
# 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)

*Esta planificación fue realizada en el curso de Gestión de proyectos entre el 29 de abril de 2025 y el 17 de junio de 2025.*

# Índice

## 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 a definir.

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, 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 líder en la gestión de préstamos para la construcción. Trabaja con múltiples entidades bancarias y financieras que otorgan financiamiento a personas que planean construír su casa, 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 en 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 revisiones manuales 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 las experiencias pasadas, sin una herramienta sistemática que permita 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, techado, acabados). 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

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 historial 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 exclusivamente a información estructurada, sin el uso de documentos escaneados o imágenes.

## Estado del arte y diferenciación de la solución

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 basándonos 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 unitarios, bases de datos estáticas por región o *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.

La regresión multivariada ha sido un enfoque base 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 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 otro tipo 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 utiliza *XGBoost* para seleccionar las variables más influyentes y estimar el costo de proyectos de edificación.

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.

Al leer los resultados de dichos proyectos, se puede apreciar que las variables de mayor influencia son la escala del proyecto (metros cuadrados construídos, número de plantas), funcionalidad (edificio residencial, comercial, salud, etc.), así como especificaciones técnicas principales (tipo de climatización, sistema estructural, acaados 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 en que:

- Utiliza modelos de aprendizaje supervisado entrenados sobre datos reales de ejecución y financiamiento, con ajuste de las predicciones al comportamiento histórico.

- Integra múltiples variables categóricas y numéricas, como ubicación geográfica, superficie construida, tipo de contratista y composición del presupuesto, con una predicción contextualizada.
- Utiliza un volumen de datos mayor a los proyectos realizados hasta el momento.
- Incorpora técnicas de explicabilidad como SHAP (Shapley Additive Explanations) para interpretar por qué el modelo predice insuficiencia o sobrecosto, con adopción por parte de analistas humanos.
- Ofrece tanto una salida binaria (¿es suficiente el préstamo?) como una salida regresiva multirrubro (estimación de costos por partida).

### Propuesta de valor e impacto esperado

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

- Reducir el porcentaje de proyectos con préstamos insuficientes y así lograr evitar renegociaciones y sobrecostos.
- Mejorar la eficiencia del análisis crediticio, donde los analistas cuentan con una herramienta predictiva basada en datos.
- Aumentar la satisfacción del cliente final, sin interrupciones en la ejecución del proyecto por errores de estimación.
- Detectar patrones de subestimación crónica en ciertos rubros o regiones, lo que podría informar futuras políticas de originación de préstamos.

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.

### Descripción funcional de la solución

La solución se compone de los siguientes bloques funcionales:

- **Módulo de extracción de datos**: identificación, scrapping y unificación de datos.
- **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 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 visualización**: presenta los resultados a través de gráficos, tablas y explicaciones interpretables para el uso por parte del equipo financiero.

En la figura 1 se presenta el diagrama en bloques del sistema. Se observa el flujo de datos desde la base histórica, las etapas de preprocesamiento, entrenamiento e inferencia, hasta la visualización de resultados por parte del analista financiero.

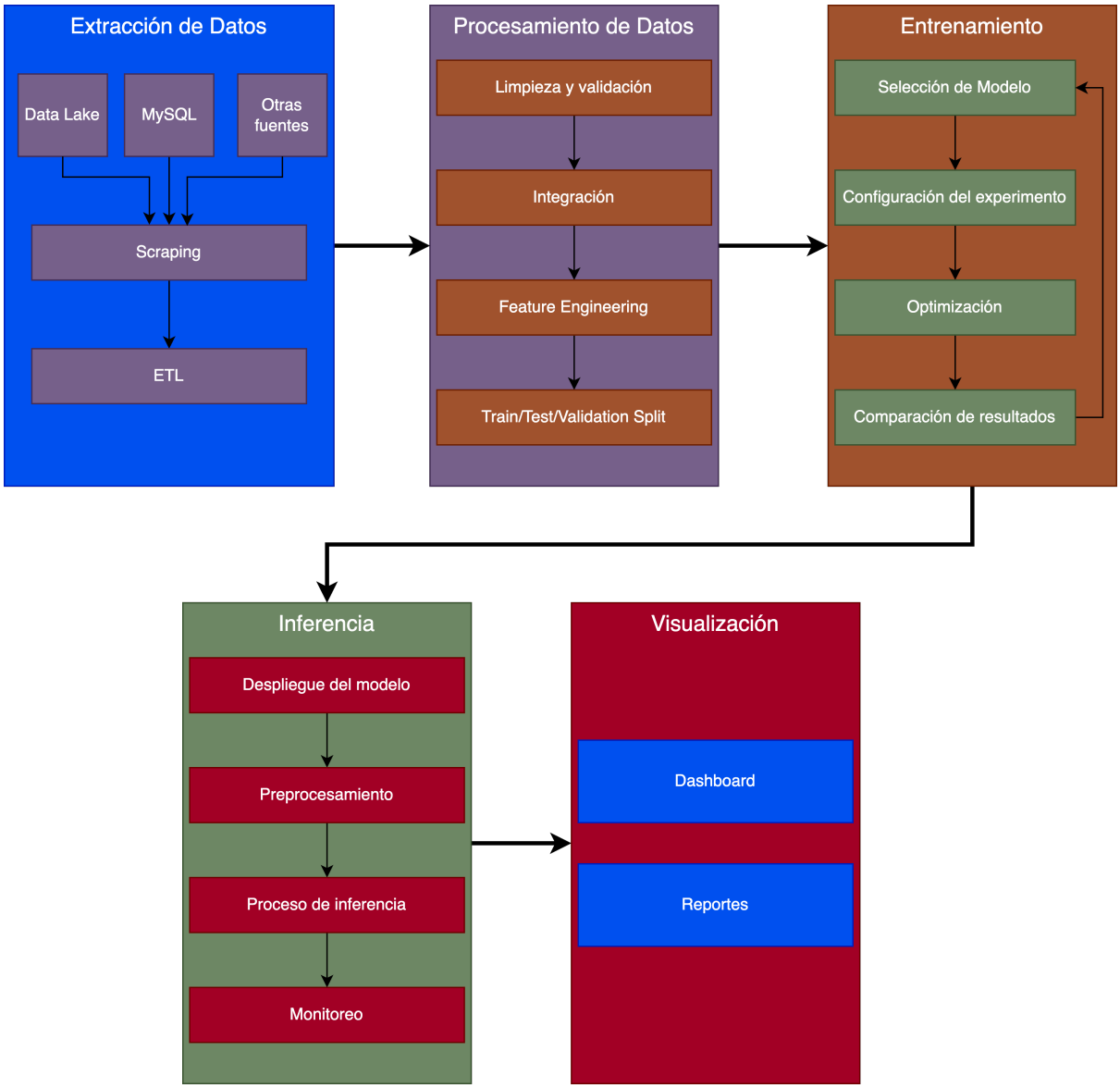


Figura 1. Diagrama en bloques del sistema.

## 2. Identificación y análisis de los interesados

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

| Rol | Nombre y Apellido | Organización | Puesto |
|---|---|---|---|
| Cliente | Martin Brambati | 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 | - | Clientes de Built | - |

- 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 del conjunto de datos.
- 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.

## 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. 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 originación de créditos.

## 4. Alcance del proyecto

El presente proyecto incluye:

- Relevamiento y comprensión del problema de negocio asociado a la evaluación de suficiencia de préstamos y estimación presupuestaria en proyectos de construcción.
- Exploración, limpieza y transformación del conjunto de datos históricos anonimizados proporcionado por Built Technologies.
- Desarrollo de un pipeline de preprocesamiento de datos, que incluirá:
  - Tratamiento de valores faltantes.
  - Codificación de variables categóricas.
  - Normalización y estandarización de variables numéricas.
  - Balanceo de clases en caso de desbalance significativo en la variable objetivo (para el modelo de clasificación).
- Entrenamiento y validación de:

  - Un modelo de clasificación binaria para predecir si el préstamo solicitado es suficiente para cubrir el presupuesto total.
  - Un modelo de regresión multivariada para estimar los costos esperados por categoría presupuestaria (por ejemplo: cimentación, electricidad, techado, etc.).
- Evaluación de los modelos desarrollados mediante métricas adecuadas (*F1-score*, *AUC-ROC*, *MAE*, *RMSE*).
- Generación de explicaciones interpretables de las predicciones con técnicas como *SHAP* o *Feature Importance*.
- Documentación del proceso completo y presentación de los resultados en un formato replicable y académico.

El presente proyecto no incluye:

- El desarrollo de una interfaz gráfica de usuario.
- La integración de los modelos desarrollados en los sistemas de producción de Built Technologies.
- La automatización completa del pipeline en un entorno de producción.
- La toma de decisiones finales sobre políticas de crédito, las cuales quedarán en manos del equipo financiero de la empresa.
- El análisis de documentos no estructurados.

## 5. Supuestos del proyecto

Para el desarrollo del presente proyecto se supone que:

- 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 requerirá solicitar acceso a datos sensibles o confidenciales.
- 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.
- 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.
- 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

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

- 1 punto: tarea simple, conocida, sin incertidumbre técnica.
- 2–3 puntos: tarea con un nivel medio de procesamiento o exploración de datos.
- 5 puntos: tarea técnica de mediana complejidad o con validaciones múltiples.
- 8+ puntos: tarea compleja o con alto nivel de incertidumbre en datos o rendimiento del modelo.

A continuación, se detallan las épicas y sus respectivas 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.
    **Prioridad: Alta — Story Points: 3**
  - **HU2:** Como estudiante, quiero actualizar y ajustar el backlog y el plan de trabajo según avances reales y obstáculos encontrados.
    **Prioridad: Media — Story Points: 2**
- **Épica 2: relevamiento y análisis de datos históricos**
  - **HU3:** 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**
  - **HU4:** 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: preprocesamiento y preparación de datos**
  - **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**
  - **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: entrenamiento y validación de modelos**
  - **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*.
    **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*.
    **Prioridad: Alta — Story Points: 5**

- **Épica 5: interpretación, documentación y validación**
  - **HU9:** Como analista, quiero aplicar técnicas de interpretabilidad (*SHAP*, *Feature Importance*) para entender las variables que más influyen en las predicciones.
    **Prioridad: Media** — **Story Points: 3**
  - **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**
- **Épica 6: sistematización y documentación técnica**
  - **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**
  - **HU12:** Como desarrollador, quiero versionar el código y registrar los experimentos de modelado para facilitar replicabilidad futura.
    **Prioridad: Media** — **Story Points: 3**
- **Épica 7: redacción del trabajo final**
  - **HU13:** Como alumno de posgrado, quiero redactar la memoria escrita del trabajo final, con antecedentes, metodología, resultados y conclusiones.
    **Prioridad: Alta** — **Story Points: 8**
  - **HU14:** Como autor, quiero adaptar el documento a los lineamientos formales de la universidad para asegurar su correcta presentación.
    **Prioridad: Alta** — 8bfStory Points: 3
- **Épica 8: preparación de la defensa oral**
  - **HU15:** Como expositor, quiero preparar una presentación clara y visual para explicar el problema, 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.
    **Prioridad: Media** — **Story Points: 3**

## 7. Criterios de aceptación de historias de usuario

- **Épica 1: planificación y organización del proyecto**
  - **Criterios de aceptación HU1**
    - Se ha elaborado un cronograma tentativo con fases alineadas al backlog.
    - El cronograma incluye tiempos estimados por tarea y asignación tentativa por sprint.
    - El plan ha sido validado con el orientador del proyecto.
  - **Criterios de aceptación HU2**
    - Se han identificado los ajustes realizados al backlog o tareas planificadas.
    - Los cambios se encuentran justificados en base a imprevistos o progresos.
    - El backlog actualizado ha sido revisado al menos en dos hitos importantes.

- **Épica 2: relevamiento y análisis de datos históricos**
  - **Criterios de aceptación HU3**
    - El conjunto de datos ha sido cargado correctamente en el entorno de trabajo.
    - Se identificaron y eliminaron valores atípicos y registros duplicados.
    - Se generó un informe exploratorio con estadísticas descriptivas y visualizaciones básicas.
  - **Criterios de aceptación HU4**
    - Se ha identificado el conjunto de variables relevantes para el modelo.
    - Las variables han sido clasificadas como numéricas o categóricas.
    - Se documentó la justificación técnica de la selección de cada variable.
- **Épica 3: preprocesamiento y preparación de datos**
  - **Criterios de aceptación HU5**
    - 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*).
    - Se verificó que el modelo no pierda precisión tras aplicar balanceo.
  - **Criterios de aceptación HU6**
    - 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.
- **Épica 4: entrenamiento y validación de modelos**
  - **Criterios de aceptación HU7**
    - El modelo de clasificación fue entrenado con un conjunto dividido en entrenamiento, validación y prueba.
    - 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.
  - **Criterios de aceptación HU8**
    - Se ha entrenado un modelo de regresión con variables de entrada preprocesadas.
    - Se calculan métricas $MAE$, $RMSE$ y $R^2$ sobre el conjunto de prueba.
    - El modelo logra un $MAE$ dentro del umbral definido según el negocio.
- **Épica 5: interpretación, documentación y validación**
  - **Criterios de aceptación HU9**
    - Se aplicó una técnica de explicabilidad (ej. $SHAP$) sobre los modelos entrenados.
    - Se identificaron las variables más influyentes en las predicciones.
    - Los resultados de interpretabilidad se incluyen en un informe gráfico y textual.
  - **Criterios de aceptación HU10**
    - Se generó un documento técnico con los pasos realizados, decisiones tomadas y resultados.
    - El documento incluye gráficas de métricas y análisis de errores.
- **Épica 6: sistematización y documentación técnica**

- **Criterios de aceptación HU11**
  - Las decisiones de preprocesamiento fueron registradas en un documento o notebook.
  - Las transformaciones realizadas son reproducibles en otros entornos.
  - El documento técnico describe el flujo completo de tratamiento de datos.
- **Criterios de aceptación HU12**
  - Se utilizó control de versiones (Git) para registrar avances del proyecto.
  - Los experimentos de modelado se guardaron en forma estructurada (scripts, parámetros).
  - La replicabilidad fue validada con el pipeline en una sesión independiente.

- **Épica 7: Redacción del trabajo final**
  - **Criterios de aceptación HU13**
    - Se redactó un borrador completo de la memoria con introducción, metodología, resultados y conclusiones.
    - Se incluyeron tablas, gráficos y citas académicas relevantes.
    - La versión final fue revisada y corregida en base a retroalimentación del orientador.
  - **Criterios de aceptación HU14**
    - El documento fue ajustado al formato requerido por la universidad.
    - Se verificó cumplimiento de normas de estilo, ortografía y citas bibliográficas.
    - El archivo final está listo para ser presentado ante la coordinación académica.
- **Épica 8: Preparación de la defensa oral**
  - **Criterios de aceptación HU15**
    - Se elaboró una presentación visual con estructura clara (problema, solución, resultados).
    - Se incorporaron visualizaciones clave del modelo y sus métricas.
    - La presentación cumple con la duración y estilo requeridos para la defensa.
  - **Criterios de aceptación HU16**
    - 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.

## 8. Fases de CRISP-DM

1. **Comprensión del negocio:** El objetivo principal del proyecto es asistir a analistas financieros de Built Technologies mediante modelos predictivos que permitan:
   - Estimar si el préstamo otorgado será suficiente para cubrir el presupuesto total de un proyecto de construcción.
   - Predecir el costo esperado por categoría presupuestaria (materiales, mano de obra, cimentación, etc.).

El valor agregado de incorporar IA radica en anticipar 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 $R^2$ para regresión), interpretabilidad y aplicabilidad práctica para el analista.

2. **Comprensión de los datos:** El dataset proviene de registros históricos anonimizados de Built Technologies, con más de 10 años de datos sobre proyectos financiados.
   - **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.
   - **Calidad:** excelente, pero con presencia de valores faltantes, codificación inconsistente de categorías y posibles *outliers*.
3. **Preparación de los datos:** Esta etapa incluye limpieza y transformación del *dataset* para su uso por los modelos:
   - Eliminación de registros con errores evidentes o duplicados.
   - Tratamiento de valores nulos mediante imputación o eliminación.
   - Codificación de variables categóricas (*one-hot encoding*).
   - Normalización y estandarización de variables numéricas.
   - Balanceo de clases para la variable objetivo en el modelo de clasificación.
   - Selección de variables relevantes mediante análisis exploratorio y técnicas automáticas (*feature importance*, selección recursiva).
4. **Modelado:** Se abordan dos problemas distintos:
   - **Probabilidad de completitud:** predecir si el préstamo será suficiente para cubrir los costos.
   - **Regresión multivariada:** estimar el costo por partida presupuestaria.

   Los algoritmos candidatos incluyen:
   - **Para clasificación:** *Random Forest*, *XGBoost*, Regresión Logística.
   - **Para regresión:** *XGBoost Regressor*, *LightGBM*, redes neuronales densas.

   Se compararán diferentes modelos y se realizará validación cruzada.
5. **Evaluación del modelo:** Se utilizarán diferentes métricas de rendimiento por tipo de modelo:
   - **Clasificación:** *F1-score*, precisión, *recall*, *AUC-ROC*.
   - **Regresión:** *MAE* (error absoluto medio), *RMSE* (raíz del error cuadrático medio), $R^2$.

   Además, 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:
   - Exportación del modelo entrenado en formato compatible (.pkl, .joblib).
   - Sugerencia de integración vía API REST o servicio batch interno.
   - Propuesta de visualización simple para interpretación de resultados.

## 9. Desglose del trabajo en tareas

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.

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.

| 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 |
| HU2 | Ajustar cronograma o redistribuir tareas en función del progreso real (replanificación) | 4 h | Media |
| HU2 | Documentar los cambios realizados sobre el plan y justificar desviaciones | 2 h | Media |
| HU3 | Importar el dataset en entorno de trabajo (Python) y validar estructura general | 4 h | Alta |
| HU3 | Identificar y cuantificar valores nulos, inconsistentes o duplicados | 4 h | Alta |
| HU3 | Realizar limpieza inicial: imputación, eliminación de duplicados y errores obvios | 6 h | Alta |
| HU3 | Analizar la distribución de variables numéricas mediante histogramas y boxplots | 6 h | Alta |
| HU3 | 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 |

| Historia de usuario | Tarea técnica | Estimación | Prioridad |
|---|---|---|---|
| HU4 | Clasificar variables por tipo (numéricas, categóricas, temporales, target) | 4 h | Alta |
| HU4 | Evaluar correlaciones entre variables numéricas y redundancia (matriz de correlación) | 6 h | Alta |
| HU4 | Analizar cardinalidad de variables categóricas y detectar categorías raras o desbalanceadas | 4 h | Media |
| HU4 | Realizar análisis de varianza (ANOVA) o chi-cuadrado para evaluar importancia de variables | 6 h | Alta |
| HU4 | Generar ranking de variables relevantes con técnicas automáticas (feature importance) | 6 h | Alta |
| HU4 | Redactar informe técnico con las variables seleccionadas, criterios y limitaciones | 6 h | Alta |
| HU4 (Opcional) | Aplicar técnicas de reducción de dimensionalidad (PCA o UMAP) y evaluar impacto | 6 h | Baja |
| HU4 (Complementaria) | Visualizar relaciones multivariadas mediante pairplots o mapas de calor | 4 h | Media |
| HU4 (Opcional) | Realizar clustering exploratorio para entender segmentos de proyectos | 6 h | Baja |
| HU5 | Analizar distribución de la variable objetivo y su desbalance (visual y cuantitativa) | 4 h | Media |
| HU5 | Implementar técnica de balanceo simple (undersampling o oversampling clásico) | 4 h | Media |
| HU5 | Evaluar impacto del balanceo sobre la distribución de variables | 4 h | Media |
| HU5 | Implementar SMOTE y variantes avanzadas (BorderlineSMOTE, ADASYN) | 6 h | Alta |
| HU5 | Comparar modelos entrenados con y sin balanceo y medir impacto en F1-score | 6 h | Alta |
| HU5 | Documentar estrategia de balanceo adoptada y razones de su elección | 4 h | Alta |
| HU5 | Validar el pipeline de balanceo con conjuntos de validación cruzada | 4 h | Alta |
| HU5 (Complementaria) | Probar técnicas de balanceo basadas en generación de ruido o augmentación sintética | 4 h | Baja |
| 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 |
| HU6 | 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 preprocesamiento (‘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 |
| HU6 (Opcional) | Probar codificación target encoding y evaluar sobreajuste con K-fold | 6 h | Baja |
| HU6 (Complementaria) | Desarrollar funciones propias reutilizables para codificación y escalamiento | 6 h | Media |
| HU6 (Opcional) | Generar versiones comprimidas del dataset para ejecución más rápida de modelos | 4 h | Baja |
| HU7 | Seleccionar y justificar algoritmos candidatos para clasificación (p. ej., RF, XGBoost) | 4 h | Alta |
| HU7 | Implementar modelo base de clasificación y entrenarlo con datos preprocesados | 6 h | Alta |
| HU7 | Realizar validación cruzada (k-fold) y medir F1-score y AUC promedio | 6 h | Alta |
| HU7 | Ajustar hiperparámetros (grid search o random search) y registrar mejoras | 6 h | Alta |
| 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 | Evaluar 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) | Evaluar 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 | Evaluar 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

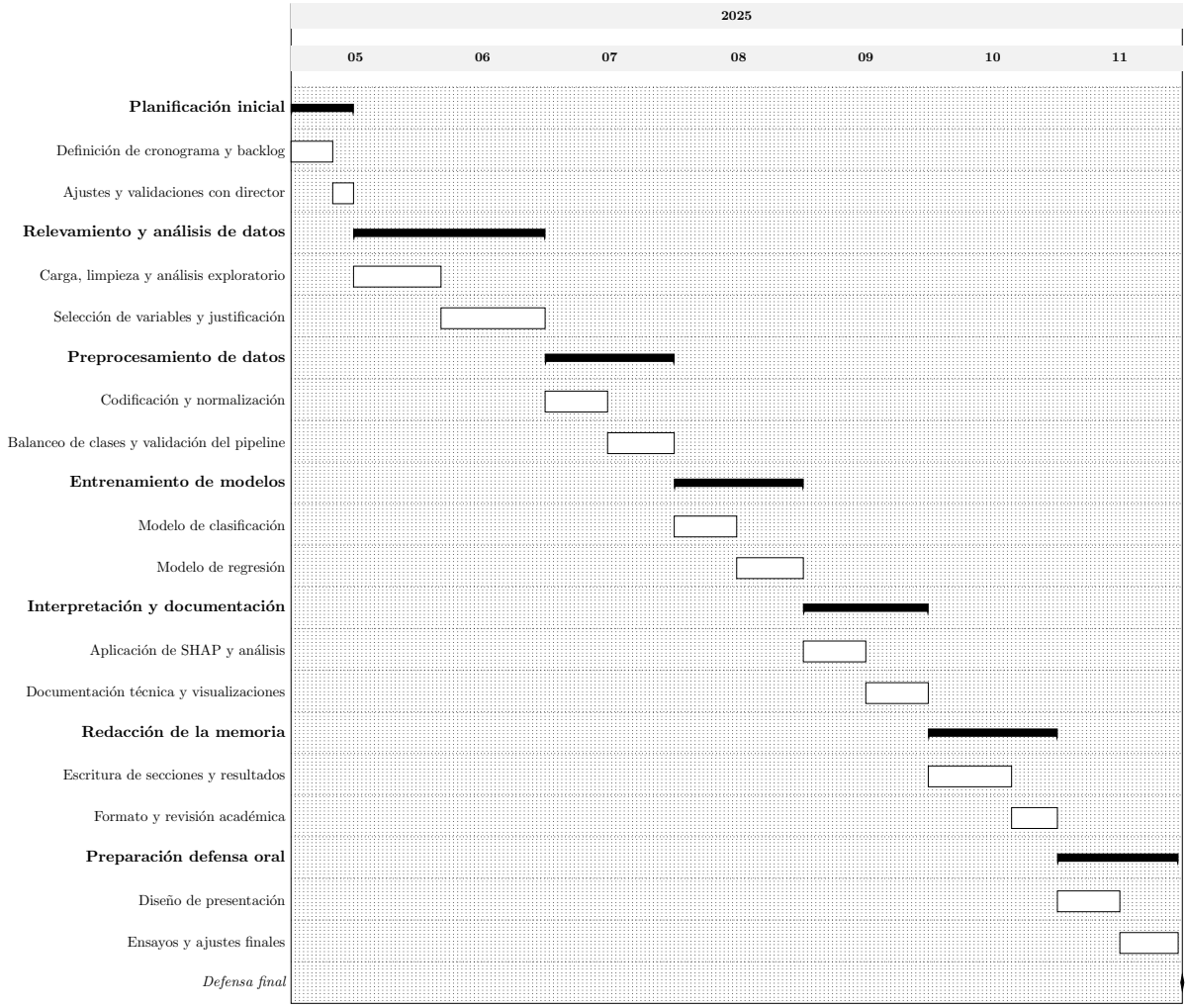


Figura 2. Diagrama de Gantt del proyecto.

## 11. Planificación de Sprints

Cuadro 1. Planificación de sprints del proyecto

| Sprint | HU o fase | Tarea | Horas | Responsable | % Completado |
|---|---|---|---|---|---|
| Sprint 0 | HU1 | Analizar el alcance del proyecto y elaborar una primera versión del cronograma general | 6 h | Alumno | 100 % |
| Sprint 0 | HU1 | Dividir el proyecto en fases e identificar dependencias | 4 h | Alumno | 100 % |
| Sprint 1 | HU1 | Diseñar plan de sprints con hitos y revisión | 6 h | Alumno | 100 % |
| Sprint 1 | HU1 | Revisar cronograma con director y ajustar plan | 2 h | Alumno | 100 % |
| Sprint 1 | HU2 | Establecer rutina de seguimiento | 2 h | Alumno | 100 % |
| Sprint 2 | HU2 | Actualizar backlog según bloqueos técnicos | 4 h | Alumno | 100 % |
| Sprint 2 | HU2 | Ajustar cronograma según progreso real | 4 h | Alumno | 100 % |
| Sprint 2 | HU2 | Documentar cambios en el plan | 2 h | Alumno | 100 % |
| Sprint 3 | HU3 | Importar y validar dataset en entorno Python | 4 h | Alumno | 100 % |
| Sprint 3 | HU3 | Identificar valores nulos y duplicados | 4 h | Alumno | 100 % |
| Sprint 4 | HU3 | Limpieza inicial (imputación, duplicados) | 6 h | Alumno | 75 % |
| Sprint 4 | HU3 | Análisis de variables con boxplots e histogramas | 6 h | Alumno | 50 % |
| Sprint 5 | HU3 | Detección y tratamiento de outliers | 6 h | Alumno | 0 % |
| Sprint 5 | HU3 | Generación de informe exploratorio | 6 h | Alumno | 0 % |
| Sprint 6 | HU4 | Clasificación de variables y correlaciones | 10 h | Alumno | 0 % |
| Sprint 7 | HU4 | ANOVA, chi-cuadrado y ranking de variables | 12 h | Alumno | 0 % |
| Sprint 8 | HU4 | Redacción de informe técnico (variables clave) | 6 h | Alumno | 0 % |
| Sprint 9 | HU5 | Balanceo: análisis, implementación y evaluación | 14 h | Alumno | 0 % |
| Sprint 10 | HU6 | Codificación y escalado de variables | 12 h | Alumno | 0 % |
| Sprint 11 | HU6 | Validación y documentación del pipeline | 10 h | Alumno | 0 % |

## 12. Normativa y cumplimiento de datos (gobernanza)

El presente proyecto hace uso de datos históricos estructurados y anonimizados provistos por Built Technologies, empresa con sede en Estados Unidos. Los datos incluyen información sobre presupuestos de construcción, características de proyectos y atributos de contratistas, sin contener identificadores personales de clientes.

### Regulación aplicable

Dado que Built Technologies opera en el mercado financiero y de tecnología aplicada a la construcción, y almacena datos sensibles en su plataforma, la empresa adhiere a estándares de cumplimiento como:

- **SOC 2**: Requiere políticas estrictas de manejo y almacenamiento de datos, aplicables a todo proveedor de servicios en la nube.
- **CCPA** (California Consumer Privacy Act): Aplica si hay usuarios residentes en California y otorga derechos sobre el uso de sus datos.
- **GDPR** (Reglamento General de Protección de Datos de la UE): En caso de usuarios o clientes europeos, la empresa debe asegurar trazabilidad, anonimización y consentimiento.

### Condiciones de uso

Built Technologies ha autorizado expresamente el uso de datos con las siguientes condiciones:

- Los datos utilizados fueron previamente **anonimizados**, eliminando cualquier identificador personal (nombres, direcciones, emails, etc.).
- No se incluirán documentos escaneados, imágenes, contratos, ni PDFs asociados a identidades específicas.
- El dataset se empleará únicamente con fines de investigación académica dentro del alcance de este trabajo final.
- No se compartirán datos ni resultados sensibles fuera del ámbito autorizado por la empresa y la universidad.

### Evaluación ética y legal

En base al análisis realizado, se concluye que el uso de los datos:

- Cumple con los lineamientos éticos del posgrado en Inteligencia Artificial.
- No vulnera derechos de privacidad de individuos, dado que el conjunto de datos ha sido previamente anonimizado.
- Se encuentra dentro del marco legal y contractual acordado con la empresa auspiciante.

En consecuencia, el proyecto es viable desde el punto de vista normativo y ético, siempre que se mantenga el tratamiento responsable de la información, sin reidentificación ni difusión pública de casos individuales.

## 13. Gestión de riesgos

### a) Identificación y análisis de riesgos

**Riesgo 1: Demora en la entrega del dataset anonimizado por parte de la empresa.**

- Severidad (S): 8
  Sin datos, el proyecto no puede avanzar. Afecta el desarrollo desde el comienzo.
- Probabilidad de ocurrencia (O): 5
  Es poco probable, pero podría retrasarse por razones administrativas o legales.

**Riesgo 2: Problemas técnicos con el volumen y la calidad de los datos históricos.**

- Severidad (S): 7
  Un dataset mal estructurado o con errores comprometería los resultados del modelo.
- Probabilidad de ocurrencia (O): 6
  Los datos son reales y extensos, lo que puede implicar inconsistencias.

**Riesgo 3: Falta de disponibilidad del director o colaboradores para revisión crítica.**

- Severidad (S): 6
  La falta de feedback oportuno puede generar retrasos o errores en la orientación.
- Probabilidad de ocurrencia (O): 6
  Las agendas laborales y académicas suelen tener conflictos.

**Riesgo 4: Subestimación del tiempo necesario para tareas de modelado.**

- Severidad (S): 6
  Podría comprometer la calidad del modelo o afectar los tiempos de entrega.
- Probabilidad de ocurrencia (O): 7
  Es común en proyectos de IA, especialmente en etapas iniciales.

**Riesgo 5: Dificultad para explicar el modelo a un público no técnico.**

- Severidad (S): 5
  Afecta la comprensión por parte del cliente o del jurado.
- Probabilidad de ocurrencia (O): 6
  Puede suceder si la explicabilidad del modelo no es adecuadamente planificada.

| Riesgo | S | O | RPN | S* | O* | RPN* |
|---|---|---|---|---|---|---|
| Demora en la entrega de datos | 8 | 5 | 40 | 6 | 3 | 18 |
| Calidad técnica del dataset | 7 | 6 | 42 | 6 | 4 | 24 |
| Falta de disponibilidad del director | 6 | 6 | 36 | 5 | 3 | 15 |
| Subestimación del esfuerzo de modelado | 6 | 7 | 42 | 5 | 5 | 25 |
| Dificultad en comunicar resultados | 5 | 6 | 30 | 4 | 4 | 16 |

### b) Tabla de gestión de riesgos

**Criterio adoptado:** Se tomarán medidas de mitigación para los riesgos con RPN ¿30.

### c) Plan de mitigación

#### Riesgo 1: Demora en la entrega de datos

- Medida: establecer una fecha límite temprana con la empresa y confirmar por escrito la disponibilidad de datos.
- Severidad (S*): 6 — con tiempo de respuesta acordado, el impacto se reduce.
- Ocurrencia (O*): 3 — con seguimiento frecuente, baja la probabilidad.

#### Riesgo 2: Calidad técnica del dataset

- Medida: planificar una etapa intensiva de validación y limpieza anticipada; preparar scripts reutilizables.
- Severidad (S*): 6 — si se detectan errores temprano, se puede corregir sin impacto crítico.
- Ocurrencia (O*): 4 — disminuye con testeo exploratorio desde el inicio.

#### Riesgo 4: Subestimación del esfuerzo de modelado

- Medida: aplicar una planificación iterativa y flexible con espacio de buffer en los sprints.
- Severidad (S*): 5 — si se gestiona adecuadamente, no afectará la entrega.
- Ocurrencia (O*): 5 — se reduce con tareas bien desglosadas y cronograma realista.

## 14. Sprint Review

| **HU seleccionada** | **Tareas asociadas** | **Entregable esperado** | **¿Cómo sabrás que está cumplida?** | **Observaciones o riesgos** |
|---|---|---|---|---|
| HU1 | Analizar el alcance del proyecto y elaborar cronograma general | Plan de proyecto estructurado | Cronograma validado por el director y usado como base para los sprints | Puede haber demora en feedback del director |
| | Diseñar plan de sprints con hitos y fases | | | |
| HU3 | Importar el dataset y validar estructura | Reporte técnico de EDA y limpieza de datos | Archivo procesado y script reproducible en Jupyter Notebook | Se requiere contar con los datos reales temprano |
| | Realizar limpieza inicial (duplicados, imputación, errores) | | | |
| HU5 | Implementar técnica de balanceo y validar su efecto | Pipeline de datos balanceados y documentado | Se comparan métricas con y sin balanceo, mejora esperada en F1-score | El balanceo puede introducir ruido si no se evalúa bien |
| | Comparar modelos con y sin balanceo | | | |
| HU7 | Implementar modelo de clasificación y validarlo | Modelo entrenado con métricas F1/AUC reportadas | Cumple umbral mínimo de F1-score definido y se documenta el desempeño | Poca capacidad de ajuste si el dataset no es representativo |
| | Ajustar hiperparámetros y documentar desempeño final | | | |

| **Sprint tipo y N°** | **¿Qué hacer más?** | **¿Qué hacer menos?** | **¿Qué mantener?** | **¿Qué empezar a hacer?** | **¿Qué dejar de hacer?** |
|---|---|---|---|---|---|
| Sprint técnico - 1 | Validar estructura y limpieza del dataset | Dejar decisiones sin documentar | Iteraciones con Jupyter Notebook | Registrar versiones de los datasets | Asumir que los datos están limpios |
| Sprint técnico - 2 | Revisión de métricas al entrenar modelos | Cambiar hiperparámetros sin planificación | Reentrenar con validación cruzada | Planificar pruebas antes de modelar | Ejecutar notebooks sin objetivos claros |
| Sprint técnico - 8 | Revisar visualizaciones explicativas | Confiar ciegamente en los valores SHAP | Análisis de interpretabilidad por sección | Relacionar interpretabilidad con métricas de negocio | Incluir visualizaciones irrelevantes |
| Sprint no técnico - 12 (defensa) | Ensayar presentación con feedback externo | Reescribir toda la memoria en la última semana | Uso de estructuras claras para exponer resultados | Dividir presentación por bloques temáticos | Usar terminología técnica sin adaptación al público |