatos incluyen:
 - Usar sprints de 1 a 3 semanas, acordes al cronograma general.
 - **Para clasificación:** *Random Forest*, *XGBoost*, Regresión Logística.
 - Actualizar el % completado durante el seguimiento del proyecto.
 - **Para regresión:** *XGBoost Regressor*, *LightGBM*, redes neuronales densas.
 - Considerar un sprint final exclusivo para pruebas, revisión y ajustes antes de la defensa. Se compararán diferentes modelos y se realizará validación cruzada.

- 5. Evaluación del modelo:** Se utilizarán diferentes métricas de rendimiento por tipo de modelo.

12. NORMATIVA Y CUMPLIMIENTO DE DATOS (GOBERNANZA)

- **Clasificación:** *F1-score*, precisión, *recall*, *AUC-ROC*.

En esta sección se debe analizar si los datos utilizados en el proyecto están sujetos a normativas de protección de datos y privacidad, y en qué condiciones se pueden emplear.

Aspectos adicionales: se aplicarán técnicas de interpretabilidad (*SHAP*, *feature importance*) para validar que los modelos son comprensibles para usuarios no técnicos.

- 6. Despliegue del modelo (opcional):** Dado que el proyecto es académico, no se contempla un despliegue a producción. No obstante, se documentará cómo podría integrarse el modelo en un pipeline de análisis crediticio dentro de Built Technologies:
- Determinar si el uso de los datos requiere consentimiento explícito de los usuarios involucrados.
 - Exportación del modelo entrenado en formato compatible (.pkl, .joblib).
 - Sugerencia de integración vía API REST o servicio batch interno.
 - Indicar si existen restricciones legales, técnicas o contractuales sobre el uso, compartición o publicación de los datos.
 - Propuesta de visualización simple para interpretación de resultados.

9. Desglose del trabajo en tareas

fuentes licenciadas, de acceso público o bajo algún tipo de autorización especial.

El siguiente desglose de tareas se realizó a partir de las historias de usuario definidas en el Product Backlog. Cada tarea fue especificada de manera técnica, concreta y medible, con el fin de facilitar la posterior planificación en sprints y la elaboración del diagrama de Gantt.

Este análisis es clave para garantizar el cumplimiento normativo y evitar conflictos legales. La estimación en horas se basa en una evaluación del grado de dificultad técnica, la complejidad algorítmica involucrada, la posible necesidad de investigación exploratoria, y el nivel de incertidumbre asociado a cada actividad.

A su vez, cada tarea ha sido asignada una prioridad relativa (Alta, Media o Baja) en función de su impacto en el cumplimiento de los criterios de aceptación, su relevancia en el ciclo de vida del modelo y su efecto habilitador sobre otras tareas posteriores.

Este desglose representa una estimación aproximada de 600 horas efectivas, donde se cubren las tareas fundamentales asociadas a la construcción del modelo, la validación técnica y la documentación final. A lo largo del desarrollo del proyecto, se podrá ajustar el detalle de tareas, subdividir algunas de mayor complejidad o incorporar tareas adicionales en función de descubrimientos o desafíos surgidos en fases intermedias.

Este enfoque estructurado garantiza la trazabilidad del avance y facilitará una gestión iterativa del trabajo a lo largo de los sprints definidos en la planificación.

Riesgo	S	O	RPN	S*	O*	RPN*

Historia de usuario	Tarea técnica	Estimación	Prioridad
HU1	Analizar el alcance del proyecto y elaborar una primera versión del cronograma general	6 h	Alta
HU1	Dividir el proyecto en fases alineadas con las épicas e identificar dependencias entre ellas	4 h	Alta
HU1	Diseñar un plan de sprints tentativo con hitos intermedios y fechas de revisión	6 h	Alta
HU1	Revisar el cronograma con el director y ajustar el plan según observaciones	2 h	Alta
HU2	Establecer una rutina de seguimiento semanal o quincenal para evaluar avances y desvíos	2 h	Media
HU2	Actualizar backlog y tareas en función de retroalimentación o bloqueos técnicos	4 h	Media
14. Sprint Review	Ajustar cronograma o redistribuir tareas en función del progreso real (replanificación)	4 h	Media
HU2	La revisión HU2 sprint (<i>Sprint Review</i>) es una revisión formal en la metodología Ágiles. Documentar los cambios realizados sobre el plan y justificar desviaciones. Consiste en revisar y evaluar el plan y justificar desviaciones.	2 h	Media
HU3	se presentan los avances y se verifica si las funcionalidades cumplen con los criterios de aceptación establecidos. También se identifican errores y se consideran ajustes si es necesario.	4 h	Alta
HU3	Aunque el proyecto aún se encuentre en etapa de planificación, esta sección permite proyectar cómo se evaluarán las funcionalidades más importantes del backlog. Esta mirada anticipada favorece la planificación enfocada en valor y permite reflexionar sobre posibles obstáculos.	4 h	Alta
HU3	Identificar y cuantificar valores nulos, inconsistentes o duplicados	6 h	Alta
Objetivo: anticipar cómo se evaluará el avance del proyecto a medida que se desarrollen las funcionalidades, utilizando como base al menos cuatro historias de usuario del <i>Product Backlog</i> .	Realizar limpieza inicial: imputación, eliminación de duplicados y errores	6 h	Alta
HU3	Analizar la distribución de variables	6 h	Alta
Seleccionar al menos 4 HU numéricas mediante Product Backlog. Para cada una, completar la siguiente tabla de revisión proyectada:	plots		
HU3 Formato sugerido:	Detectar y tratar outliers con técnicas estadísticas (IQR, Z-score)	6 h	Alta
HU3	Generar un informe exploratorio con visualizaciones y estadísticas descriptivas	6 h	Alta
HU3	Dividir el dataset en subconjuntos por tipo de proyecto o año (si aplica)	4 h	Media
HU3	Documentar el pipeline de limpieza con justificación de decisiones	4 h	Alta
HU3 (Complementaria)	Investigar e incorporar información contextual externa (región, inflación, etc.)	6 h	Media

HU	Tareas	Tarea técnica	¿Cómo sabrás que el entregable es esperado?	Estimación	Prioridad
HU4 seleccionada		Clasificar variables por tipo (numéricas, categóricas, temporales, target)	(numéricas, categóricas, temporales, target) cumplida?	4 h o riesgos	Alta
HU4 HU1	Tarea 1	Evaluar correlaciones entre variables numéricas y redundancia (matriz de correlación)	variables seriadas y redundancia (matriz adén definidos)	6 h	Alta Falta validar con el tutor
	Tarea 2				
HU4 HU3	Tarea 1	Analizar cardinalidad de variables categóricas	Exportación y detectar categorías raras o desbalanceadas	4 h	Media Requiere datos reales
	Tarea 2				
HU4 HU5	Tarea 1	Realizar análisis de varianza (ANOVA) o chi-cuadrado para evaluar importancias de variables	R (ANOVA) o chi-cuadrado para evaluar importancias operativas	6 h	Alta Riesgo en integración
	Tarea 2	Generar ranking de variables relevantes con técnicas automáticas (feature importance)	Informe trimestral PDF con gráficos y evolución	6 h	Alta Puede faltar tiempo para ajustes
HU4		Redactar informe técnico con las variables seleccionadas, criterios y limitaciones		6 h	Alta
15. Sprint Retrospective					
HU4 (Opcional) La retrospectiva de sprint es una práctica orientada a la mejora continua. Al finalizar un sprint, el equipo (o el alumno, si trabaja de forma individual) reflexiona sobre lo que funcionó bien, lo que puede mejorar y qué acciones concretas pueden implementarse para trabajar mejor en el futuro.		Aplicar técnicas de reducción de dimensionalidad (PCA o UMAP) y evaluar impacto		6 h	Baja
HU4 (Complementaria) Objetivo: reflexionar sobre las condiciones iniciales del proyecto, identificando anticipadamente posibles dificultades y estrategias de mejora, incluso antes del inicio del desarrollo.		Visualizar relaciones multivariadas mediante pairplots o mapas de calor		4 h	Media
HU4 (Opcional) Durante la cursada se propone entenderse de la Estrella de los proyectos. Durante la retrospectiva, que organiza la reflexión en cinco ejes.		Realizar clustering exploratorio para entender segmentos		6 h	Baja
HU5		Analizar distribución de la variable objetivo y su desbalance (visual y cuantitativa)		4 h	Media
■ ¿Qué hacer más?		HU5	Implementar técnica de balanceo simple (undersampling o oversampling clásico)	4 h	Media
■ ¿Qué hacer menos?		HU5	Evaluar impacto del balanceo sobre la distribución de variables	4 h	Media
■ ¿Qué mantener?		HU5	Implementar SMOTE y variantes avanzadas (BorderlineSMOTE, ADASYN)	6 h	Alta
■ ¿Qué empezar a hacer?		HU5	Comparar modelos entrenados con y sin balanceo y medir impacto en F1-score	6 h	Alta
■ ¿Qué dejar de hacer?		HU5	Documentar estrategia de balanceo	4 h	Alta
Aun en una etapa temprana adoptada y razones de su elección		HU5	Ayudará a planificar su forma de trabajar, identificando anticipadamente posibles dificultades y diseño de estrategias de organización personal.	4 h	Alta
Objetivo: reflexionar sobre las condiciones iniciales del proyecto, identificando fortalezas, posibles dificultades y estrategias de mejora, incluso antes del inicio del desarrollo.		HU5 (Complementaria)	Probar técnicas de balanceo basadas en generación de ruido o augmentación sintética	4 h	Baja
Completar la siguiente tabla tomando como referencia los cinco ejes de la Estrella de la Retrospectiva (Starfish o estrella de mar). Esta instancia te ayudará a definir buenas prácticas desde el inicio y prepararte para enfrentar tu trabajo de forma organizada y flexible. Se deberá completar la tabla al menos para 3 sprints técnicos y 1 no técnico.		HU6	Codificar variables categóricas nominales con one-hot encoding	4 h	Alta
		HU6	Codificar variables categóricas ordinales con label encoding o mappings personalizados	4 h	Alta
Formato sugerido:			Normalizar variables numéricas (min-max) y evaluar escalas	4 h	Alta

Historia de usuario		Tarea técnica	Estimación	Prioridad	
HU6		Estandarizar variables numéricas (z-score) y comparar con normalización	4 h	Alta	
HU6		Analizar la sensibilidad de los modelos a diferentes esquemas de codificación	4 h	Media	
HU6		Evaluar correlaciones post-codificación para evitar colinealidades artificiales	4 h	Media	
HU6		Implementar un pipeline reproducible de preprocessamiento ('Pipeline' de scikit-learn)	6 h	Alta	
HU6		Validar integridad de los datos procesados (dimensiones, tipos, escalas esperadas)	4 h	Alta	
HU6		Documentar las decisiones de transformación de datos, con gráficos de antes y después	4 h	Alta	
Sprint tipo y N°	¿Qué hacer más?	¿Qué hacer menos?	¿Qué mantener?	¿Qué empezar a hacer?	¿Qué dejar de hacer?
HU6 (Opcional)				6 h	Baja
Sprint técnico - 1	Validaciones continuas con el alumno	Cambios sin versión registrada	Pruebas con datos simulados	Documentar cambios propuestos	Ajustes sin análisis de impacto
HU6 (Opcional)				4 h	Baja
Sprint técnico - 2	Verificar configuraciones	Mapear ejecución de los modelos	Obtener más rápidas métricas reutilizables	User logs para	Repetir pruebas
HU7 en múltiples escenarios		Seleccionar y justificar algoritmos candidatos para clasificación (p. ej., RF, XGBoost)	configuración	4 h	Alta innecesarias
Sprint técnico - 8	Comparar correlaciones con casos previos	Cambiar parámetros sin justificar	Revisión cruzada de los datos preprocesados	Anotar configuraciones usadas	Trabajar sin respaldo de datos
HU7	Ensayos	Realizar validación cruzada (k-fold) y medir F1-score y AUC promedio	6 h	Alta	
Sprint técnico - 12 (por ej.: "Defensa")	Entrales con feedback	Ajustar hiperparámetros (grid search o random search) y registrar mejoras por bloques	Dividir la presentación	Agregar gráficos	
HU7		Evaluar modelo en conjunto de prueba (hold-out) y registrar métricas finales	4 h	Alta	
HU7		Analizar matriz de confusión y distribución de errores por clase	4 h	Media	
HU7		Comparar resultados con modelo base (regresión logística o árbol simple)	4 h	Media	
HU7		Documentar arquitectura, parámetros y desempeño del modelo final de clasificación	4 h	Alta	
HU7 (Complementaria)		Entrenar variante del modelo de clasificación con LightGBM	4 h	Media	
HU7 (Opcional)		Implementar curva Precision-Recall y optimizar umbral de decisión	4 h	Baja	
HU8		Seleccionar algoritmos de regresión adecuados (XGBoost, LightGBM, regresión múltiple)	4 h	Alta	

Historia de usuario	Tarea técnica	Estimación	Prioridad
HU8	Implementar modelo base de regresión y entrenar con datos preprocesados	6 h	Alta
HU8	Realizar validación cruzada (k-fold) y evaluar MAE, RMSE, R ²	6 h	Alta
HU8	Aplicar técnicas de regularización (Lasso, Ridge) y comparar impacto	4 h	Media
HU8	Afinar hiperparámetros y evaluar mejora sobre conjunto de validación	6 h	Alta
HU8	Evaluando modelo en conjunto de prueba, generar gráfico de dispersión real vs. predicho	4 h	Alta
HU8	Analizar errores por categoría presupuestaria y posibles sesgos	4 h	Media
HU8	Documentar resultados del modelo final y su aplicabilidad práctica	4 h	Alta
HU8 (Complementaria)	Comparar modelo de regresión con red neuronal simple (MLP)	6 h	Baja
HU8 (Opcional)	Evaluando sensibilidad del modelo a outliers y aplicar técnicas de robustez	4 h	Baja
HU9	Implementar método de interpretabilidad basado en SHAP para el modelo de clasificación	6 h	Alta
HU9	Visualizar los valores SHAP globales (summary plot) e identificar variables más influyentes	4 h	Alta
HU9	Generar interpretaciones locales de predicciones individuales (force plots)	4 h	Media
HU9	Aplicar análisis de importancia de variables por método de permutación (como alternativa)	4 h	Media
HU9	Comparar resultados de SHAP con Feature Importance tradicional (XGBoost, LightGBM)	4 h	Media
HU9	Evaluando consistencia entre modelos de clasificación y regresión respecto a variables clave	4 h	Media
HU9	Preparar gráficos explicativos de interpretabilidad para uso en el informe final	4 h	Alta
HU9 (Complementaria)	Explorar herramientas de explicabilidad adicionales (LIME, ELI5) y comparar resultados	6 h	Baja
HU9 (Opcional)	Generar dashboard interactivo de interpretabilidad con SHAP o Plotly Dash	6 h	Baja
HU10	Redactar sección metodológica detallada sobre interpretabilidad y validación del modelo	6 h	Alta
HU10	Documentar errores comunes, desviaciones y limitaciones técnicas del enfoque usado	6 h	Alta

Historia de usuario	Tarea técnica	Estimación	Prioridad
HU10	Organizar todos los resultados intermedios en un repositorio de evidencia (figuras, métricas, logs)	4 h	Alta
HU10	Preparar anexos técnicos (tablas, configuraciones, parámetros de entrenamiento)	4 h	Media
HU10	Escribir resumen ejecutivo de hallazgos clave para perfil no técnico	4 h	Alta
HU10	Validar consistencia de resultados con al menos un reentrenamiento completo del pipeline	6 h	Alta
HU10	Consolidar todas las visualizaciones y tablas en formato compatible con el trabajo final	6 h	Alta
HU10	Revisión cruzada de los datos usados, modelos entrenados y outputs finales para garantizar trazabilidad	4 h	Alta
HU10 (Complementaria)	Crear checklist de calidad para evaluación reproducible del proyecto	4 h	Media
HU10 (Opcional)	Preparar versión resumida tipo “executive deck” para stakeholders empresariales	4 h	Baja
HU11	Redactar documento técnico sobre el pipeline de preprocesamiento (paso a paso)	4 h	Alta
HU11	Incluir justificación de decisiones de limpieza, codificación y normalización de variables	4 h	Alta
HU11	Documentar criterios para selección de variables y eliminación de atributos redundantes	4 h	Alta
HU11	Crear diagrama de flujo del pipeline de datos para incluir en el informe técnico	4 h	Media
HU12	Configurar un repositorio de control de versiones (Git) con estructura organizada por módulos	4 h	Alta
HU12	Registrar versiones clave de scripts de modelado y preprocesamiento (commits etiquetados)	4 h	Media
HU12	Estandarizar nombre de archivos y estructuras de carpetas para reproducibilidad	3 h	Media
HU12	Documentar cada experimento de entrenamiento en un log estructurado (fecha, modelo, métricas)	3 h	Media
HU13	Redactar sección de introducción y justificación del proyecto	4 h	Alta
HU13	Redactar el estado del arte y antecedentes técnicos con citas académicas	5 h	Alta

Historia de usuario	Tarea técnica	Estimación	Prioridad
HU13	Describir la metodología, con el enfoque CRISP-DM y diseño experimental	4 h	Alta
HU13	Redactar sección de resultados con métricas, gráficas y análisis comparativo	4 h	Alta
HU13	Redactar las conclusiones, recomendaciones y posibles líneas futuras de trabajo	3 h	Alta
HU14	Adaptar el documento al formato oficial del posgrado (estructura, márgenes, tipografía)	3 h	Alta
HU14	Revisar y corregir estilo, ortografía y redacción académica del documento completo	4 h	Alta
HU14	Incorporar referencias en formato académico (BibTeX o APA) y verificar su consistencia	3 h	Alta
HU15	Diseñar el esquema de la presentación (estructura narrativa y bloques temáticos)	3 h	Alta
HU15	Crear diapositivas para la introducción, problema y contexto del proyecto	4 h	Alta
HU15	Elaborar visualizaciones para explicar la metodología y modelos aplicados	4 h	Alta
HU15	Incluir resultados clave, métricas, interpretabilidad y conclusiones en formato visual	4 h	Alta
HU15	Revisar estilo gráfico, legibilidad, formato y duración estimada de la presentación	2 h	Alta
HU16	Preparar guion detallado para la exposición oral (con tiempos por sección)	3 h	Media
HU16	Realizar al menos dos ensayos de defensa con cronómetro y grabación propia	4 h	Media
HU16	Practicar respuestas a posibles preguntas técnicas y de evaluación crítica	3 h	Media
HU16	Ajustar discurso, transiciones y tiempos en función de los ensayos	3 h	Media

10. Diagrama de Gantt

El diagrama de Gantt debe representar de forma visual y cronológica todas las tareas del proyecto, abarcando aproximadamente 600 horas totales, de las cuales entre 480 y 500 deben destinarse a tareas técnicas (desarrollo, pruebas, implementación) y entre 100 y 120 a tareas no técnicas (planificación, documentación, escritura de memoria y preparación de la defensa).

Consignas y recomendaciones:

- Incluir tanto tareas técnicas derivadas de las HU como tareas no técnicas generales del proyecto.
- El eje vertical debe listar las tareas y el eje horizontal representar el tiempo en semanas o fechas.
- Utilizar colores diferenciados para distinguir tareas técnicas y no técnicas.
- Las tareas deben estar ordenadas cronológicamente y reflejar todo el ciclo del proyecto.
- Iniciar con la planificación del proyecto (coincidente con el inicio de Gestión de Proyectos) y finalizar con la defensa, próxima a la fecha de cierre del trabajo.
- Configurar el software para mostrar los códigos del desglose de tareas y los nombres junto a cada barra.
- Asegurarse de que la fecha final coincida con la del Acta Constitutiva.
- Evitar tareas genéricas o ambiguas y asegurar una secuencia lógica y realista.
- Las fechas pueden ser aproximadas; ajustar el ancho del diagrama según el texto y el parámetro `x unit`. Para mejorar la apariencia del diagrama, es necesario ajustar este valor y, quizás, acortar los nombres de las tareas.

Herramientas sugeridas:

- Planner, GanttProject, Trello + plugins
<https://blog.trello.com/es/diagrama-de-gantt-de-un-proyecto>
- Creately (colaborativa online)
<https://creately.com/diagram/example/ieb3p3ml/LaTeX>
- LaTeX con pgfgantt:
<http://ctan.dcc.uchile.cl/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf>

Incluir una imagen legible del diagrama de Gantt. Si es muy ancho, presentar primero la tabla y luego el gráfico de barras.

11. Planificación de Sprints

Organizar las tareas técnicas del proyecto en sprints de trabajo que permitan distribuir de forma equilibrada la carga horaria total, estimada en 600 horas.

Consigna:

- Completar una tabla que relacione sprints con HU y tareas técnicas correspondientes.
- Incluir estimación en horas para cada tarea.
- Indicar responsable y porcentaje de avance estimado o completado.
- Contemplar también tareas de planificación, documentación, redacción de memoria y preparación de defensa.

Conceptos clave:

- Una épica es una unidad funcional amplia; una historia de usuario es una funcionalidad concreta; un sprint es una unidad de tiempo donde se ejecutan tareas.
- Las tareas son el nivel más desagregado: permiten estimar tiempos, asignar responsables y monitorear progreso.

Duración sugerida:

- Para un proyecto de 600 h, se recomienda planificar entre 10 y 12 sprints de aproximadamente 2 semanas cada uno.
- Asignar entre 45 y 50 horas efectivas por sprint a tareas técnicas.
- Reservar 100 a 120 h para actividades no técnicas (planificación, escritura, reuniones, defensa).

Importante:

- En proyectos individuales, el responsable suele ser el propio autor.
- Aun así, desagregar tareas facilita el seguimiento y mejora continua.

Conversión opcional de Story Points a horas:

- 1 SP ≈ 2 h como referencia flexible.
- Tener en cuenta aproximaciones tipo Fibonacci.

Recomendaciones:

- Verificar que la carga horaria por sprint sea equilibrada.
- Usar sprints de 1 a 3 semanas, acordes al cronograma general.
- Actualizar el % completado durante el seguimiento del proyecto.
- Considerar un sprint final exclusivo para pruebas, revisión y ajustes antes de la defensa.

Cuadro 1. Formato sugerido

Sprint	HU o fase	Tarea	Horas / SP	Responsable	% Completado
Sprint 0	Planificación	Definir alcance y cronograma	10 h	Alumno	100 %
Sprint 0	Planificación	Reunión con tutor/cliente	5 h	Alumno	50 %
Sprint 0	Planificación	Ajuste de entregables	6 h	Alumno	25 %
Sprint 1	HU1	Tarea 1 HU1	6 h / 3 SP	Alumno	0 %
Sprint 1	HU1	Tarea 2 HU1	10 h / 5 SP	Alumno	0 %
Sprint 2	HU2	Tarea 1 HU2	7 h / 5 SP	Alumno	0 %
...
Sprint 5	Escritura	Redacción memoria	50 h / 34 SP	Alumno	0 %
Sprint 6	Defensa	Preparación exposición	20 h / 13 SP	Alumno	0 %

12. Normativa y cumplimiento de datos (gobernanza)

En esta sección se debe analizar si los datos utilizados en el proyecto están sujetos a normativas de protección de datos y privacidad, y en qué condiciones se pueden emplear.

Aspectos a considerar:

- Evaluar si los datos están regulados por normativas como GDPR, Ley 25.326 de Protección de Datos Personales en Argentina, HIPAA u otras según jurisdicción y temática.
- Determinar si el uso de los datos requiere consentimiento explícito de los usuarios involucrados.
- Indicar si existen restricciones legales, técnicas o contractuales sobre el uso, compartición o publicación de los datos.
- Aclarar si los datos provienen de fuentes licenciadas, de acceso público o bajo algún tipo de autorización especial.
- Analizar la viabilidad del proyecto desde el punto de vista legal y ético, considerando la gobernanza de los datos.

Este análisis es clave para garantizar el cumplimiento normativo y evitar conflictos legales durante el desarrollo y publicación del proyecto.

13. Gestión de riesgos

a) Identificación de los riesgos (al menos cinco) y estimación de sus consecuencias:

Riesgo 1: detallar el riesgo (riesgo es algo que si ocurre altera los planes previstos de forma negativa)

- Severidad (S): mientras más severo, más alto es el número (usar números del 1 al 10).
Justificar el motivo por el cual se asigna determinado número de severidad (S).
- Probabilidad de ocurrencia (O): mientras más probable, más alto es el número (usar del 1 al 10).
Justificar el motivo por el cual se asigna determinado número de (O).

Riesgo 2:

- Severidad (S): X.
Justificación...
- Ocurrencia (O): Y.
Justificación...

Riesgo 3:

- Severidad (S): X.
Justificación...
- Ocurrencia (O): Y.
Justificación...

b) Tabla de gestión de riesgos: (El RPN se calcula como $RPN=S \times O$)

Riesgo	S	O	RPN	S*	O*	RPN*

Criterio adoptado:

Se tomarán medidas de mitigación en los riesgos cuyos números de RPN sean mayores a...

Nota: los valores marcados con (*) en la tabla corresponden luego de haber aplicado la mitigación.

c) Plan de mitigación de los riesgos que originalmente excedían el RPN máximo establecido:

Riesgo 1: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación). Nueva asignación de S y O, con su respectiva justificación:

- Severidad (S*): mientras más severo, más alto es el número (usar números del 1 al 10).
Justificar el motivo por el cual se asigna determinado número de severidad (S).
- Probabilidad de ocurrencia (O*): mientras más probable, más alto es el número (usar del 1 al 10). Justificar el motivo por el cual se asigna determinado número de (O).

Riesgo 2: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación).

Riesgo 3: plan de mitigación (si por el RPN fuera necesario elaborar un plan de mitigación).

14. Sprint Review

La revisión de sprint (*Sprint Review*) es una práctica fundamental en metodologías ágiles. Consiste en revisar y evaluar lo que se ha completado al finalizar un sprint. En esta instancia, se presentan los avances y se verifica si las funcionalidades cumplen con los criterios de aceptación establecidos. También se identifican entregables parciales y se consideran ajustes si es necesario.

Aunque el proyecto aún se encuentre en etapa de planificación, esta sección permite proyectar cómo se evaluarán las funcionalidades más importantes del backlog. Esta mirada anticipada favorece la planificación enfocada en valor y permite reflexionar sobre posibles obstáculos.

Objetivo: anticipar cómo se evaluará el avance del proyecto a medida que se desarrollen las funcionalidades, utilizando como base al menos cuatro historias de usuario del *Product Backlog*.

Seleccionar al menos 4 HU del Product Backlog. Para cada una, completar la siguiente tabla de revisión proyectada:

Formato sugerido:

HU seleccionada	Tareas asociadas	Entregable esperado	¿Cómo sabrás que está cumplida?	Observaciones o riesgos
HU1	Tarea 1	Módulo funcional	Cumple criterios de aceptación definidos	Falta validar con el tutor
	Tarea 2			
HU3	Tarea 1	Reporte generado	Exportación disponible y clara	Requiere datos reales
	Tarea 2			
HU5	Tarea 1	Panel de gestión	Roles diferenciados operativos	Riesgo en integración
	Tarea 2			
HU7	Tarea 1	Informe trimestral	PDF con gráficos y evolución	Puede faltar tiempo para ajustes
	Tarea 2			

15. Sprint Retrospective

La retrospectiva de sprint es una práctica orientada a la mejora continua. Al finalizar un sprint, el equipo (o el alumno, si trabaja de forma individual) reflexiona sobre lo que funcionó bien, lo que puede mejorarse y qué acciones concretas pueden implementarse para trabajar mejor en el futuro.

Durante la cursada se propuso el uso de la **Estrella de la Retrospectiva**, que organiza la reflexión en torno a cinco ejes:

- ¿Qué hacer más?
- ¿Qué hacer menos?

- ¿Qué mantener?
- ¿Qué empezar a hacer?
- ¿Qué dejar de hacer?

Aun en una etapa temprana, esta herramienta permite que el alumno planifique su forma de trabajar, identifique anticipadamente posibles dificultades y diseñe estrategias de organización personal.

Objetivo: reflexionar sobre las condiciones iniciales del proyecto, identificando fortalezas, posibles dificultades y estrategias de mejora, incluso antes del inicio del desarrollo.

Completar la siguiente tabla tomando como referencia los cinco ejes de la Estrella de la Retrospectiva (*Starfish* o estrella de mar). Esta instancia te ayudará a definir buenas prácticas desde el inicio y prepararte para enfrentar el trabajo de forma organizada y flexible. Se deberá completar la tabla al menos para 3 sprints técnicos y 1 no técnico.

Formato sugerido:

Sprint tipo y N°	¿Qué hacer más?	¿Qué hacer menos?	¿Qué mantener?	¿Qué empezar a hacer?	¿Qué dejar de hacer?
Sprint técnico - 1	Validaciones continuas con el alumno	Cambios sin versión registrada	Pruebas con datos simulados	Documentar cambios propuestos	Ajustes sin análisis de impacto
Sprint técnico - 2	Verificar configuraciones en múltiples escenarios	Modificar parámetros sin guardar historial	Perfiles reutilizables	Usar logs para configuración	Repetir pruebas manuales innecesarias
Sprint técnico - 8	Comparar correlaciones con casos previos	Cambiar parámetros sin justificar	Revisión cruzada de métricas	Anotar configuraciones usadas	Trabajar sin respaldo de datos
Sprint no técnico - 12 (por ej.: “Defensa”)	Ensayos orales con feedback	Cambiar contenidos en la memoria	Material visual claro	Dividir la presentación por bloques	Agregar gráficos difíciles de explicar