PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN EN CIENCIA DE DATOS

Predicción del riesgo de retraso en actividades a partir de patrones estructurales en cronogramas de proyectos

Autor:

Estalin Andrés Vásquez Torres

Tutor:

Rafael Melgarejo

Programa académico:

Maestría en Sistemas de Información mención en Ciencia de Datos

Fecha de entrega:

16/08/2026

AGRADECIMIENTOS

Expreso mi más sincero agradecimiento a mi madre, por su apoyo incondicional, sus enseñanzas y su constante motivación a lo largo de mi formación académica y profesional.

A mi esposa, por su amor, comprensión y paciencia durante cada etapa de este proceso, así como por brindarme la fortaleza necesaria para alcanzar esta meta.

A mis hijos, quienes han sido mi mayor fuente de inspiración y motivación para seguir creciendo personal y profesionalmente. Este logro también es para ustedes.

Agradezco a mi tutor y a los docentes de la Maestría en Sistemas de Información mención en Ciencia de Datos por su orientación, conocimientos y acompañamiento durante el desarrollo de esta investigación.

Finalmente, expreso mi reconocimiento a la organización que facilitó el acceso a los datos utilizados en este estudio, haciendo posible el desarrollo de la presente investigación.

INDICE

CAPÍTULO 1. INTRODUCCIÓN 1
1.1 Introducción 1
1.2 Objetivos 2
1.2.1 Objetivo general 2
1.2.2 Objetivos específicos 2
1.3 Justificación 2
1.4 Pregunta de investigación 3
CAPÍTULO 2. MARCO TEORICO 4
2.1 Propósito del capítulo 4
2.2 Estrategia de revisión de literatura 4
2.2.1 Bases de datos y fuentes académicas 5
2.2.2 Cadenas de búsqueda 6
2.2.3 Criterios de inclusión y exclusión 7
2.2.4 Proceso de selección de literatura 7
2.2.5 Matriz de extracción y análisis de literatura 8
2.3 Redacción de antecedentes 9
2.3.1 Contexto general del fenómeno 9
2.3.2 Estudios sobre variables estructurales del cronograma 9
2.3.3 Estudios sobre métodos y modelos predictivos 10
2.3.4 Limitaciones de la literatura y necesidad de investigación 10
2.4 Bases conceptuales 11
2.4.1 Ciencia de Datos aplicada al análisis de cronogramas 13
2.4.2 Riesgo de retraso en actividades 13
2.4.3 Machine Learning y análisis predictivo 14
2.4.4 Modelos predictivos 14
2.4.5 Variables estructurales del cronograma 14
2.4.6 Clasificación como problema de aprendizaje supervisado 15
2.4.7 Importancia de las variables en el proceso predictivo 15

2.5 Teorías, modelos y enfoques relacionados 16
2.5.1 Enfoque de análisis de datos aplicado a cronogramas 16
2.5.2 Modelos estadísticos como referencia 16
2.5.3 Enfoque de Machine Learning predictivo 17
2.5.4 Aprendizaje supervisado como enfoque metodológico 17
2.5.5 Modelos de Machine Learning utilizados en la investigación 17
2.5.6 Integración de enfoques 18
2.6 Variables 19
2.6.1 Clasificación de variables 19
2.6.2 Descripción de las variables utilizadas 20
2.6.3 Justificación de las variables seleccionadas 21
2.6.4 Relación de las variables con el modelo analítico 22
2.6.5 Importancia de las variables en la investigación 22
2.7 Brecha de investigación 22
2.7.1 Introducción de la brecha 22
2.7.2 Conocimiento existente 23
2.7.3 Limitaciones identificadas en la literatura 23
2.7.4 Necesidad del estudio 24
2.7.5 Aporte de la investigación 24
2.8 Síntesis del capítulo 25
CAPÍTULO 3. METODOLOGÍA 27
3.1 Enfoque de investigación 27
3.2 Diseño de investigación 27
3.3 Fuente de datos 28
3.4 Unidad de análisis 29
3.5 Variables de estudio 29
3.5.1 Variable dependiente 30
3.5.2 Variables independientes 30
3.5.3 Resumen de variables utilizadas 31

3.5.4 Justificación de las variables seleccionadas 31
3.6 Preparación de datos 32
3.6.1 Consolidación de la información 32
3.6.2 Selección de actividades para el análisis 32
3.6.3 Construcción de la variable objetivo 33
3.6.4 Construcción de variables derivadas 33
3.6.5 Verificación de calidad de datos 34
3.6.6 Tratamiento de valores extremos 34
3.7 Análisis Exploratorio de Datos (EDA) 35
3.7.1 Distribución de la variable objetivo 35
3.7.2 Análisis de las variables predictoras 35
3.7.3 Identificación de valores extremos 36
3.7.4 Relación entre variables y riesgo de retraso 36
3.7.5 Análisis de correlación 36
3.7.6 Conclusiones del análisis exploratorio 37
3.8 Construcción de modelos predictivos 37
3.8.1 Regresión logística 38
3.8.2 Árbol de decisión 38
3.8.3 Random Forest 38
3.8.4 XGBoost 39
3.8.5 Comparación de modelos 39
3.9 Métricas de evaluación 40
3.9.1 Accuracy 40
3.9.2 Precision 40
3.9.3 Recall 40
3.9.4 F1 Score 41
3.9.5 ROC-AUC 41
3.9.6 Matriz de confusión 41
3.9.7 Selección del modelo final 42

CAPÍTULO 4. RESULTADOS 43
4.1 Descripción del conjunto de datos 43
4.2 Resultados del Análisis Exploratorio de Datos 44
4.2.1 Distribución del riesgo de retraso 44
4.2.2 Análisis de la duración de las actividades 45
4.2.3 Duración y riesgo de retraso 46
4.2.4 Número de predecesores y riesgo de retraso 47
4.2.5 Número de sucesores y riesgo de retraso 48
4.2.6 Análisis de correlación 49
4.2.7 Principales hallazgos del análisis exploratorio 50
4.3 Resultados de la Regresión Logística 50
4.3.1 Desempeño general del modelo 50
4.3.2 Matriz de confusión 51
4.3.3 Interpretación de los resultados 52
4.3.4 Conclusión de la Regresión Logística 52
4.4 Resultados del Árbol de Decisión 52
4.4.1 Desempeño general del modelo 53
4.4.2 Interpretación de los resultados 53
4.5 Resultados de Random Forest 53
4.5.1 Desempeño general del modelo 54
4.5.2 Interpretación de los resultados 54
4.6 Resultados de XGBoost 54
4.6.1 XGBoost con umbral 0.50 55
4.6.2 Optimización del umbral de clasificación 55
4.6.3 Interpretación de los resultados 56
4.7 Comparación de modelos 56
4.8 Selección del modelo final 57
4.8.1 Curva ROC del modelo seleccionado 58
4.8.2 Importancia de variables 58

4.8.3 Matriz de confusión del modelo seleccionado 59
4.9 Aplicación práctica del modelo mediante Power BI 61
4.9.1 Predicción de actividades futuras 61
4.9.2 Dashboard de visualización 61
4.9.3 Distribución de niveles de riesgo 62
4.9.4 Análisis por proyecto 63
4.9.5 Utilidad para la gestión de proyectos 64
4.10 Síntesis del capítulo 65
CAPÍTULO 5. DISCUSIÓN 66
5.1 Interpretación general de los resultados 66
5.2 Cumplimiento de la pregunta de investigación 66
5.3 Discusión sobre las variables estructurales utilizadas 67
5.4 Discusión sobre los modelos predictivos 67
5.5 Discusión sobre la optimización del umbral de clasificación 68
5.6 Relación de los resultados con la literatura revisada 69
5.7 Limitaciones de la investigación 70
5.8 Implicaciones prácticas y académicas 72
5.9 Síntesis del capítulo 72
CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES 74
6.1 Conclusiones 74
6.2 Recomendaciones 75
6.3 Síntesis final de la investigación 76
7 REFERENCIAS 77
8 ANEXOS 78
8.1 Anexo A. Dashboard Power BI 78
8.2 Anexo B. Código principal utilizado para el desarrollo del modelo predictivo 78
8.3 Anexo C. Data set anonimizado. 78

Lista de Tablas

Table 1 Bases de datos y fuentes académicas 5
Table 2 Cadenas de búsqueda 6
Table 3 Criterios de inclusión y exclusión 7
Table 4 Selección de literatura 7
Table 5 Matriz de extracción y análisis de literatura 8
Table 6 Bases conceptuales 12
Table 7 Variables 20
Table 8 Resumen de variables utilizadas 31
Table 9 Caracteristicas del conjunto de Datos 44
Table 10 Resultados de la Regresión Logística 51
Table 11 Desempeño general Árbol de Decisión 53
Table 12 Desempeño general Random Forest 54
Table 13 Resultados de XGBoost con umbral 0.5 55
Table 14 Resultados de XGBoost con umbral optimizado 55
Table 15 Comparación de modelos 56
Table 16 Matriz de confusión del modelo seleccionado 60

Lista de Figuras

Ilustración 1 Distribución de actividades retrasadas 45
Ilustración 2 Distribución de duración de actividades 46
Ilustración 3 Boxplot de duración 46
Ilustración 4 Duración según riesgo de retraso 47
Ilustración 5 Predecesores según riesgo de retraso 48
Ilustración 6 Sucesores según riesgo de retraso 48
Ilustración 7 Heatmap de correlaciones 49
Ilustración 8 Matriz de Confusión - Regresión Logística 51
Ilustración 9 Curva ROC - XGBoost 58
Ilustración 10 Importancia de Variables - XGBoost 59
Ilustración 11 Matriz de Confusión Final 60
Ilustración 12 Dashboard Power BI 62
Ilustración 13 Distribución de niveles de riesgo 63
Ilustración 14 Riesgo promedio por proyecto 64

CAPÍTULO 1. INTRODUCCIÓN

1.1 Introducción

La gestión del tiempo constituye uno de los factores más relevantes para el éxito de los proyectos, ya que el cumplimiento de los cronogramas influye directamente en la coordinación de actividades, la utilización de recursos y el logro de los objetivos establecidos. Sin embargo, en la práctica es común que las actividades experimenten desviaciones respecto a las fechas originalmente planificadas, generando retrasos que afectan el desempeño general del proyecto.

Los cronogramas de proyectos contienen una gran cantidad de información relacionada con la estructura de las actividades, sus duraciones y las relaciones de dependencia existentes entre ellas. Esta información representa una valiosa fuente de datos que puede ser utilizada para identificar patrones asociados al comportamiento de las actividades durante la ejecución de los proyectos.

En los últimos años, el desarrollo de técnicas de Ciencia de Datos y Machine Learning ha permitido transformar grandes volúmenes de datos históricos en conocimiento útil para la toma de decisiones. Estas técnicas han demostrado su capacidad para identificar patrones complejos y construir modelos predictivos capaces de anticipar eventos futuros en diferentes áreas de aplicación.

En este contexto, la presente investigación analiza datos históricos provenientes de cronogramas desarrollados en Primavera P6 con el propósito de identificar patrones estructurales asociados al riesgo de retraso en actividades. Para ello, se utilizan variables derivadas de la estructura del cronograma, tales como la duración de las actividades, el número de predecesores y el número de sucesores, con el fin de desarrollar modelos predictivos capaces de estimar la probabilidad de retraso de una actividad.

El estudio adopta un enfoque basado en datos históricos reales y compara diversos modelos de Machine Learning para determinar cuál presenta mejor desempeño en la predicción del riesgo de retraso. De esta manera, se busca contribuir tanto al campo de la Ciencia de Datos aplicada a la gestión de proyectos como al desarrollo de herramientas que permitan una gestión más proactiva del cronograma.

1.2 Objetivos

1.2.1 Objetivo general

Predecir el riesgo de retraso en actividades utilizando técnicas de Ciencia de Datos aplicadas a datos históricos de cronogramas de proyectos.

1.2.2 Objetivos específicos

- Identificar patrones estructurales en cronogramas asociados al riesgo de retraso en actividades.
- Desarrollar modelos predictivos para estimar el riesgo de retraso de actividades a partir de variables estructurales del cronograma.
- Comparar el desempeño de distintos modelos predictivos aplicados al problema de clasificación.
- Evaluar el desempeño de los modelos mediante métricas de clasificación y seleccionar la alternativa más adecuada para la detección de actividades con riesgo de retraso.

1.3 Justificación

Los retrasos en las actividades representan uno de los problemas más frecuentes en la gestión de proyectos, debido a que afectan el cumplimiento de plazos y dificultan el control efectivo del cronograma. Aunque existen diversas herramientas para planificar y monitorear proyectos, gran parte de los esfuerzos de seguimiento se realizan una vez que las desviaciones ya han ocurrido, limitando la capacidad de respuesta de los equipos de planificación.

En este contexto, la aplicación de técnicas de Ciencia de Datos ofrece una oportunidad para utilizar información histórica con el objetivo de anticipar posibles retrasos antes de que estos se materialicen. La posibilidad de identificar actividades con mayor probabilidad de retraso permite priorizar esfuerzos de seguimiento y mejorar la toma de decisiones durante la ejecución del proyecto.

Desde una perspectiva académica, esta investigación contribuye al análisis de cronogramas mediante la aplicación de modelos de Machine Learning sobre datos reales provenientes de proyectos ejecutados. Asimismo, aporta evidencia sobre la utilidad de las

variables estructurales del cronograma como fuente de información para la construcción de modelos predictivos.

Desde una perspectiva práctica, los resultados obtenidos pueden servir como apoyo para la gestión preventiva del cronograma, permitiendo identificar actividades potencialmente problemáticas y facilitando una gestión más proactiva basada en datos.

Adicionalmente, el estudio incorpora una aplicación práctica mediante la utilización de herramientas de visualización que permiten presentar las predicciones de manera comprensible para los responsables de la planificación y el control de proyectos.

1.4 Pregunta de investigación

¿Cómo pueden los modelos de Ciencia de Datos predecir el riesgo de retraso en actividades a partir de patrones estructurales presentes en cronogramas de proyectos?

CAPÍTULO 2. MARCO TEORICO

2.1 Propósito del capítulo

El presente capítulo tiene como propósito fundamentar teóricamente la investigación relacionada con la predicción del riesgo de retraso en actividades a partir de datos históricos de cronogramas de proyectos mediante técnicas de Ciencia de Datos y Machine Learning.

Para ello, se revisan los principales conceptos, enfoques metodológicos, antecedentes de investigación y modelos analíticos relevantes para comprender el fenómeno de estudio. Asimismo, se analizan las variables estructurales presentes en los cronogramas y su relación con el desempeño de las actividades dentro de los proyectos.

El capítulo se estructura en diferentes secciones que permiten establecer una conexión entre el problema de investigación, los objetivos planteados y el enfoque metodológico adoptado. En primer lugar, se describe la estrategia de revisión de literatura utilizada para la identificación y selección de estudios relevantes. Posteriormente, se presentan los antecedentes y bases conceptuales relacionadas con la gestión de cronogramas, el riesgo de retraso, la Ciencia de Datos y el Machine Learning.

De igual manera, se revisan los principales enfoques y modelos utilizados en investigaciones previas para la predicción de retrasos en proyectos, así como las variables empleadas en este tipo de estudios. Finalmente, se identifican las principales brechas existentes en la literatura y se justifica la relevancia de la presente investigación.

En conjunto, este capítulo proporciona el soporte teórico necesario para comprender el desarrollo metodológico y la construcción de los modelos predictivos presentados en los capítulos posteriores.

2.2 Estrategia de revisión de literatura

La revisión de la literatura se desarrolló con el propósito de identificar, analizar y sintetizar estudios científicos relevantes relacionados con la predicción de retrasos en actividades

de proyectos mediante técnicas de Ciencia de Datos y Machine Learning. Este proceso permitió fundamentar teóricamente la investigación, identificar variables relevantes y reconocer los principales enfoques metodológicos utilizados en investigaciones similares. La estrategia de revisión se diseñó bajo un enfoque sistemático, considerando el uso de bases de datos académicas reconocidas, cadenas de búsqueda estructuradas, criterios de inclusión y exclusión claramente definidos y un proceso sistemático de selección de literatura. De esta manera, se buscó garantizar la calidad, pertinencia y trazabilidad de las fuentes utilizadas durante el desarrollo del estudio.

2.2.1 Bases de datos y fuentes académicas

Fuente	Justificación de uso	Fecha de búsqueda
Scopus	Base multidisciplinaria con artículos científicos revisados por pares, relevante para ingeniería, minería y Data Science.	Junio 2026
Web of Science	Permite identificar estudios de alto impacto y rastrear la evolución de la investigación mediante citas.	Junio 2026
IEEE Xplore	Fuente especializada en ingeniería, inteligencia artificial y Machine Learning.	Junio 2026
ACM Digital Library	Proporciona literatura relevante en algoritmos, ciencia de datos y sistemas computacionales.	Julio 2026
Google Scholar	Complementa la búsqueda incluyendo tesis, artículos y literatura académica amplia.	Julio 2026

Table 1 Bases de datos y fuentes académicas

Las bases de datos seleccionadas fueron elegidas debido a su reconocimiento académico, cobertura temática y calidad de las publicaciones indexadas. Su utilización permitió acceder a investigaciones relacionadas con gestión de proyectos, análisis de cronogramas, Machine Learning y Ciencia de Datos, proporcionando una base teórica sólida para el desarrollo de la investigación.

2.2.2 Cadenas de búsqueda

Bloque	Términos sugeridos	Términos del estudiante
Fenómeno	delay OR deviation OR duration	schedule deviation OR project delay OR duration deviation
Contexto	projects OR construction OR mining	project management OR mining projects
Método	machine learning OR predictive models	machine learning OR data science OR predictive analytics
Datos	historical data OR schedule data	historical project data OR Primavera P6
Cadena final	generic example	("schedule deviation" OR "project delay") AND ("machine learning" OR "predictive models") AND ("project management" OR mining)

Table 2 Cadenas de búsqueda

Las cadenas de búsqueda fueron construidas mediante la combinación de palabras clave relacionadas con retrasos de proyectos, análisis de cronogramas, modelos predictivos y técnicas de Machine Learning. Para ampliar o restringir los resultados según la necesidad, se utilizaron operadores booleanos como AND y OR, facilitando la identificación de estudios directamente relacionados con el problema de investigación.

2.2.3 Criterios de inclusión y exclusión

Criterio	Incluir	Excluir
Temporalidad	Estudios entre 2020–2026 y autores clave anteriores	Estudios desactualizados sin relevancia
Pertinencia	Relacionados con predicción retrasos	Estudios generales sin relación directa
Calidad	Publicaciones indexadas en Scopus, WoS, IEEE	Blogs o fuentes sin rigor científico
Aplicabilidad	Estudios con modelos ML o análisis estadístico	Estudios sin datos comparables

Table 3 Criterios de inclusión y exclusión

Los criterios de inclusión y exclusión permitieron seleccionar estudios alineados con los objetivos de la investigación. Se priorizaron publicaciones científicas indexadas, estudios recientes y trabajos relacionados con la predicción de retrasos mediante enfoques cuantitativos, técnicas estadísticas o algoritmos de aprendizaje automático.

Por otra parte, se excluyeron documentos con escaso rigor científico, estudios sin una metodología claramente definida o investigaciones que no guardaban una relación directa con la predicción de retrasos en proyectos o actividades.

2.2.4 Proceso de selección de literatura

Etapa	Cantidad	Descripción
Registros identificados	15	Total de documentos encontrados en bases de datos
Duplicados eliminados	3	Registros repetidos eliminados
Registros cribados	12	Revisados por título y resumen
Textos completos evaluados	8	Análisis detallado
Estudios incluidos	4	Estudios finales utilizados

Table 4 Selección de literatura

El proceso de selección de literatura se desarrolló de manera progresiva. Inicialmente se identificó un conjunto amplio de publicaciones, posteriormente se eliminaron registros duplicados y se realizó una revisión preliminar basada en títulos y resúmenes. Finalmente, los documentos seleccionados fueron analizados en profundidad considerando su relevancia, calidad metodológica y aporte al tema de investigación.

Este procedimiento permitió obtener un conjunto de referencias representativas del estado actual del conocimiento en el ámbito de la predicción de retrasos y la aplicación de técnicas de Ciencia de Datos en gestión de proyectos.

2.2.5 Matriz de extracción y análisis de literatura

Los estudios seleccionados fueron analizados considerando diversos aspectos relevantes para la investigación, tales como:

- Tipo de datos utilizados.
- Variables empleadas.
- Algoritmos o modelos predictivos aplicados.
- Resultados obtenidos.
- Limitaciones identificadas.
- Aportes para futuras investigaciones.

Este análisis permitió identificar tendencias comunes, metodologías recurrentes y vacíos existentes en la literatura. Asimismo, facilitó la comparación de los estudios revisados y proporcionó una base sólida para la definición del enfoque metodológico adoptado en la presente investigación.

N°	Autor/año	Título	Contexto	Datos/variables	Método	Hallazgos	Limitaciones	Aporte
1	Lishner (2022)	Project Duration Prediction	Ingeniería	Datos históricos	Redes neuronales	Alta precisión	Requiere muchos datos	Modelos complejos
2	Pang (2023)	Hybrid ML Models	Construcción	Tiempo y costo	Modelo híbrido	Mejora predicción	Complejidad	Integración técnicas
3	Kothapalli (2024)	Schedule Prediction	Ingeniería	Datos cronograma	ML	+27% precisión	Dataset limitado	Validación empírica
4	Restrepo (2025)	Delay Prediction	Infraestructura	Desviaciones	ML	Alta precisión	Contexto civil	Aplicación real

Table 5 Matriz de extracción y análisis de literatura

2.3 Redacción de antecedentes

2.3.1 Contexto general del fenómeno

La predicción de retrasos en proyectos ha sido objeto de estudio en diversas disciplinas relacionadas con la ingeniería, la construcción y la gestión de proyectos, debido a la influencia que el cumplimiento de los cronogramas tiene sobre el desempeño de las organizaciones y el éxito de los proyectos.

Los retrasos representan uno de los problemas más frecuentes durante la ejecución de proyectos, ya que pueden generar incumplimientos de plazos, afectaciones operativas y dificultades en la coordinación de actividades. Como consecuencia, existe un creciente interés por desarrollar herramientas que permitan anticipar estos eventos y apoyar la toma de decisiones de manera oportuna.

Tradicionalmente, la gestión de cronogramas se ha basado en técnicas determinísticas y en la experiencia de los equipos de planificación. Sin embargo, el aumento en la disponibilidad de datos históricos y el desarrollo de técnicas de análisis de datos han permitido explorar nuevos enfoques orientados a identificar patrones asociados al comportamiento de las actividades dentro de los proyectos.

En este contexto, la Ciencia de Datos y el Machine Learning han surgido como alternativas capaces de analizar grandes volúmenes de información histórica y construir modelos predictivos orientados a anticipar eventos futuros, incluyendo el riesgo de retraso en actividades y proyectos.

2.3.2 Estudios sobre variables estructurales del cronograma

Diversas investigaciones han analizado la influencia que tienen las características estructurales de los cronogramas sobre el comportamiento de las actividades y el cumplimiento de los plazos establecidos.

Entre las variables más estudiadas se encuentran la duración de las actividades y las relaciones de dependencia existentes dentro de la red del cronograma. Estas características permiten representar la estructura interna de un proyecto y describir el nivel de complejidad e interdependencia entre las actividades.

La literatura reporta que actividades con mayores niveles de interdependencia suelen presentar una mayor exposición al riesgo de retraso, debido a que dependen del desempeño de otras actividades dentro de la red lógica del cronograma. De manera similar, la duración de las actividades ha sido utilizada frecuentemente como una variable explicativa para analizar diferencias en el comportamiento temporal de las tareas.

La relevancia de estas variables radica en que pueden obtenerse directamente a partir de los cronogramas, sin necesidad de incorporar información externa relacionada con costos, recursos o productividad. Esto favorece la construcción de modelos más generalizables y aplicables en distintos tipos de proyectos.

2.3.3 Estudios sobre métodos y modelos predictivos

La evolución de las técnicas de análisis de datos ha generado una transición progresiva desde métodos estadísticos tradicionales hacia algoritmos de Machine Learning para la predicción de eventos relacionados con el desempeño de proyectos.

Los modelos estadísticos han sido ampliamente utilizados como herramientas de referencia para evaluar relaciones entre variables y construir modelos explicativos. No obstante, estos enfoques presentan limitaciones cuando los datos contienen relaciones complejas o comportamientos no lineales.

Como alternativa, diversos estudios han incorporado algoritmos de aprendizaje automático capaces de capturar patrones más complejos presentes en los datos. Entre los métodos más utilizados se encuentran los árboles de decisión, Random Forest y otros modelos basados en ensambles, los cuales han demostrado resultados favorables en problemas de clasificación y predicción.

Estos enfoques permiten identificar interacciones entre variables, mejorar la capacidad predictiva y detectar patrones difíciles de representar mediante métodos tradicionales, contribuyendo al desarrollo de herramientas de apoyo para la gestión de proyectos.

2.3.4 Limitaciones de la literatura y necesidad de investigación

A pesar del crecimiento de las investigaciones relacionadas con la predicción de retrasos, aún existen importantes oportunidades de desarrollo dentro de este campo.

En primer lugar, una parte importante de los estudios disponibles se ha desarrollado utilizando conjuntos de datos limitados o datos simulados, reduciendo la posibilidad de validar los resultados en escenarios reales de ejecución de proyectos.

En segundo lugar, numerosos trabajos se centran en el análisis del desempeño a nivel de proyecto, prestando menor atención al comportamiento individual de las actividades que conforman el cronograma. Esta situación limita la identificación de patrones específicos asociados al riesgo de retraso dentro de la red de actividades.

Asimismo, muchas investigaciones incorporan variables relacionadas con recursos, costos, productividad u otras características que no siempre se encuentran disponibles o estandarizadas entre organizaciones. Esto dificulta la aplicación práctica de los modelos desarrollados en entornos reales.

Finalmente, existe una necesidad de estudios que evalúen y comparen diferentes algoritmos predictivos utilizando datos históricos reales y variables derivadas directamente de la estructura del cronograma.

En respuesta a estas limitaciones, la presente investigación propone analizar datos históricos provenientes de cronogramas desarrollados en Primavera P6 y construir modelos predictivos basados en variables estructurales de las actividades, con el objetivo de estimar el riesgo de retraso a nivel individual y comparar el desempeño de diferentes algoritmos de Machine Learning.

2.4 Bases conceptuales

En esta sección se presentan los conceptos fundamentales que sustentan el desarrollo de la investigación. Estos conceptos permiten comprender el fenómeno de estudio y establecer una base teórica para la selección de variables, métodos analíticos y modelos predictivos.

A diferencia de una recopilación aislada de definiciones, los conceptos se presentan en función de su aplicación dentro de esta investigación, la cual se enfoca en la identificación de patrones estructurales en cronogramas de proyectos y en la predicción del riesgo de retraso en actividades mediante técnicas de Ciencia de Datos y Machine Learning.

Concepto	Autores base	Definición sintetizada	Uso en la investigación
Data Science	Géron (2022)	Disciplina que integra estadística y programación para analizar datos y extraer conocimiento	Base para el análisis de datos históricos de cronogramas
Riesgo de retraso	Adaptado	Probabilidad de que una actividad no cumpla su fecha planificada	Variable dependiente del estudio
Machine Learning	Hastie et al. (2009)	Técnicas que permiten aprender patrones a partir de datos para realizar predicciones	Método principal para modelar el riesgo de retraso
Modelo predictivo	James et al. (2013)	Herramientas que permiten estimar comportamientos futuros a partir de datos históricos	Se utilizan para predecir el riesgo de retraso
Variables estructurales del cronograma	Adaptado	Características cuantificables de las actividades (duración y relaciones)	Variables independientes del modelo
Clasificación	Adaptado	Tipo de problema donde se predicen categorías (ej. atraso/no atraso)	Enfoque del modelo predictivo
Random Forest	Breiman (2001)	Algoritmo basado en múltiples árboles de decisión	Modelo principal de predicción
Árbol de decisión	Breiman et al.	Modelo basado en reglas para clasificación de datos	Modelo interpretable para identificar patrones

Table 6 Bases conceptuales

2.4.1 Ciencia de Datos aplicada al análisis de cronogramas

La Ciencia de Datos constituye un enfoque interdisciplinario que combina técnicas estadísticas, programación y análisis de datos con el objetivo de transformar información en conocimiento útil para la toma de decisiones.

En el ámbito de la gestión de proyectos, la Ciencia de Datos permite analizar grandes volúmenes de información histórica provenientes de cronogramas y registros de ejecución, facilitando la identificación de patrones asociados al comportamiento de las actividades.

La aplicación de este enfoque permite pasar de modelos de análisis descriptivo a modelos predictivos capaces de anticipar eventos futuros. En el contexto de esta investigación, la Ciencia de Datos proporciona el marco metodológico para analizar actividades históricas y desarrollar modelos capaces de estimar el riesgo de retraso a partir de características estructurales del cronograma.

2.4.2 Riesgo de retraso en actividades

El riesgo de retraso puede definirse como la probabilidad de que una actividad no cumpla con la fecha establecida dentro de un cronograma de proyecto. En el contexto de la gestión de proyectos, este concepto se asocia a la posibilidad de que una desviación en la ejecución afecte el cumplimiento de los plazos planificados y genere impactos sobre el desempeño del cronograma.

En esta investigación, el riesgo se interpreta como una medida prospectiva asociada a la posibilidad de incumplimiento de la fecha planificada de una actividad. Desde una perspectiva operativa, dicho riesgo se representa mediante la variable Delay_Risk, construida a partir de la comparación entre la fecha real de finalización y la fecha de finalización establecida en la línea base del cronograma.

Desde una perspectiva analítica, este concepto se modela como un problema de clasificación binaria que permite distinguir entre actividades que presentan retraso y actividades que cumplen su planificación original. De esta manera, el riesgo de retraso constituye la variable objetivo del estudio y sirve como base para el entrenamiento y evaluación de los modelos predictivos desarrollados.

2.4.3 Machine Learning y análisis predictivo

El Machine Learning corresponde a un conjunto de técnicas capaces de identificar patrones a partir de datos históricos para generar predicciones sobre eventos futuros.

A diferencia de los enfoques tradicionales basados únicamente en reglas predefinidas, estos algoritmos aprenden relaciones presentes en los datos y utilizan dicho conocimiento para clasificar o estimar comportamientos futuros.

En el caso de los cronogramas de proyectos, los modelos de Machine Learning permiten analizar simultáneamente múltiples variables relacionadas con la estructura de las actividades y generar estimaciones sobre su probabilidad de retraso.

Debido a esta capacidad, el Machine Learning se ha convertido en una herramienta relevante para apoyar la toma de decisiones en problemas donde existe incertidumbre y grandes volúmenes de datos históricos disponibles.

2.4.4 Modelos predictivos

Los modelos predictivos son herramientas analíticas que utilizan información histórica para estimar el comportamiento futuro de una variable de interés.

Su objetivo principal es identificar patrones presentes en los datos y transformarlos en reglas o funciones capaces de realizar predicciones sobre nuevas observaciones.

Dentro del contexto de esta investigación, los modelos predictivos son utilizados para estimar la probabilidad de retraso de una actividad a partir de variables estructurales derivadas del cronograma.

La comparación de distintos algoritmos permite determinar cuál de ellos presenta una mejor capacidad para identificar actividades potencialmente retrasadas.

2.4.5 Variables estructurales del cronograma

Las variables estructurales del cronograma corresponden a características cuantificables de las actividades que describen la forma en que estas se encuentran organizadas dentro de la red de planificación.

Estas variables permiten representar aspectos relacionados con la duración de las actividades y el nivel de dependencia existente entre ellas.

En esta investigación se consideran principalmente variables asociadas a:

- Duración de la actividad.
- Número de actividades predecesoras.
- Número de actividades sucesoras.

Estas características fueron seleccionadas debido a que están presentes de forma consistente en los cronogramas analizados y permiten representar la estructura de las actividades sin depender de información externa relacionada con recursos, costos o productividad.

2.4.6 Clasificación como problema de aprendizaje supervisado

La clasificación es un tipo de problema de aprendizaje supervisado cuyo objetivo consiste en asignar una observación a una categoría previamente definida.

Para desarrollar este proceso es necesario contar con ejemplos históricos para los cuales ya se conoce el resultado real. A partir de dichos ejemplos, el algoritmo aprende patrones que posteriormente utiliza para clasificar nuevos registros.

En esta investigación, el problema es abordado como una tarea de clasificación binaria, donde cada actividad puede pertenecer a una de las siguientes categorías:

- Actividad sin retraso.
- Actividad retrasada.

Este enfoque permite utilizar métricas específicas de clasificación para evaluar el desempeño de los modelos desarrollados.

2.4.7 Importancia de las variables en el proceso predictivo

La capacidad predictiva de un modelo depende en gran medida de la calidad y relevancia de las variables utilizadas durante el entrenamiento.

Por esta razón, la selección de variables constituye un elemento esencial dentro del proceso de análisis. Variables que representan adecuadamente el comportamiento de las actividades permiten capturar patrones significativos y mejorar el desempeño de los modelos.

En el presente estudio, las variables estructurales seleccionadas buscan representar características propias de la planificación de actividades, permitiendo evaluar hasta qué

punto la estructura del cronograma contiene información suficiente para anticipar retrasos durante la ejecución de proyectos.

2.5 Teorías, modelos y enfoques relacionados

2.5.1 Enfoque de análisis de datos aplicado a cronogramas

El análisis de cronogramas ha evolucionado desde enfoques tradicionales basados únicamente en técnicas de planificación hacia metodologías fundamentadas en el uso de datos históricos. Este enfoque busca identificar patrones ocultos en la información generada durante la ejecución de proyectos, permitiendo comprender mejor el comportamiento de las actividades y apoyar la toma de decisiones.

Desde esta perspectiva, los cronogramas pueden ser analizados como fuentes de datos estructurados, donde cada actividad contiene información que puede ser utilizada para estudiar fenómenos relacionados con el cumplimiento de plazos, el desempeño temporal y el riesgo de retraso.

La aplicación de métodos de análisis de datos permite transformar registros históricos en conocimiento útil, generando información que puede contribuir a una gestión más proactiva de los proyectos.

2.5.2 Modelos estadísticos como referencia

Los modelos estadísticos han sido ampliamente utilizados para analizar relaciones entre variables y explicar fenómenos observados en diversos contextos.

Dentro del ámbito predictivo, modelos como la regresión logística constituyen una referencia importante debido a su simplicidad e interpretabilidad. Estos modelos permiten estimar probabilidades y analizar la influencia de diferentes variables sobre un evento de interés.

Sin embargo, cuando los datos presentan comportamientos complejos o relaciones no lineales, los enfoques estadísticos tradicionales pueden presentar limitaciones en su capacidad predictiva. Por esta razón, suelen utilizarse como modelos de referencia para comparar enfoques más avanzados basados en Machine Learning.

2.5.3 Enfoque de Machine Learning predictivo

El Machine Learning constituye un enfoque orientado a la construcción de modelos capaces de aprender patrones a partir de datos históricos y utilizar dicho conocimiento para generar predicciones sobre nuevos casos.

A diferencia de los enfoques tradicionales, estos algoritmos son capaces de identificar relaciones complejas entre múltiples variables sin necesidad de establecer reglas explícitas de forma manual.

Este enfoque resulta particularmente útil cuando se trabaja con grandes volúmenes de datos históricos y cuando las relaciones entre variables presentan comportamientos no lineales o difíciles de modelar mediante técnicas estadísticas convencionales.

En el contexto de esta investigación, el Machine Learning permite analizar características estructurales de las actividades y estimar su probabilidad de retraso a partir de patrones identificados en cronogramas históricos.

2.5.4 Aprendizaje supervisado como enfoque metodológico

El aprendizaje supervisado corresponde a una categoría de algoritmos de Machine Learning que utilizan ejemplos históricos para aprender relaciones entre variables de entrada y una variable objetivo previamente conocida.

Durante el proceso de entrenamiento, el algoritmo analiza observaciones etiquetadas y construye un modelo capaz de clasificar nuevos registros o generar predicciones futuras.

La predicción del riesgo de retraso puede abordarse como un problema de clasificación supervisada, donde cada actividad pertenece a una de dos categorías posibles:

- Actividad sin retraso.
- Actividad retrasada.

Este enfoque permite entrenar modelos utilizando información histórica y posteriormente aplicar el conocimiento adquirido sobre actividades que aún no han finalizado.

2.5.5 Modelos de Machine Learning utilizados en la investigación

Con el fin de evaluar distintas estrategias predictivas, la presente investigación considera varios algoritmos ampliamente utilizados en problemas de clasificación.

- Regresión logística

La regresión logística se utiliza como modelo base de comparación debido a su simplicidad, interpretabilidad y amplia aplicación en problemas de clasificación binaria.

- Árboles de decisión

Los árboles de decisión permiten representar reglas de clasificación mediante estructuras jerárquicas que facilitan la interpretación de los resultados.

Entre sus principales ventajas se encuentran la facilidad de comprensión y la capacidad para identificar relaciones no lineales entre variables.

- Random Forest

Random Forest corresponde a un algoritmo basado en conjuntos de árboles de decisión que busca mejorar la capacidad predictiva individual de cada árbol mediante técnicas de agregación.

Este enfoque permite reducir el riesgo de sobreajuste y mejorar la estabilidad de las predicciones.

- XGBoost

XGBoost es un algoritmo basado en técnicas de Gradient Boosting que combina múltiples modelos débiles para construir un modelo predictivo más robusto.

Este algoritmo ha demostrado un desempeño destacado en numerosos problemas de clasificación debido a su capacidad para capturar relaciones complejas y optimizar progresivamente el proceso de aprendizaje.

2.5.6 Integración de enfoques

La presente investigación integra conceptos provenientes de la gestión de proyectos, el análisis de datos y el Machine Learning.

La gestión de proyectos proporciona el contexto en el que se generan los cronogramas y las actividades analizadas. El análisis de datos permite explorar y comprender el comportamiento histórico de dichas actividades. Finalmente, el Machine Learning aporta las herramientas necesarias para transformar esta información en modelos predictivos capaces de estimar el riesgo de retraso.

La combinación de estos enfoques permite desarrollar una metodología basada en evidencia, orientada a la construcción de modelos predictivos aplicados al análisis de cronogramas de proyectos.

2.6 Variables

En estudios cuantitativos orientados al desarrollo de modelos predictivos, la definición de variables constituye un elemento fundamental para estructurar el análisis y establecer relaciones entre los fenómenos observados.

En la presente investigación, las variables fueron definidas a partir de información histórica proveniente de cronogramas desarrollados en Primavera P6. La selección se realizó considerando únicamente variables estructurales del cronograma que se encontraban disponibles de forma consistente en los proyectos analizados.

A diferencia de otros enfoques que incorporan información relacionada con costos, recursos, productividad o rendimiento de los equipos de trabajo, este estudio se enfoca exclusivamente en características derivadas de la estructura de las actividades dentro del cronograma.

En consecuencia, se definió una variable objetivo relacionada con el riesgo de retraso y un conjunto de variables predictoras asociadas a la estructura de las actividades.

2.6.1 Clasificación de variables

Variable dependiente

La variable dependiente corresponde al riesgo de retraso de una actividad.

Para efectos del modelado, esta variable fue representada mediante una clasificación binaria denominada Delay_Risk, definida de la siguiente manera:

- Delay_Risk = 1: actividad retrasada.
- Delay_Risk = 0: actividad no retrasada.

La clasificación se realizó mediante la comparación entre la fecha real de finalización de cada actividad y la fecha de finalización establecida en la línea base del cronograma.

Esta variable representa el fenómeno principal de estudio y constituye el objetivo de los modelos predictivos desarrollados durante la investigación.

Variables independientes

Las variables independientes corresponden a características estructurales extraídas directamente de los cronogramas analizados.

Las variables finalmente utilizadas fueron:

- Duración original de la actividad.
- Número de predecesores.
- Número de sucesores.

Estas variables permiten representar aspectos relacionados con la complejidad estructural de una actividad y su posición dentro de la red lógica del cronograma.

Variable	Definición conceptual	Indicador	Fuente
Riesgo de retraso (dependiente)	Probabilidad de que una actividad no cumpla su fecha programada	0 = no retraso / 1 = retraso	Adaptado
Duración	Tiempo asociado a la ejecución de una actividad	Días	Literatura general
Nº de predecesores	Cantidad de actividades previas que condicionan una actividad	Número	Análisis de redes
Nº de sucesores	Cantidad de actividades dependientes posteriores	Número	Análisis de redes

Table 7 Variables

2.6.2 Descripción de las variables utilizadas

- Duración original de la actividad

Representa el tiempo planificado inicialmente para la ejecución de una actividad dentro del cronograma.

Esta variable se expresa en días y permite caracterizar la magnitud temporal de la actividad.

Se considera una de las variables más utilizadas en estudios relacionados con planificación y control de proyectos debido a su influencia potencial sobre el comportamiento de las actividades.

- Número de predecesores

Corresponde a la cantidad de actividades que deben cumplirse previamente para que una actividad pueda iniciar o continuar su ejecución.

Esta variable representa el nivel de dependencia que posee una actividad respecto a otras actividades dentro de la red lógica del cronograma.

Un mayor número de predecesores puede incrementar la exposición de una actividad a retrasos acumulados provenientes de actividades anteriores.

- Número de sucesores

Corresponde a la cantidad de actividades dependientes de una actividad determinada.

Esta variable permite representar el nivel de conectividad de una actividad dentro del cronograma y su influencia potencial sobre actividades posteriores.

La cantidad de sucesores puede considerarse un indicador de relevancia estructural dentro de la red de planificación.

2.6.3 Justificación de las variables seleccionadas

La selección de variables se fundamentó en la necesidad de representar la estructura de las actividades mediante información disponible de manera consistente en los cronogramas analizados.

La duración original permite caracterizar el tamaño temporal de las actividades, mientras que el número de predecesores y sucesores permite representar las relaciones de dependencia existentes dentro de la red lógica del cronograma.

La utilización exclusiva de variables estructurales responde al interés de evaluar hasta qué punto la propia configuración del cronograma contiene información suficiente para anticipar la ocurrencia de retrasos, sin recurrir a variables externas relacionadas con recursos, costos o condiciones específicas de cada proyecto.

2.6.4 Relación de las variables con el modelo analítico

Desde una perspectiva analítica, el modelo busca identificar patrones existentes entre las características estructurales de las actividades y la ocurrencia de retrasos durante la ejecución de los proyectos.

El supuesto principal de la investigación es que ciertas configuraciones dentro de la estructura del cronograma pueden estar asociadas a una mayor probabilidad de retraso.

Por esta razón, las variables seleccionadas fueron utilizadas como predictores dentro de distintos modelos de Machine Learning con el objetivo de estimar el riesgo de retraso de actividades individuales.

2.6.5 Importancia de las variables en la investigación

Las variables definidas en esta sección constituyen la base del proceso analítico desarrollado en los capítulos posteriores.

Estas variables fueron utilizadas durante:

- La preparación de los datos.
- El análisis exploratorio de datos (EDA).
- El entrenamiento de los modelos predictivos.
- La evaluación y comparación del desempeño de los algoritmos.

Su correcta definición garantiza la coherencia entre el marco teórico, la metodología y los resultados obtenidos en la presente investigación.

2.7 Brecha de investigación

2.7.1 Introducción de la brecha

A partir de la revisión de literatura realizada, se evidencia un creciente interés por la aplicación de técnicas de Ciencia de Datos y Machine Learning en el análisis del desempeño de proyectos y la predicción de retrasos. Diversas investigaciones han demostrado que los datos históricos pueden utilizarse para identificar patrones asociados al comportamiento temporal de proyectos y actividades.

Sin embargo, también se observa que gran parte de las investigaciones disponibles presentan limitaciones relacionadas con la disponibilidad de datos, el nivel de análisis considerado y las variables utilizadas para construir los modelos predictivos.

Estas limitaciones muestran la necesidad de continuar desarrollando investigaciones orientadas a comprender el comportamiento de las actividades dentro de los cronogramas utilizando enfoques basados en datos reales y metodologías reproducibles.

2.7.2 Conocimiento existente

La literatura revisada muestra que los modelos predictivos basados en Machine Learning poseen capacidad para identificar patrones asociados al desempeño de proyectos y cronogramas.

Asimismo, diferentes investigaciones han evidenciado que ciertas características de las actividades y de los cronogramas contienen información relevante para anticipar desviaciones en los plazos planificados.

También existe consenso respecto al potencial de algoritmos de clasificación como los árboles de decisión, Random Forest y otros modelos de aprendizaje supervisado para abordar problemas relacionados con la predicción de retrasos y el análisis de riesgos en proyectos.

En términos generales, la evidencia disponible sugiere que los datos históricos constituyen una fuente valiosa de información para apoyar la toma de decisiones y mejorar la capacidad de anticipación ante posibles desviaciones en la ejecución de proyectos.

2.7.3 Limitaciones identificadas en la literatura

A pesar de los avances observados, se identifican diversas limitaciones en los estudios revisados.

En primer lugar, una parte importante de las investigaciones utiliza conjuntos de datos simulados o bases de datos de tamaño reducido, limitando la validación de los modelos en contextos reales de ejecución de proyectos.

En segundo lugar, numerosos estudios centran su análisis en indicadores globales de proyecto, prestando menor atención al comportamiento individual de las actividades que conforman los cronogramas.

En tercer lugar, varios modelos dependen de variables relacionadas con recursos, costos, productividad o características específicas de las organizaciones, las cuales no siempre se encuentran disponibles o estandarizadas entre distintos proyectos.

Finalmente, existe una cantidad limitada de investigaciones que comparen distintos algoritmos predictivos utilizando datos históricos reales provenientes de sistemas de planificación ampliamente utilizados en la industria, como Primavera P6.

2.7.4 Necesidad del estudio

Las limitaciones identificadas evidencian la necesidad de desarrollar investigaciones orientadas al análisis de actividades individuales utilizando datos históricos reales y variables disponibles directamente en los cronogramas.

En este contexto, resulta pertinente analizar si las características estructurales de las actividades contienen información suficiente para anticipar la ocurrencia de retrasos, sin depender de variables externas asociadas al contexto particular de cada proyecto.

Asimismo, se considera relevante comparar diferentes algoritmos de Machine Learning bajo un mismo conjunto de datos para determinar cuáles presentan mejor capacidad predictiva frente al problema estudiado.

La presente investigación responde a esta necesidad mediante el análisis de actividades extraídas de cronogramas históricos desarrollados en Primavera P6 y la construcción de modelos predictivos orientados a estimar el riesgo de retraso a nivel de actividad.

2.7.5 Aporte de la investigación

La contribución de esta investigación se concentra en varios aspectos.

Uso de datos históricos reales.

La investigación utiliza información proveniente de cronogramas reales desarrollados y gestionados mediante Primavera P6, permitiendo evaluar el desempeño de los modelos en condiciones cercanas a un entorno operativo.

Análisis a nivel de actividades.

El estudio se enfoca en actividades individuales en lugar de analizar únicamente indicadores agregados de proyecto, permitiendo identificar patrones específicos asociados al riesgo de retraso.

Uso de variables estructurales del cronograma.

La investigación evalúa la capacidad predictiva de variables derivadas directamente de la estructura del cronograma, tales como la duración de las actividades y sus relaciones de dependencia.

Comparación de modelos predictivos.

Se comparan distintos algoritmos de clasificación con el fin de determinar cuál ofrece el mejor desempeño para la predicción del riesgo de retraso.

Aplicación práctica de los resultados.

Los resultados obtenidos se complementan con la implementación de una visualización orientada a facilitar la identificación de actividades con mayor probabilidad de retraso, contribuyendo al uso práctico de los modelos desarrollados.

2.8 Síntesis del capítulo

El desarrollo del presente capítulo permitió establecer los fundamentos teóricos y conceptuales necesarios para abordar el problema de investigación relacionado con la predicción del riesgo de retraso en actividades mediante técnicas de Ciencia de Datos y Machine Learning.

A través de la revisión de la literatura se identificó que el análisis del comportamiento de actividades dentro de los cronogramas constituye un tema de creciente interés en los ámbitos de la gestión de proyectos, la ingeniería y la analítica de datos. Asimismo, se evidenció que los retrasos representan uno de los problemas más frecuentes durante la ejecución de proyectos, generando impactos significativos sobre el cumplimiento de los plazos planificados.

Los antecedentes revisados muestran que el uso de datos históricos y modelos predictivos permite identificar patrones asociados al comportamiento de actividades y proyectos. Sin embargo, también se observaron limitaciones relacionadas con la utilización de conjuntos de datos reducidos, el uso de información simulada y la escasez de estudios enfocados en actividades individuales utilizando datos reales provenientes de sistemas de planificación.

Las bases conceptuales desarrolladas permitieron definir los principales elementos que sustentan la investigación, incluyendo conceptos relacionados con Ciencia de Datos,

Machine Learning, análisis predictivo, riesgo de retraso y variables estructurales del cronograma. Estos conceptos constituyen el marco de referencia para la construcción y evaluación de los modelos desarrollados posteriormente.

De igual forma, la revisión de los principales enfoques y modelos analíticos permitió comprender las ventajas y limitaciones de los métodos estadísticos tradicionales y de los algoritmos de aprendizaje automático aplicados a problemas de clasificación. En este contexto, se justificó la utilización de técnicas de Machine Learning para identificar patrones complejos presentes en los datos históricos de cronogramas.

Asimismo, se definieron las variables consideradas en la investigación, destacando la duración de las actividades, el número de predecesores y el número de sucesores como representaciones de la estructura interna del cronograma. Estas variables constituyen la base para el análisis aplicado en los capítulos posteriores.

Finalmente, la identificación de las brechas existentes en la literatura permitió justificar la pertinencia de este estudio, particularmente en relación con el uso de datos históricos reales, el análisis a nivel de actividades individuales y la comparación de distintos modelos predictivos dentro de un mismo contexto de investigación.

En conjunto, los elementos desarrollados en este capítulo proporcionan el soporte conceptual necesario para el desarrollo de la metodología presentada en el capítulo siguiente, donde se describen los procedimientos utilizados para la recopilación, preparación y análisis de los datos, así como la construcción y evaluación de los modelos predictivos empleados en la investigación.

CAPÍTULO 3. METODOLOGÍA

3.1 Enfoque de investigación

La presente investigación adopta un enfoque cuantitativo, debido a que se fundamenta en el análisis de datos numéricos extraídos de cronogramas históricos de proyectos. Este enfoque permite estudiar de manera objetiva las características de las actividades y su relación con la ocurrencia de retrasos mediante el uso de técnicas estadísticas y algoritmos de Machine Learning.

El estudio se basa en la utilización de información histórica proveniente de cronogramas desarrollados en Primavera P6, a partir de la cual se identifican patrones estructurales asociados al comportamiento de las actividades. Posteriormente, dichos patrones son utilizados para construir modelos predictivos orientados a estimar el riesgo de retraso.

El enfoque cuantitativo resulta adecuado para esta investigación debido a que permite medir variables específicas del cronograma, evaluar relaciones entre ellas y desarrollar modelos capaces de generar predicciones sobre nuevas observaciones. Asimismo, facilita la comparación objetiva entre diferentes algoritmos de clasificación mediante métricas de desempeño definidas previamente.

En consecuencia, la investigación se desarrolla bajo una perspectiva analítica y predictiva, orientada a transformar datos históricos en información útil para apoyar la gestión del cronograma y la toma de decisiones en proyectos.

3.2 Diseño de investigación

La presente investigación adopta un diseño no experimental, debido a que las variables analizadas no fueron manipuladas por el investigador, sino que fueron observadas y analizadas a partir de información histórica existente en cronogramas de proyectos.

Asimismo, el estudio posee un carácter retrospectivo, ya que utiliza datos históricos registrados en proyectos previamente ejecutados para identificar patrones asociados al retraso de actividades. Este enfoque permite analizar el comportamiento real de las actividades y construir modelos predictivos basados en evidencia histórica.

Desde el punto de vista analítico, la investigación presenta un alcance correlacional con componente predictivo. El componente correlacional se fundamenta en el análisis de las

relaciones existentes entre las variables estructurales del cronograma y el riesgo de retraso de las actividades. Por su parte, el componente predictivo busca utilizar dichas relaciones para desarrollar modelos capaces de estimar la probabilidad de retraso de nuevas actividades.

La investigación se desarrolla mediante la aplicación de técnicas de Ciencia de Datos y Machine Learning sobre información proveniente de cronogramas históricos, permitiendo identificar patrones recurrentes y evaluar la capacidad de distintos algoritmos para predecir el riesgo de retraso.

En consecuencia, el diseño seleccionado resulta coherente con los objetivos planteados, ya que permite analizar datos históricos reales, identificar relaciones relevantes entre variables y construir modelos predictivos aplicables a la gestión de cronogramas de proyectos.

3.3 Fuente de datos

La presente investigación se fundamenta en el análisis de datos históricos provenientes de cronogramas de proyectos gestionados mediante el software Primavera P6. Esta herramienta es ampliamente utilizada para la planificación, programación y control de proyectos, permitiendo registrar información detallada sobre actividades, duraciones, relaciones lógicas y estado de avance.

Los datos utilizados corresponden a cronogramas reales ejecutados dentro de una organización, los cuales fueron extraídos y consolidados para conformar una base de datos única para el análisis. Con el propósito de preservar la confidencialidad de la información, los identificadores de proyectos y actividades fueron anonimizados antes de la presentación de resultados y la elaboración de productos derivados de la investigación.

Inicialmente se consolidó información proveniente de diez proyectos diferentes, obteniendo un conjunto de datos compuesto por 17.679 actividades. Posteriormente, se aplicaron procesos de selección y depuración para identificar aquellas actividades que disponían de la información necesaria para determinar de forma objetiva la ocurrencia o no de retrasos.

La utilización de datos históricos reales constituye una de las principales fortalezas de la investigación, ya que permite evaluar el comportamiento de las actividades en condiciones reales de ejecución y desarrollar modelos predictivos basados en situaciones efectivamente observadas dentro de los proyectos analizados.

Asimismo, el uso de información proveniente directamente de los cronogramas permite garantizar la consistencia de las variables analizadas y facilita la reproducibilidad de la metodología empleada.

3.4 Unidad de análisis

La unidad de análisis de la presente investigación corresponde a las actividades individuales registradas dentro de los cronogramas de proyectos.

La elección de las actividades como unidad de análisis responde al objetivo de identificar patrones estructurales asociados al riesgo de retraso a un nivel de detalle más específico que el obtenido mediante indicadores globales de proyecto. Este enfoque permite estudiar el comportamiento particular de cada actividad y evaluar cómo determinadas características estructurales pueden influir en la ocurrencia de retrasos.

A partir de la consolidación inicial de los cronogramas se obtuvo un conjunto de 17.679 actividades. Posteriormente, se seleccionaron las actividades con estado Completed, debido a que únicamente estas permitían comparar las fechas reales de finalización con las fechas de referencia establecidas en la línea base del cronograma.

Como resultado de este proceso, se identificaron 8.877 actividades completadas. Luego de verificar la disponibilidad y consistencia de la información necesaria para construir la variable objetivo, el conjunto final utilizado para el modelado quedó conformado por 8.092 actividades.

Este conjunto de datos permitió desarrollar el análisis exploratorio, entrenar los modelos predictivos y evaluar su desempeño utilizando una muestra suficientemente amplia y representativa para los propósitos de la investigación.

3.5 Variables de estudio

Las variables utilizadas en la presente investigación fueron definidas a partir de la información disponible en los cronogramas históricos analizados. Con base en los objetivos

planteados y en la disponibilidad de datos consistentes, se estableció una variable dependiente asociada al riesgo de retraso de las actividades y un conjunto de variables independientes de carácter estructural.

La selección de variables se realizó considerando únicamente información derivada directamente del cronograma, con el fin de evaluar la capacidad predictiva de la propia estructura de planificación sin incorporar variables externas relacionadas con recursos, costos o productividad.

3.5.1 Variable dependiente

La variable dependiente corresponde al riesgo de retraso de una actividad, representado mediante la variable binaria denominada Delay_Risk.

La construcción de esta variable se realizó mediante la comparación entre la fecha real de finalización de cada actividad y la fecha de finalización establecida en la línea base del cronograma.

La clasificación utilizada fue la siguiente:

- Delay_Risk = 1: actividad retrasada.
- Delay_Risk = 0: actividad no retrasada.

De esta manera, una actividad fue considerada retrasada cuando su fecha real de finalización fue posterior a la fecha de finalización definida en la línea base.

Esta variable constituye el objetivo principal de los modelos predictivos desarrollados durante la investigación.

3.5.2 Variables independientes

Las variables independientes corresponden a características estructurales extraídas de los cronogramas utilizados en el estudio.

- **Duración original de la actividad**

Representa la duración planificada inicialmente para cada actividad dentro del cronograma, expresada en días.

Esta variable permite caracterizar la magnitud temporal de las actividades y constituye una de las variables más utilizadas en el análisis de cronogramas.

- **Número de predecesores**

Corresponde a la cantidad de actividades que deben completarse previamente para permitir el inicio o continuidad de una actividad determinada.

Esta variable representa el nivel de dependencia existente dentro de la red lógica del cronograma.

- **Número de sucesores**

Corresponde a la cantidad de actividades que dependen de una actividad específica para su ejecución.

Esta variable permite representar el grado de conectividad de una actividad dentro del cronograma y su relación con otras actividades posteriores.

3.5.3 Resumen de variables utilizadas

Variable	Tipo	Descripción
Delay_Risk	Dependiente	Actividad retrasada (1) o no retrasada (0)
Original Duration(d)	Independiente	Duración original de la actividad
Num_Predecessors	Independiente	Número de actividades predecesoras
Num_Successors	Independiente	Número de actividades sucesoras

Table 8 Resumen de variables utilizadas

3.5.4 Justificación de las variables seleccionadas

Las variables utilizadas fueron seleccionadas debido a que representan características estructurales presentes de forma consistente en todos los cronogramas analizados.

La duración original permite caracterizar el tamaño temporal de las actividades, mientras que el número de predecesores y sucesores permite representar la complejidad de las relaciones lógicas dentro de la red de planificación.

La selección de estas variables se encuentra alineada con el objetivo de la investigación, el cual busca evaluar si la propia estructura del cronograma contiene información suficiente para anticipar el riesgo de retraso de las actividades.

Asimismo, la utilización exclusiva de variables derivadas directamente del cronograma favorece la reproducibilidad de la metodología y facilita su aplicación en diferentes contextos de gestión de proyectos.

3.6 Preparación de datos

La preparación de los datos constituyó una etapa fundamental dentro de la investigación, ya que permitió transformar la información histórica proveniente de los cronogramas en un conjunto de datos adecuado para el análisis y el desarrollo de modelos predictivos.

El proceso incluyó actividades de consolidación, limpieza, transformación y validación de la información, con el propósito de garantizar la calidad y consistencia de los datos utilizados durante el estudio.

3.6.1 Consolidación de la información

La información fue obtenida a partir de cronogramas históricos desarrollados en Primavera P6, correspondientes a diez proyectos diferentes.

Los archivos exportados fueron integrados en una única base de datos maestra, permitiendo consolidar toda la información necesaria para el análisis en un solo conjunto de datos.

Como resultado de esta etapa se obtuvo una base inicial compuesta por 17.679 actividades provenientes de los proyectos analizados.

3.6.2 Selección de actividades para el análisis

Debido a que el objetivo de la investigación consistía en predecir el riesgo de retraso de las actividades, fue necesario identificar aquellas observaciones que permitieran determinar de manera objetiva la ocurrencia o no de retrasos.

Por esta razón, se seleccionaron únicamente las actividades con estado Completed, ya que estas disponían de información histórica suficiente para comparar las fechas reales de finalización con las fechas establecidas en la línea base del cronograma.

Mediante este procedimiento se obtuvo un subconjunto de 8.877 actividades completadas.

Posteriormente, se eliminaron aquellas actividades que presentaban información incompleta en las variables utilizadas para determinar el retraso, obteniéndose un conjunto final de 8.092 actividades aptas para el modelado.

3.6.3 Construcción de la variable objetivo

La variable objetivo fue definida con el propósito de representar el riesgo de retraso de las actividades mediante una clasificación binaria.

Para ello se comparó la fecha real de finalización de cada actividad con la fecha de finalización correspondiente a la línea base del cronograma.

La clasificación utilizada fue la siguiente:

- Delay_Risk = 1: actividad retrasada.
- Delay_Risk = 0: actividad no retrasada.

De esta manera, las actividades cuya fecha real de finalización fue posterior a la fecha prevista en la línea base fueron clasificadas como retrasadas.

Este procedimiento permitió transformar el problema de investigación en un problema de clasificación supervisada.

3.6.4 Construcción de variables derivadas

Además de las variables originalmente disponibles en los cronogramas, se construyeron variables derivadas destinadas a representar la estructura de relaciones existente entre las actividades.

- **Número de predecesores**

El número de predecesores fue calculado a partir de la información de relaciones lógicas registrada en el cronograma.

Esta variable representa la cantidad de actividades de las cuales depende una actividad determinada.

- **Número de sucesores**

El número de sucesores fue calculado mediante el conteo de las actividades dependientes asociadas a cada actividad.

Esta variable permite cuantificar el nivel de conexión de una actividad dentro de la red del cronograma.

La incorporación de estas variables permitió representar aspectos estructurales de la planificación que no se encontraban disponibles de forma directa dentro del conjunto original de datos.

3.6.5 Verificación de calidad de datos

Una vez construidas las variables de estudio, se realizó una revisión de calidad con el fin de identificar posibles problemas relacionados con valores faltantes, inconsistencias o errores de registro.

El análisis confirmó la ausencia de valores nulos en las variables utilizadas para el entrenamiento de los modelos predictivos, lo que permitió trabajar con un conjunto de datos completamente utilizable sin necesidad de aplicar técnicas de imputación.

Asimismo, se verificó la consistencia de las variables seleccionadas y la correcta construcción de la variable objetivo antes de iniciar las etapas de análisis exploratorio y modelado.

3.6.6 Tratamiento de valores extremos

Durante el análisis preliminar se identificó la presencia de actividades con duraciones significativamente superiores al comportamiento habitual observado en el conjunto de datos.

Sin embargo, tras su revisión se determinó que dichos registros correspondían a actividades reales de los proyectos analizados y no a errores de captura o inconsistencias de información.

Por esta razón, se decidió conservar los valores extremos dentro del conjunto de datos, permitiendo que los modelos aprendieran el comportamiento real presente en los cronogramas históricos.

La conservación de estos registros contribuye a mantener la representatividad del conjunto de datos y favorece la capacidad de generalización de los modelos desarrollados.

3.7 Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (Exploratory Data Analysis, EDA) se realizó con el propósito de comprender las características generales del conjunto de datos, identificar posibles patrones asociados al riesgo de retraso y verificar la calidad de la información antes de iniciar el proceso de modelado.

Esta etapa permitió analizar la distribución de las variables, detectar valores extremos, examinar la presencia de datos faltantes y estudiar las relaciones existentes entre las variables predictoras y la variable objetivo.

3.7.1 Distribución de la variable objetivo

Una vez construida la variable Delay_Risk, se analizó la distribución de las actividades retrasadas y no retrasadas dentro del conjunto de datos.

Los resultados mostraron que de las 8.092 actividades utilizadas para modelado:

- 5.533 actividades fueron clasificadas como no retrasadas.
- 2.559 actividades fueron clasificadas como retrasadas.

Esto representa aproximadamente:

- 68,4 % de actividades no retrasadas.
- 31,6 % de actividades retrasadas.

La distribución observada indica un nivel de balance adecuado para el desarrollo de modelos supervisados de clasificación, permitiendo entrenar y evaluar los algoritmos sin aplicar técnicas especiales de balanceo de clases.

3.7.2 Análisis de las variables predictoras

Se realizó un análisis descriptivo de las variables independientes seleccionadas para el estudio:

- Duración original de la actividad.
- Número de predecesores.
- Número de sucesores.

El análisis evidenció que la duración de las actividades presenta una distribución asimétrica positiva, donde la mayoría de las actividades poseen duraciones relativamente cortas y un grupo reducido presenta duraciones significativamente mayores.

Por su parte, las variables relacionadas con la estructura de dependencias mostraron una concentración importante de actividades con pocos predecesores y sucesores, coexistiendo con un número reducido de actividades altamente conectadas dentro de la red lógica del cronograma.

3.7.3 Identificación de valores extremos

El análisis de la duración de las actividades permitió identificar la presencia de valores extremos asociados a actividades de larga duración.

Sin embargo, la revisión de estos registros mostró que correspondían a actividades reales presentes en los cronogramas históricos y no a errores de captura o inconsistencias en la información.

Por esta razón, se decidió conservar dichos registros dentro del conjunto de datos, permitiendo que los modelos aprendieran a partir de la variabilidad real existente en los proyectos analizados.

3.7.4 Relación entre variables y riesgo de retraso

Con el fin de identificar patrones preliminares en los datos, se realizaron análisis gráficos y estadísticos comparando las variables predictoras con la variable objetivo.

Los resultados sugirieron que la duración de las actividades, por sí sola, no presenta una relación claramente diferenciadora entre actividades retrasadas y no retrasadas.

Por otro lado, el número de predecesores mostró una tendencia más consistente, observándose que las actividades retrasadas suelen presentar una mayor cantidad de dependencias previas en comparación con las actividades que cumplieron sus fechas planificadas.

Estos hallazgos proporcionaron evidencia inicial acerca de la posible influencia que la estructura del cronograma puede ejercer sobre el comportamiento de las actividades.

3.7.5 Análisis de correlación

Finalmente, se calculó una matriz de correlación entre las variables incluidas en el estudio con el propósito de explorar posibles asociaciones lineales.

Los resultados evidenciaron que el número de predecesores presentó la asociación más relevante con la variable Delay_Risk, aunque con una magnitud moderada. Las variables duración y número de sucesores mostraron relaciones más débiles con el riesgo de retraso. Si bien las correlaciones observadas no fueron suficientemente altas para explicar por sí solas el comportamiento de la variable objetivo, los resultados sugirieron la existencia de relaciones más complejas que podrían ser capturadas mediante algoritmos de Machine Learning.

3.7.6 Conclusiones del análisis exploratorio

El análisis exploratorio permitió confirmar la calidad y consistencia del conjunto de datos utilizado en la investigación. Asimismo, evidenció que las variables estructurales del cronograma contienen información potencialmente útil para la predicción del riesgo de retraso.

Los hallazgos obtenidos justificaron la continuidad hacia la etapa de modelado predictivo, donde se evaluaron distintos algoritmos de clasificación con el objetivo de identificar patrones que no pueden ser observados únicamente mediante técnicas descriptivas tradicionales.

3.8 Construcción de modelos predictivos

Con el propósito de evaluar la capacidad de las variables estructurales del cronograma para predecir el riesgo de retraso en actividades, se desarrollaron distintos modelos de clasificación supervisada utilizando técnicas de Machine Learning.

Como variables predictoras se utilizaron la duración original de la actividad, el número de predecesores y el número de sucesores. Por su parte, la variable objetivo correspondió a la clasificación binaria Delay_Risk, la cual identifica actividades retrasadas y no retrasadas.

Antes del entrenamiento de los modelos, el conjunto de datos fue dividido en dos subconjuntos:

- Conjunto de entrenamiento (80 %).
- Conjunto de prueba (20 %).

La división se realizó mediante un procedimiento de muestreo estratificado, con el fin de mantener la misma proporción de actividades retrasadas y no retrasadas en ambos conjuntos de datos.

Posteriormente, se entrenaron y evaluaron diferentes algoritmos de clasificación, con el objetivo de comparar su desempeño y seleccionar la alternativa más adecuada para el problema de investigación.

3.8.1 Regresión logística

La regresión logística fue utilizada como modelo base de comparación debido a su amplia aplicación en problemas de clasificación binaria y a su facilidad de interpretación.

Este modelo estima la probabilidad de ocurrencia de un evento utilizando una función logística, permitiendo establecer relaciones entre la variable objetivo y las variables predictoras.

La incorporación de este algoritmo permitió disponer de una referencia inicial para evaluar el desempeño de modelos más avanzados de Machine Learning.

3.8.2 Árbol de decisión

El árbol de decisión corresponde a un algoritmo de clasificación basado en reglas jerárquicas que segmenta los datos mediante sucesivas divisiones de las variables predictoras.

Este enfoque permite representar de forma intuitiva las reglas de clasificación utilizadas por el modelo y facilita la interpretación de los resultados.

Además de su carácter explicativo, el algoritmo posee la capacidad de capturar relaciones no lineales entre las variables analizadas.

3.8.3 Random Forest

Random Forest es un modelo basado en conjuntos de árboles de decisión que combina múltiples clasificadores para mejorar la estabilidad y precisión de las predicciones.

El algoritmo genera numerosos árboles utilizando subconjuntos aleatorios de datos y variables, integrando posteriormente sus resultados mediante un mecanismo de votación.

Esta estrategia permite reducir el riesgo de sobreajuste y mejorar el desempeño predictivo en comparación con un único árbol de decisión.

Para el entrenamiento del modelo se utilizaron 200 árboles de decisión (n_estimators = 200) y una semilla aleatoria fija (random_state = 42), con el propósito de garantizar la reproducibilidad de los resultados obtenidos.

3.8.4 XGBoost

XGBoost corresponde a un algoritmo basado en técnicas de Gradient Boosting que construye de forma secuencial múltiples modelos con el propósito de corregir los errores generados por iteraciones anteriores.

Este algoritmo es ampliamente utilizado en problemas de análisis predictivo debido a su capacidad para modelar relaciones complejas y obtener altos niveles de desempeño en tareas de clasificación.

En la presente investigación, XGBoost fue incorporado como alternativa de alto rendimiento para evaluar su capacidad de identificar patrones asociados al riesgo de retraso en actividades.

Los hiperparámetros utilizados fueron definidos mediante pruebas exploratorias realizadas durante el proceso de desarrollo del modelo. Se evaluaron distintas configuraciones hasta identificar una combinación que ofreciera un equilibrio adecuado entre desempeño predictivo, estabilidad de resultados y simplicidad del modelo. Con el fin de garantizar la reproducibilidad de los resultados, se empleó una semilla aleatoria fija (random_state = 42).

Para el modelo XGBoost se utilizaron los siguientes hiperparámetros: n_estimators = 200, max_depth = 4, learning_rate = 0.05, random_state = 42 y eval_metric = "logloss". La selección final se realizó considerando el desempeño observado sobre los datos de evaluación y la capacidad del modelo para discriminar entre actividades retrasadas y no retrasadas.

3.8.5 Comparación de modelos

Una vez entrenados los diferentes algoritmos, se realizó una comparación de desempeño utilizando el conjunto de prueba.

La evaluación permitió analizar la capacidad de cada modelo para clasificar correctamente actividades retrasadas y no retrasadas, así como identificar la alternativa con mejor desempeño para los objetivos de la investigación.

Los resultados detallados de esta comparación se presentan en el Capítulo 4.

3.9 Métricas de evaluación

La evaluación del desempeño de los modelos predictivos se realizó mediante métricas ampliamente utilizadas en problemas de clasificación supervisada. Estas métricas permiten analizar la capacidad de los algoritmos para identificar correctamente actividades retrasadas y no retrasadas, proporcionando una medida objetiva de su rendimiento.

La utilización de múltiples métricas resulta necesaria debido a que cada una evalúa aspectos diferentes del proceso de clasificación. De esta manera, se obtiene una visión más completa del comportamiento de los modelos desarrollados.

3.9.1 Accuracy

La métrica de Accuracy representa la proporción de predicciones correctas realizadas por un modelo respecto al total de observaciones evaluadas.

Esta métrica proporciona una visión general del desempeño del algoritmo y permite conocer la capacidad global del modelo para clasificar correctamente las actividades.

Sin embargo, su interpretación debe realizarse con cautela, ya que en algunos casos puede verse influenciada por la distribución de las clases presentes en el conjunto de datos.

3.9.2 Precision

La Precision mide la proporción de actividades clasificadas como retrasadas que efectivamente presentaron retraso.

Esta métrica permite evaluar la confiabilidad de las predicciones positivas realizadas por el modelo y resulta útil cuando el objetivo es reducir la cantidad de falsas alarmas generadas durante el proceso de clasificación.

Un valor elevado de Precision indica que la mayoría de las actividades identificadas como riesgosas realmente presentan retraso.

3.9.3 Recall

El Recall mide la capacidad del modelo para identificar correctamente las actividades que realmente presentan retraso.

Esta métrica fue considerada especialmente relevante para la investigación, debido a que el objetivo principal consiste en detectar oportunamente actividades con riesgo potencial de atraso.

Un valor elevado de Recall indica que el modelo logra identificar una mayor proporción de actividades retrasadas, reduciendo la cantidad de casos que pasan desapercibidos.

Desde la perspectiva de gestión de proyectos, la detección temprana de posibles retrasos resulta más valiosa que una clasificación excesivamente conservadora que deje sin identificar actividades problemáticas.

3.9.4 F1 Score

El F1 Score corresponde a una medida que combina Precision y Recall en un único indicador.

Esta métrica permite evaluar el equilibrio entre la capacidad de detección de retrasos y la confiabilidad de las predicciones realizadas por el modelo.

Su utilización resulta especialmente útil cuando se busca balancear ambos objetivos y evitar depender exclusivamente de una sola métrica de evaluación.

3.9.5 ROC-AUC

La métrica ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) mide la capacidad global de un modelo para diferenciar entre actividades retrasadas y no retrasadas.

A diferencia de otras métricas que dependen de un umbral específico de clasificación, el ROC-AUC se calcula utilizando las probabilidades generadas por el modelo y permite evaluar su capacidad de discriminación de forma independiente a la regla de decisión adoptada.

Valores más altos de ROC-AUC indican una mejor capacidad para distinguir entre ambas clases.

3.9.6 Matriz de confusión

La matriz de confusión constituye una herramienta fundamental para analizar en detalle los resultados de clasificación obtenidos por los modelos.

Esta matriz permite identificar:

- Verdaderos positivos (actividades retrasadas correctamente identificadas).
- Verdaderos negativos (actividades no retrasadas correctamente identificadas).
- Falsos positivos (actividades clasificadas como retrasadas sin estarlo realmente).
- Falsos negativos (actividades retrasadas que no fueron detectadas por el modelo).

El análisis conjunto de estos resultados permite comprender mejor los tipos de errores cometidos por cada algoritmo y evaluar su utilidad práctica dentro del contexto de la gestión de cronogramas.

3.9.7 Selección del modelo final

Una vez calculadas las métricas de evaluación para cada algoritmo, se realizó una comparación integral de resultados con el propósito de identificar la alternativa más adecuada para los objetivos de la investigación.

La selección del modelo final consideró no solamente el desempeño global, sino también su capacidad para detectar actividades retrasadas, aspecto especialmente importante desde una perspectiva de gestión preventiva del cronograma.

Los resultados obtenidos y el proceso de comparación entre modelos se presentan y analizan en detalle en el Capítulo 4.

CAPÍTULO 4. RESULTADOS

4.1 Descripción del conjunto de datos

La presente investigación se desarrolló utilizando información histórica proveniente de cronogramas de proyectos gestionados mediante Primavera P6. Durante la etapa de preparación de datos se consolidó información correspondiente a diez proyectos diferentes, obteniéndose una base inicial compuesta por 17.679 actividades.

Posteriormente, se seleccionaron únicamente las actividades con estado Completed, debido a que estas permitían determinar de forma objetiva la existencia o no de retrasos mediante la comparación entre las fechas reales de finalización y las fechas establecidas en la línea base del cronograma.

Como resultado de este proceso se identificaron 8.877 actividades completadas. Luego de excluir los registros que no disponían de información válida para la construcción de la variable objetivo, se obtuvo un conjunto final de 8.092 actividades utilizadas en el análisis y el modelado predictivo.

La variable objetivo fue definida mediante la creación de la variable binaria Delay_Risk, la cual clasifica las actividades de la siguiente manera:

- Delay_Risk = 1: actividad retrasada.
- Delay_Risk = 0: actividad no retrasada.

La distribución obtenida mostró que 5.533 actividades fueron clasificadas como no retrasadas, mientras que 2.559 actividades fueron clasificadas como retrasadas.

En términos relativos, la distribución de clases fue:

- 68,4 % de actividades no retrasadas.
- 31,6 % de actividades retrasadas.

Esta distribución indica un nivel adecuado de representación para ambas clases, permitiendo desarrollar modelos supervisados de clasificación sin requerir técnicas adicionales de balanceo de datos.

Las variables predictoras utilizadas durante la investigación correspondieron a características estructurales del cronograma:

- Duración original de la actividad.

- Número de predecesores.
- Número de sucesores.

Estas variables fueron seleccionadas debido a su disponibilidad consistente en todos los cronogramas analizados y a su capacidad para representar la estructura de las actividades dentro de la red lógica de planificación.

La Tabla 9 resume las principales características del conjunto de datos utilizado en la investigación.

Indicador	Valor
Actividades originales	17.679
Actividades Completed	8.877
Actividades utilizadas para modelado	8.092
Actividades no retrasadas	5.533
Actividades retrasadas	2.559
Variables predictoras	3

Table 9 Caracteristicas del conjunto de Datos

4.2 Resultados del Análisis Exploratorio de Datos

Con el fin de comprender las características del conjunto de datos y explorar posibles patrones asociados al riesgo de retraso, se realizó un Análisis Exploratorio de Datos (EDA) sobre las variables seleccionadas para la investigación.

Este proceso permitió estudiar la distribución de la variable objetivo, analizar el comportamiento de las variables predictoras y examinar las relaciones existentes entre dichas variables y el riesgo de retraso.

4.2.1 Distribución del riesgo de retraso

La distribución de la variable objetivo mostró que aproximadamente el 68,4 % de las actividades no presentaron retraso, mientras que el 31,6 % fueron clasificadas como retrasadas.

Esta distribución evidencia que la ocurrencia de retrasos constituye un fenómeno frecuente dentro de los cronogramas analizados, lo que justifica la aplicación de técnicas predictivas orientadas a su identificación temprana.

Asimismo, la proporción observada permitió disponer de una representación adecuada de ambas clases para el entrenamiento y evaluación de los modelos de clasificación.

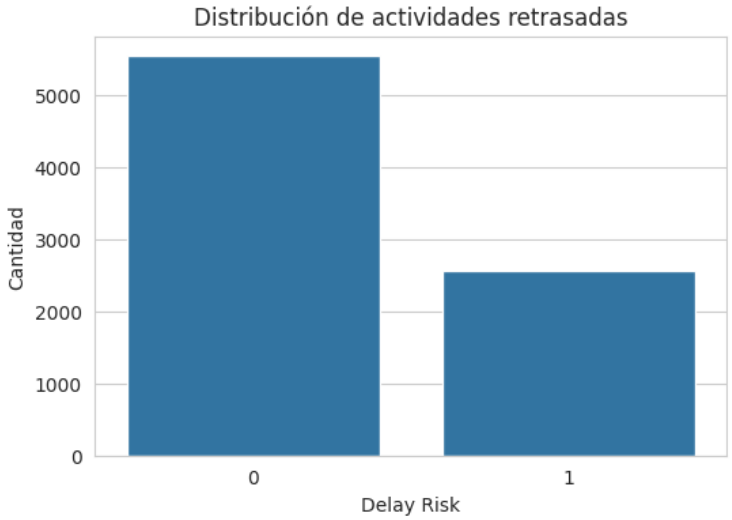


Ilustración 1 Distribución de actividades retrasadas

4.2.2 Análisis de la duración de las actividades

El análisis descriptivo de la duración original de las actividades mostró una distribución fuertemente asimétrica hacia la derecha.

Los principales estadísticos obtenidos fueron:

- Media: 24,88 días.
- Mediana: 10 días.
- Percentil 90: 60 días.
- Percentil 95: 101 días.
- Percentil 99: 210 días.
- Valor máximo: 728 días.

Estos resultados indican que la mayoría de las actividades poseen duraciones relativamente cortas, mientras que existe un grupo reducido de actividades de larga duración que generan una cola extensa en la distribución.

Este comportamiento es consistente con la naturaleza de los cronogramas de proyectos, donde normalmente coexisten actividades operativas de corta duración con actividades de planificación, construcción o gestión que pueden extenderse durante períodos considerablemente más largos.

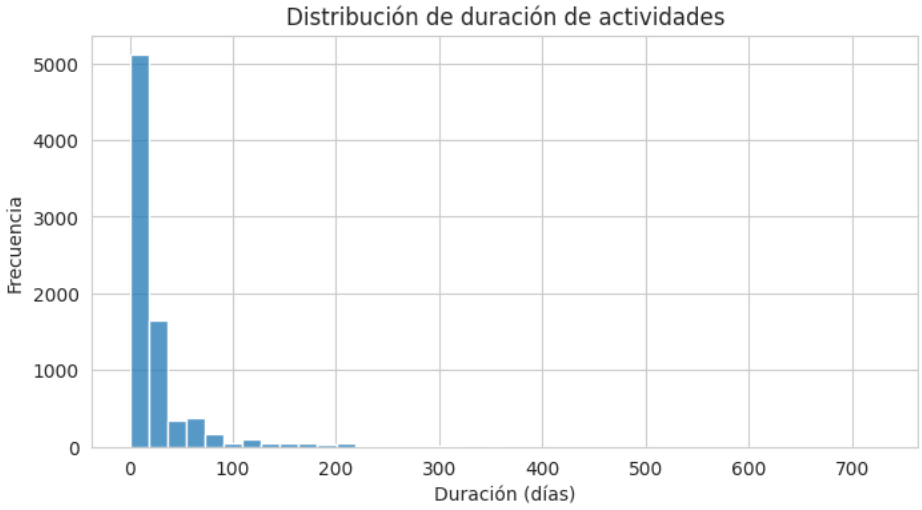


Ilustración 2 Distribución de duración de actividades

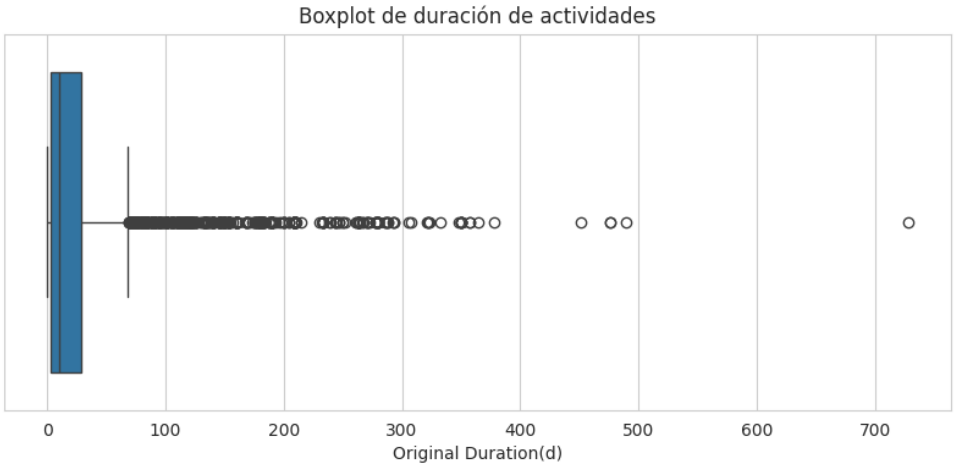


Ilustración 3 Boxplot de duración

4.2.3 Duración y riesgo de retraso

Con el propósito de explorar la relación entre la duración de las actividades y la ocurrencia de retrasos, se compararon las distribuciones de duración para las actividades retrasadas y no retrasadas.

El análisis gráfico mostró que ambos grupos presentan distribuciones similares y con una importante superposición entre sus rangos de valores.

Aunque se observan actividades retrasadas de larga duración, la duración por sí sola no parece generar una separación clara entre actividades retrasadas y no retrasadas.

Este hallazgo sugiere que la duración individual de una actividad no constituye un indicador suficiente para explicar completamente la ocurrencia de retrasos.

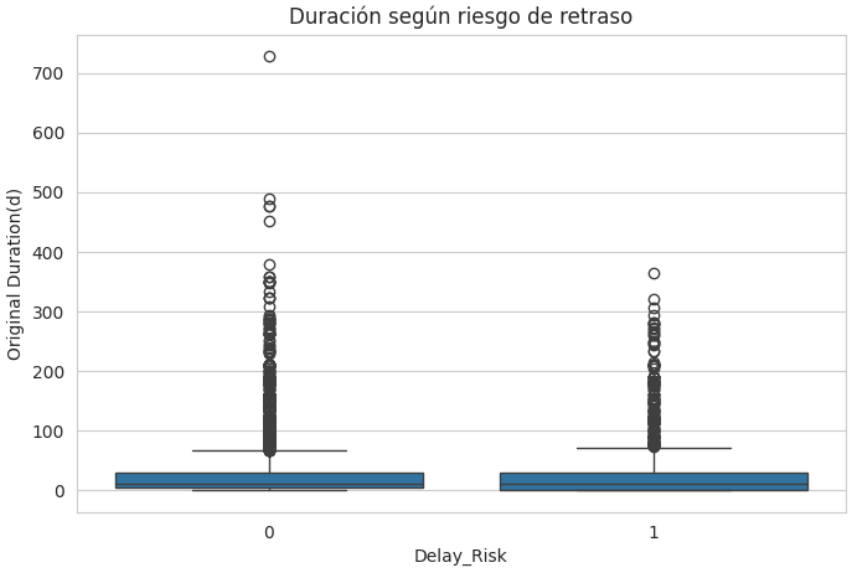


Ilustración 4 Duración según riesgo de retraso

4.2.4 Número de predecesores y riesgo de retraso

El análisis del número de predecesores mostró resultados particularmente relevantes para la investigación.

Las actividades clasificadas como retrasadas tendieron a presentar una mayor cantidad de predecesores en comparación con aquellas que cumplieron la planificación original.

Este comportamiento sugiere que las actividades con mayores dependencias previas se encuentran más expuestas a retrasos acumulados provenientes de otras actividades dentro de la red lógica del cronograma.

Desde una perspectiva de gestión de proyectos, este resultado es coherente con el hecho de que una actividad altamente dependiente puede verse afectada por desviaciones ocurridas en etapas anteriores de la planificación.

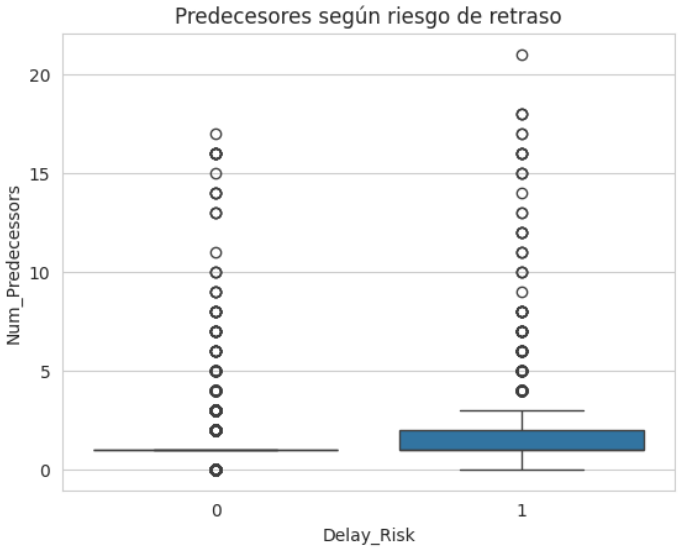


Ilustración 5 Predecesores según riesgo de retraso

4.2.5 Número de sucesores y riesgo de retraso

El análisis de la variable correspondiente al número de sucesores mostró diferencias menos pronunciadas entre actividades retrasadas y no retrasadas.

Si bien se observaron ciertas variaciones entre ambos grupos, la capacidad de esta variable para discriminar visualmente el riesgo de retraso fue inferior a la observada para el número de predecesores.

No obstante, debido a que representa una característica estructural relevante dentro de la red de planificación, la variable fue conservada para las etapas posteriores de modelado.

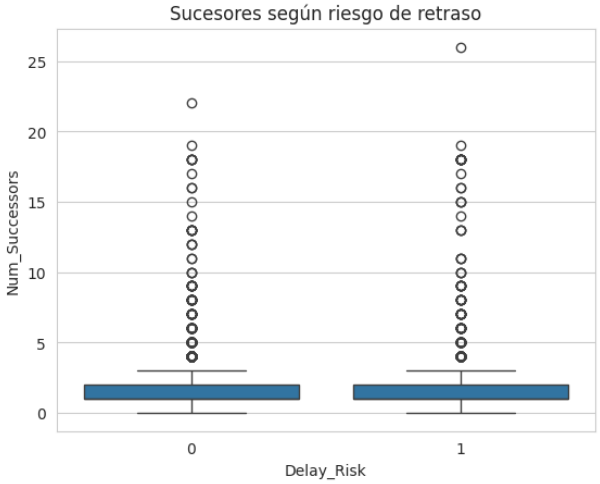


Ilustración 6 Sucesores según riesgo de retraso

4.2.6 Análisis de correlación

Con el propósito de explorar asociaciones lineales entre las variables analizadas, se construyó una matriz de correlación utilizando las variables predictoras y la variable objetivo.

Los resultados mostraron que:

- Delay_Risk y Duration presentaron una correlación de -0,016.
- Delay_Risk y Num_Predecessors presentaron una correlación de 0,120.
- Delay_Risk y Num_Successors presentaron una correlación de 0,049.

La correlación más elevada fue observada para el número de predecesores, lo que refuerza la evidencia obtenida mediante el análisis gráfico.

Aunque ninguna de las variables presentó una correlación fuerte con la variable objetivo, los resultados sugieren la existencia de relaciones más complejas que podrían ser identificadas mediante algoritmos de Machine Learning capaces de capturar patrones no lineales.

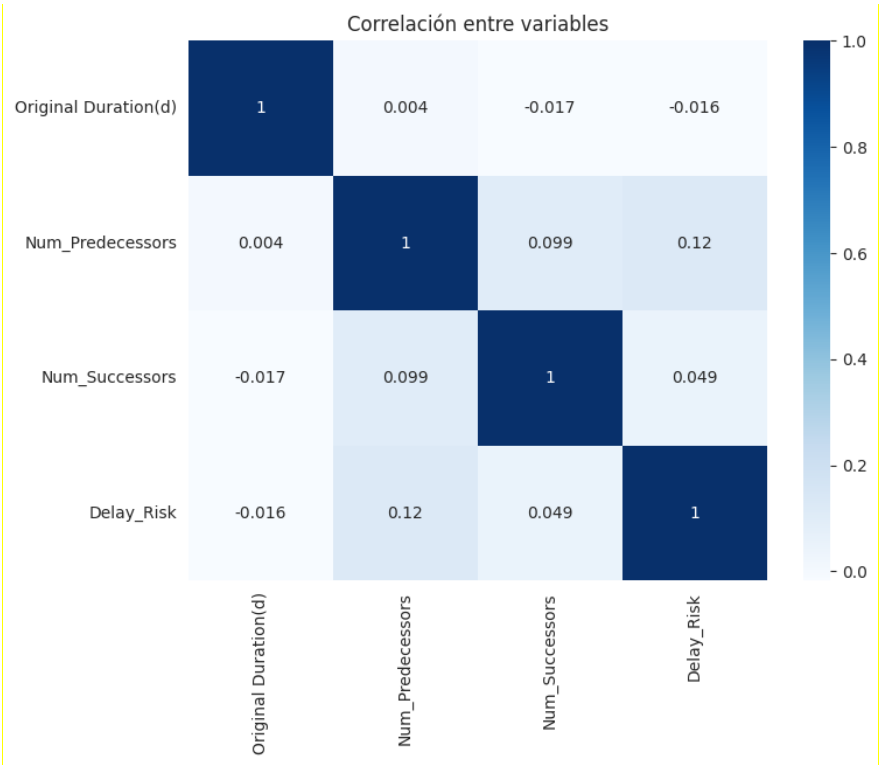


Ilustración 7 Heatmap de correlaciones

4.2.7 Principales hallazgos del análisis exploratorio

Los resultados obtenidos durante el análisis exploratorio permiten destacar los siguientes hallazgos:

- El conjunto de datos presenta una distribución adecuada para el desarrollo de modelos de clasificación.
- La duración de las actividades muestra una distribución asimétrica con presencia de valores extremos reales.
- El número de predecesores constituye la variable con mayor asociación observada respecto al riesgo de retraso.
- Las correlaciones lineales encontradas son relativamente bajas, sugiriendo la existencia de relaciones más complejas entre las variables.
- Las variables estructurales del cronograma contienen información potencialmente útil para la construcción de modelos predictivos.

En función de estos resultados, se procedió a la etapa de modelado con el propósito de evaluar la capacidad de distintos algoritmos para identificar patrones asociados al riesgo de retraso en actividades.

4.3 Resultados de la Regresión Logística

La regresión logística fue utilizada como modelo base de clasificación con el propósito de evaluar la capacidad de las variables estructurales del cronograma para predecir el riesgo de retraso en actividades.

El modelo fue entrenado utilizando las variables duración original de la actividad, número de predecesores y número de sucesores. Posteriormente, su desempeño fue evaluado sobre el conjunto de prueba mediante diferentes métricas de clasificación.

4.3.1 Desempeño general del modelo

Los resultados obtenidos por la regresión logística se presentan en la Tabla 10.

Métrica	Valor
Accuracy	68,75 %
Precision	60,71 %
Recall	3,32 %
F1 Score	6,30 %
ROC-AUC	0,5702

Table 10 Resultados de la Regresión Logística

Los resultados muestran que el modelo alcanzó un Accuracy de 68,75 %. Sin embargo, esta métrica debe interpretarse con precaución debido a la distribución de clases presente en el conjunto de datos.

La principal limitación observada fue la baja capacidad del modelo para identificar actividades retrasadas, evidenciada por un Recall de únicamente 3,32 %. Esto significa que la gran mayoría de las actividades que efectivamente presentaron retraso no fueron detectadas correctamente por el algoritmo.

Asimismo, el valor de ROC-AUC obtenido fue de 0,5702, indicando una capacidad de discriminación apenas superior al azar.

4.3.2 Matriz de confusión

Con el propósito de analizar el comportamiento del modelo con mayor detalle, se construyó una matriz de confusión utilizando las predicciones realizadas sobre el conjunto de prueba.

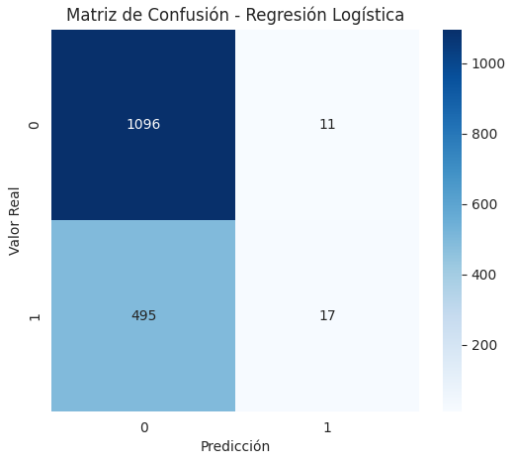


Ilustración 8 Matriz de Confusión - Regresión Logística

La matriz de confusión permitió identificar una elevada cantidad de actividades retrasadas clasificadas incorrectamente como actividades sin retraso.

Este comportamiento explica los bajos valores obtenidos para Recall y F1 Score, evidenciando que el modelo presenta dificultades para detectar actividades con riesgo de retraso utilizando únicamente relaciones lineales entre las variables analizadas.

4.3.3 Interpretación de los resultados

La regresión logística cumplió adecuadamente su función como modelo base de comparación. No obstante, los resultados obtenidos sugieren que las relaciones entre las variables estructurales del cronograma y el riesgo de retraso no son predominantemente lineales.

La baja capacidad de detección observada indica que este algoritmo no logra capturar de manera efectiva los patrones presentes en los datos históricos.

En consecuencia, los resultados justifican la evaluación de modelos más complejos capaces de representar relaciones no lineales entre variables, tales como árboles de decisión y algoritmos basados en ensambles.

4.3.4 Conclusión de la Regresión Logística

La regresión logística obtuvo un desempeño limitado para la predicción del riesgo de retraso en actividades, particularmente en términos de capacidad de detección de actividades retrasadas.

Aunque el modelo permitió identificar cierta relación entre las variables estructurales y la variable objetivo, los resultados obtenidos muestran que la estructura del problema requiere enfoques capaces de representar patrones más complejos.

Por esta razón, se procedió a la evaluación de modelos basados en árboles de decisión con el objetivo de mejorar el desempeño predictivo y explorar relaciones no lineales presentes en los datos.

4.4 Resultados del Árbol de Decisión

Con el propósito de evaluar la capacidad de los modelos basados en reglas para identificar actividades con riesgo de retraso, se implementó un Árbol de Decisión utilizando las variables estructurales seleccionadas para la investigación.

A diferencia de la Regresión Logística, este algoritmo permite representar relaciones no lineales entre las variables y generar reglas de clasificación de fácil interpretación.

4.4.1 Desempeño general del modelo

Los resultados obtenidos se presentan en la Tabla 11.

Métrica	Valor
Accuracy	69,36 %
Precision	54,35 %
Recall	19,53 %
F1 Score	28,74 %
ROC-AUC	0,6304

Table 11 Desempeño general Árbol de Decisión

En comparación con la Regresión Logística, el Árbol de Decisión mostró una mejora significativa en la capacidad para detectar actividades retrasadas. El Recall aumentó de 3,32 % a 19,53 %, mientras que el F1 Score pasó de 0,0630 a 0,2874.

Asimismo, el ROC-AUC aumentó de 0,5702 a 0,6304, indicando una mejor capacidad de discriminación entre actividades retrasadas y no retrasadas.

4.4.2 Interpretación de los resultados

Los resultados sugieren que las relaciones existentes entre las variables estructurales y el riesgo de retraso presentan un comportamiento no lineal que puede ser aprovechado por algoritmos basados en árboles.

Aunque la mejora observada es considerable respecto al modelo base, el nivel de detección de actividades retrasadas continúa siendo limitado para una aplicación práctica orientada a la prevención de retrasos.

Por esta razón, se procedió a evaluar modelos más robustos basados en conjuntos de árboles.

4.5 Resultados de Random Forest

Con el propósito de incrementar la estabilidad y capacidad predictiva del Árbol de Decisión, se implementó el algoritmo Random Forest.

Este modelo combina múltiples árboles de decisión y genera una clasificación final mediante mecanismos de agregación, reduciendo la sensibilidad del modelo a variaciones particulares en los datos de entrenamiento.

4.5.1 Desempeño general del modelo

Los resultados obtenidos se presentan en la siguiente tabla .

Métrica	Valor
Accuracy	68,68 %
Precision	51,28 %
Recall	19,53 %
F1 Score	28,29 %
ROC-AUC	0,6396

Table 12 Desempeño general Random Forest

Los resultados muestran un desempeño muy similar al obtenido por el Árbol de Decisión.

Si bien el Accuracy disminuyó ligeramente, el modelo alcanzó un ROC-AUC de 0,6396, superior al obtenido por el Árbol de Decisión (0,6304).

4.5.2 Interpretación de los resultados

La principal ventaja observada en Random Forest fue su mayor capacidad de discriminación global, reflejada en el incremento del ROC-AUC.

Sin embargo, la capacidad para detectar actividades retrasadas permaneció prácticamente sin cambios respecto al Árbol de Decisión, manteniendo un Recall de 19,53 %.

Estos resultados indican que, aunque Random Forest logra representar de mejor manera las relaciones presentes en los datos, aún existen dificultades para detectar una proporción significativa de actividades retrasadas utilizando el umbral de clasificación tradicional.

4.6 Resultados de XGBoost

Con el objetivo de evaluar un algoritmo de mayor capacidad predictiva, se implementó XGBoost utilizando el mismo conjunto de entrenamiento y prueba utilizado en los modelos anteriores.

Inicialmente se evaluó el desempeño utilizando el umbral de clasificación convencional de 0,50.

4.6.1 XGBoost con umbral 0.50

Los resultados obtenidos se presentan en la siguiente Tabla

Métrica	Valor
Accuracy	68,87 %
Precision	53,12 %
Recall	13,28 %
F1 Score	21,25 %
ROC-AUC	0,6587

Table 13 Resultados de XGBoost con umbral 0.5

El modelo obtuvo el mejor valor de ROC-AUC entre todos los algoritmos evaluados, alcanzando 0,6587.

Sin embargo, el Recall obtenido fue relativamente bajo, indicando que una proporción importante de actividades retrasadas continuaba sin ser detectada.

4.6.2 Optimización del umbral de clasificación

Debido a que el objetivo principal de la investigación consiste en identificar actividades con riesgo de retraso, se realizó una evaluación de diferentes umbrales de clasificación con el fin de incrementar la capacidad de detección del modelo.

Como resultado de este proceso se seleccionó un umbral de clasificación de 0,30, el cual permitió mejorar significativamente el Recall.

Los resultados obtenidos se presentan en la siguiente tabla.

Métrica	Valor
Accuracy	60,96 %
Precision	42,56 %
Recall	66,99 %
F1 Score	52,05 %

Table 14 Resultados de XGBoost con umbral optimizado

El ajuste del umbral permitió incrementar el Recall desde 13,28 % hasta 66,99 %, representando una mejora sustancial en la capacidad de detección de actividades retrasadas.

Asimismo, el F1 Score aumentó de 21,25 % a 52,05 %, convirtiéndose en el mejor resultado obtenido entre todos los modelos evaluados.

4.6.3 Interpretación de los resultados

Los resultados muestran que XGBoost posee la mejor capacidad para capturar patrones estructurales asociados al riesgo de retraso.

La optimización del umbral permitió adaptar el modelo a los objetivos específicos de la investigación, priorizando la detección de actividades potencialmente retrasadas por encima del desempeño global medido únicamente mediante Accuracy.

En consecuencia, XGBoost fue seleccionado como el modelo final para la investigación.

4.7 Comparación de modelos

Con el propósito de identificar el algoritmo con mejor desempeño para la predicción del riesgo de retraso, se realizó una comparación entre los diferentes modelos evaluados durante la investigación.

La Tabla siguiente presenta los resultados obtenidos para cada algoritmo utilizando el conjunto de prueba.

Modelo	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Regresión Logística	68,75 %	60,71 %	3,32 %	0,0630	0,5702
Árbol de Decisión	69,36 %	54,35 %	19,53 %	0,2874	0,6304
Random Forest	68,68 %	51,28 %	19,53 %	0,2829	0,6396
XGBoost (0.50)	68,87 %	53,12 %	13,28 %	0,2125	0,6587
XGBoost (0.30)	60,96 %	42,56 %	66,99 %	0,5205	0,6587

Table 15 Comparación de modelos

Los resultados evidencian que los modelos basados en árboles superaron claramente a la Regresión Logística en términos de capacidad predictiva.

La Regresión Logística obtuvo el menor desempeño global, particularmente en la detección de actividades retrasadas, alcanzando un Recall de apenas 3,32 %. Esto indica que el modelo fue incapaz de identificar una proporción significativa de las actividades que efectivamente presentaron retraso.

Por su parte, el Árbol de Decisión y Random Forest mostraron mejoras importantes en Recall y F1 Score, demostrando una mejor capacidad para capturar relaciones no lineales entre las variables estructurales del cronograma y el riesgo de retraso.

Finalmente, XGBoost obtuvo el mejor desempeño en términos de capacidad de discriminación, alcanzando un ROC-AUC de 0,6587. Posteriormente, mediante la optimización del umbral de clasificación, logró incrementar significativamente la detección de actividades retrasadas, obteniendo el mejor Recall y el mejor F1 Score de todos los modelos evaluados.

4.8 Selección del modelo final

La selección del modelo final se realizó considerando tanto el desempeño estadístico como la utilidad práctica dentro del contexto de la gestión de cronogramas.

Si bien los modelos evaluados presentaron niveles de Accuracy relativamente similares, esta métrica no fue utilizada como criterio principal de selección debido a que el objetivo de la investigación consiste en identificar actividades con riesgo de retraso y no únicamente maximizar el número total de clasificaciones correctas.

En este sentido, las métricas Recall y F1 Score adquirieron una importancia especial, ya que permiten evaluar la capacidad del modelo para detectar actividades retrasadas y mantener un equilibrio adecuado entre detección y precisión.

Los resultados obtenidos evidenciaron que XGBoost presentó la mejor capacidad global de discriminación, obteniendo un ROC-AUC de 0,6587. Adicionalmente, la optimización del umbral de clasificación permitió incrementar significativamente la detección de actividades retrasadas.

Como resultado, el modelo XGBoost con umbral de clasificación de 0,30 alcanzó:

- Recall = 66,99 %
- Precision = 42,56 %
- F1 Score = 52,05 %

Estos resultados representan el mejor equilibrio observado entre capacidad de detección y desempeño general entre todos los modelos evaluados.

Por esta razón, XGBoost con umbral de clasificación de 0,30 fue seleccionado como el modelo final de la investigación.

4.8.1 Curva ROC del modelo seleccionado

La curva ROC del modelo seleccionado presenta un área bajo la curva (ROC-AUC) de 0,659.

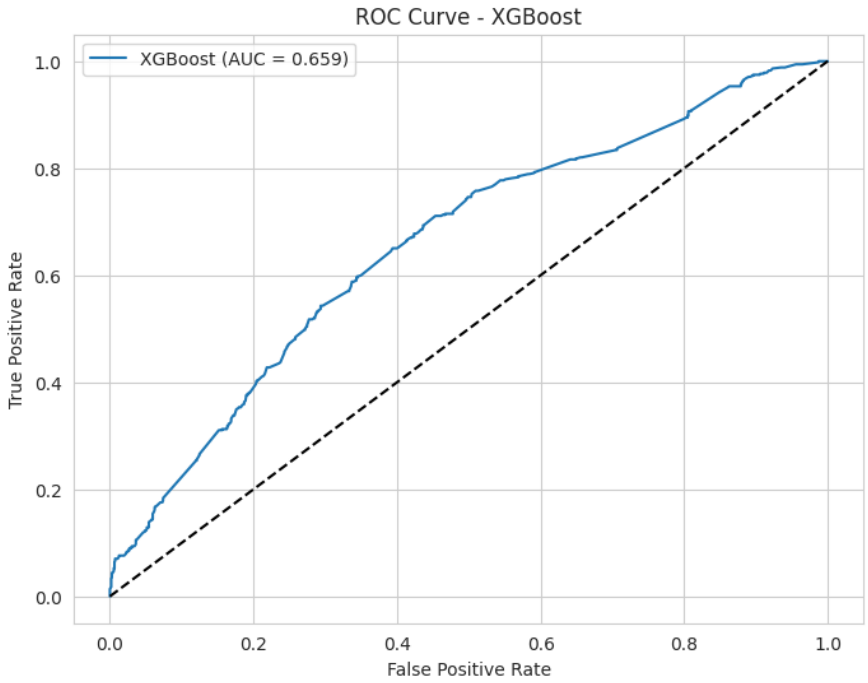


Ilustración 9 Curva ROC - XGBoost

El valor obtenido indica que el modelo posee una capacidad de discriminación superior al azar para distinguir entre actividades retrasadas y no retrasadas.

Aunque el nivel de discriminación alcanzado no puede considerarse excelente, los resultados demuestran que las variables estructurales del cronograma contienen información útil para la predicción del riesgo de retraso.

4.8.2 Importancia de variables

El análisis de importancia de variables permitió identificar cuáles de las características estructurales utilizadas tuvieron una mayor contribución en las predicciones realizadas por el modelo.

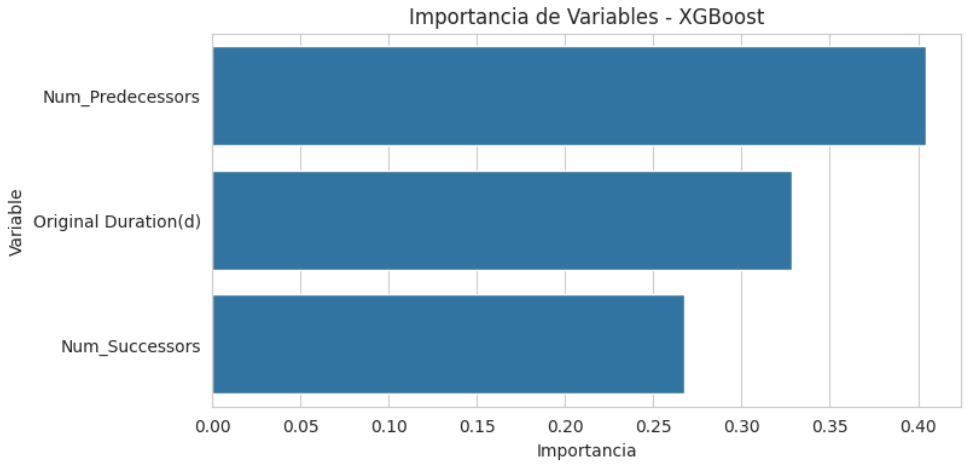


Ilustración 10 Importancia de Variables - XGBoost

Los resultados muestran el siguiente orden de importancia:

1. Número de predecesores.
2. Duración original de la actividad.
3. Número de sucesores.

El número de predecesores fue la variable más influyente, indicando que las dependencias existentes dentro del cronograma constituyen un factor relevante para la identificación de actividades con riesgo de retraso.

Este resultado es consistente con los hallazgos observados durante el análisis exploratorio de datos y refuerza la hipótesis de que la estructura de relaciones del cronograma contiene información valiosa para la predicción.

4.8.3 Matriz de confusión del modelo seleccionado

La matriz de confusión del modelo XGBoost con umbral de clasificación de 0,30 mostró los siguientes resultados:

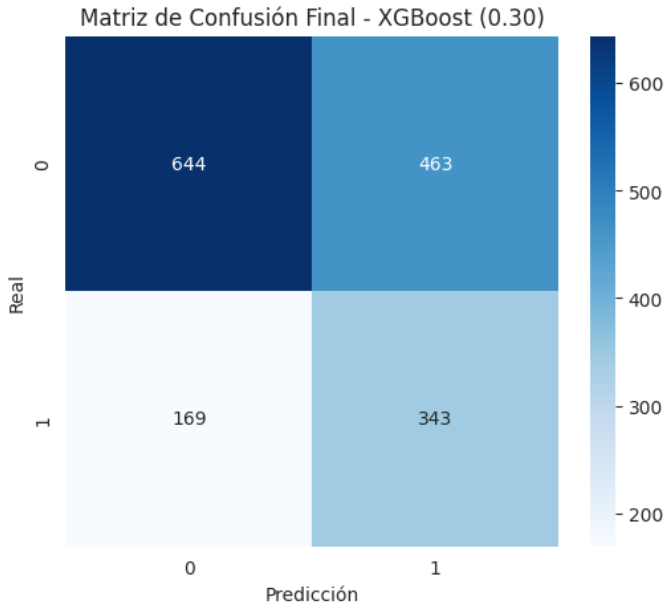


Ilustración 11 Matriz de Confusión Final

	Predicción No Retraso	**Predicción Retraso**
No retraso	644	463
Retraso	169	343

Table 16 Matriz de confusión del modelo seleccionado

El modelo logró identificar correctamente 343 actividades retrasadas de un total de 512 actividades que efectivamente presentaron retraso.

Este resultado representa un Recall de 66,99 %, indicando que aproximadamente dos de cada tres actividades retrasadas fueron detectadas correctamente por el modelo.

Asimismo, se registraron 463 falsos positivos, correspondientes a actividades identificadas como riesgosas que finalmente no presentaron retraso.

Desde la perspectiva de gestión de proyectos, esta configuración resulta adecuada debido a que el objetivo principal consiste en alertar oportunamente sobre actividades potencialmente problemáticas. En este contexto, el costo asociado a revisar actividades sospechosas suele ser significativamente menor que el impacto generado por no detectar actividades que terminarán afectando el cronograma.

4.9 Aplicación práctica del modelo mediante Power BI

Con el propósito de demostrar una posible aplicación operativa de los resultados obtenidos, el modelo seleccionado fue aplicado a actividades con estado Not Started e In Progress, generando una probabilidad de retraso para cada actividad analizada. Las probabilidades estimadas fueron posteriormente integradas en un dashboard desarrollado en Power BI, permitiendo visualizar de forma dinámica la distribución del riesgo y facilitar la identificación de actividades potencialmente críticas dentro de los cronogramas analizados. Esta aplicación tiene carácter demostrativo y no constituye un proceso adicional de validación estadística del modelo.

4.9.1 Predicción de actividades futuras

Una vez seleccionado el modelo XGBoost con umbral de clasificación de 0,30, se aplicó dicho modelo sobre un conjunto de 8.802 actividades que aún no habían finalizado su ejecución.

Para cada actividad se calculó una probabilidad estimada de retraso (Probability_Delay) y posteriormente se clasificó en tres niveles de riesgo:

- Riesgo Alto.
- Riesgo Medio.
- Riesgo Bajo.

Esta clasificación permitió transformar los resultados generados por el modelo en información fácilmente interpretable para fines de planificación y control.

4.9.2 Dashboard de visualización

El dashboard desarrollado en Power BI presenta indicadores y visualizaciones orientadas a apoyar el análisis de riesgo de las actividades.

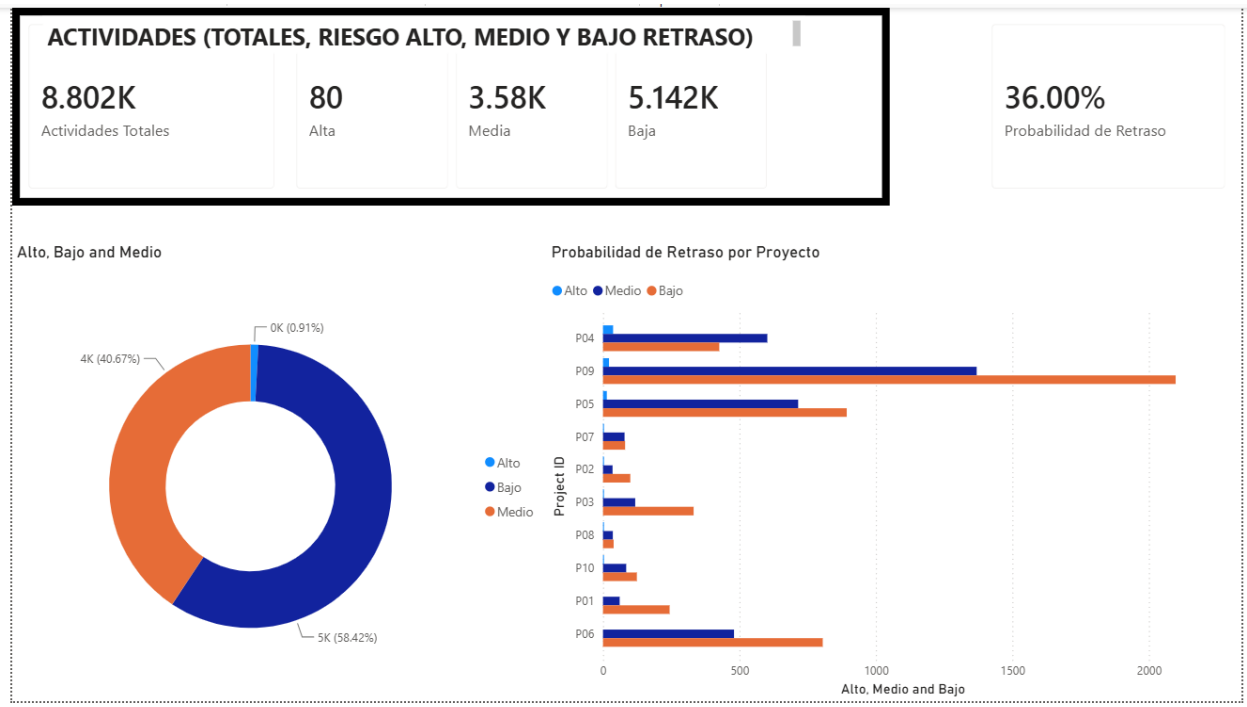


Ilustración 12 Dashboard Power BI

Los indicadores principales muestran:

- 8.802 actividades analizadas.
- 80 actividades clasificadas como riesgo alto.
- 3580 actividades clasificadas como riesgo medio.
- 5.142 actividades clasificadas como riesgo bajo.
- Probabilidad promedio de retraso del 36 %.

Estos resultados permiten obtener una visión general del comportamiento del conjunto de actividades evaluadas.

4.9.3 Distribución de niveles de riesgo

La distribución de actividades por nivel de riesgo muestra que la mayor parte de las actividades se concentra en las categorías de riesgo bajo y riesgo medio.

Específicamente:

- 58,42 % corresponden a riesgo bajo.
- 40,67 % corresponden a riesgo medio.
- 0,91 % corresponden a riesgo alto.

La baja proporción de actividades clasificadas como riesgo alto evidencia que el modelo identifica un grupo relativamente reducido de actividades que requieren atención prioritaria.

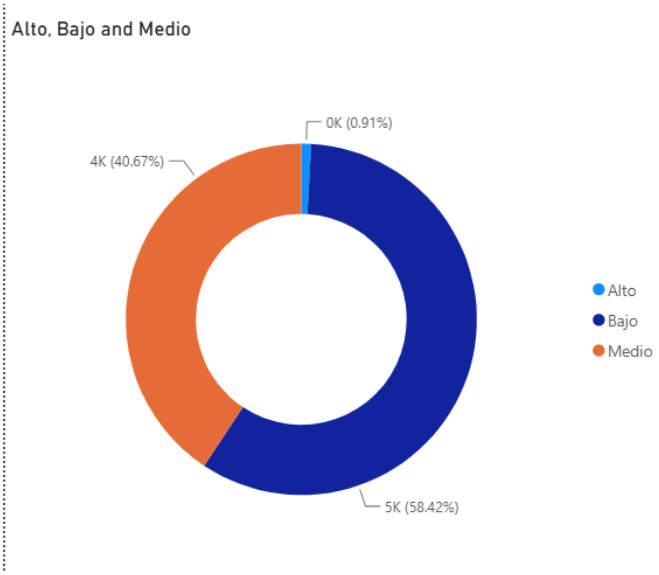


Ilustración 13 Distribución de niveles de riesgo

4.9.4 Análisis por proyecto

El dashboard también permite analizar la distribución del riesgo por proyecto.

Los resultados muestran diferencias en la probabilidad promedio de retraso entre los proyectos analizados. Entre los valores más altos destacan:

- P04: 0,44
- P08: 0,39
- P07: 0,39

Mientras que los valores más bajos corresponden a:

- P01: 0,29
- P06: 0,30
- P02: 0,33

Estos resultados permiten identificar proyectos donde la estructura de las actividades presenta una mayor exposición potencial al riesgo de retraso.

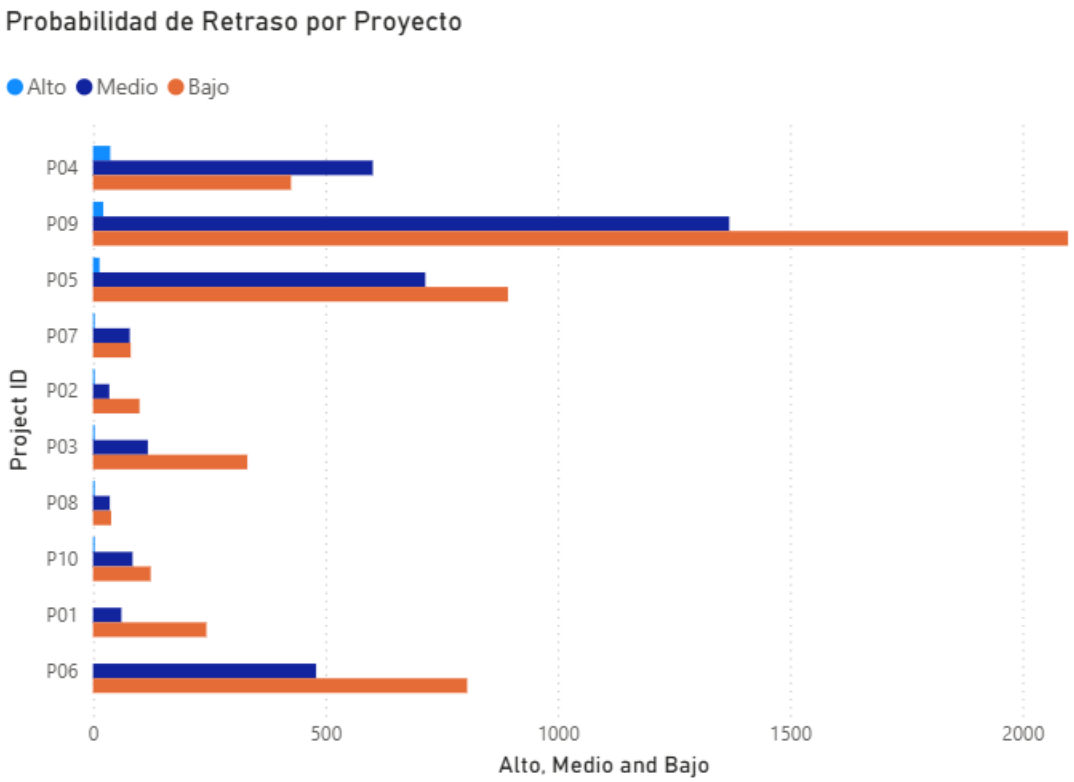


Ilustración 14 Riesgo promedio por proyecto

4.9.5 Utilidad para la gestión de proyectos

La integración del modelo predictivo con herramientas de visualización permite transformar los resultados obtenidos en información útil para la toma de decisiones.

Desde una perspectiva práctica, el dashboard facilita:

- La identificación temprana de actividades potencialmente críticas.
- La priorización de actividades con mayor riesgo.
- El monitoreo de proyectos con mayor exposición al retraso.
- El apoyo a procesos de planificación y seguimiento basados en datos.

De esta manera, los modelos desarrollados dejan de ser únicamente herramientas analíticas y se convierten en un mecanismo de apoyo para la gestión preventiva de cronogramas.

4.10 Síntesis del capítulo

Los resultados presentados en este capítulo permitieron evaluar la capacidad de diferentes modelos de Machine Learning para predecir el riesgo de retraso en actividades utilizando exclusivamente variables estructurales del cronograma.

El análisis exploratorio evidenció que las variables seleccionadas contienen información relevante para el estudio del fenómeno, destacando particularmente la importancia del número de predecesores dentro de la red lógica de planificación.

La comparación de modelos mostró que los algoritmos basados en árboles superaron el desempeño de la Regresión Logística, evidenciando la existencia de relaciones no lineales entre las variables analizadas y el riesgo de retraso.

Entre los modelos evaluados, XGBoost obtuvo la mejor capacidad global de discriminación (ROC-AUC = 0,659). Adicionalmente, la optimización del umbral de clasificación permitió incrementar significativamente la detección de actividades retrasadas, alcanzando un Recall de 66,99 % y un F1 Score de 52,05 %, resultados que motivaron su selección como modelo final de la investigación.

Finalmente, la implementación de un dashboard en Power BI demostró la aplicabilidad práctica de los resultados obtenidos, permitiendo transformar las predicciones generadas por el modelo en información útil para la identificación temprana de actividades con riesgo potencial de retraso.

CAPÍTULO 5. DISCUSIÓN

5.1 Interpretación general de los resultados

La presente investigación tuvo como objetivo predecir el riesgo de retraso en actividades mediante el uso de técnicas de Ciencia de Datos aplicadas a información histórica proveniente de cronogramas de proyectos. Los resultados obtenidos evidencian que las variables estructurales del cronograma contienen información útil para identificar patrones asociados a la ocurrencia de retrasos.

A partir del análisis realizado sobre 8.092 actividades históricas, fue posible desarrollar modelos predictivos capaces de diferenciar actividades con y sin retraso utilizando únicamente información relacionada con la duración de las actividades y sus relaciones de dependencia dentro de la red de planificación.

Los resultados obtenidos muestran que las relaciones estructurales existentes entre las actividades aportan capacidad predictiva al proceso de clasificación, confirmando que la configuración interna del cronograma contiene señales asociadas al comportamiento futuro de las actividades.

Aunque ninguno de los modelos alcanzó niveles de predicción perfectos, los resultados obtenidos demuestran que es posible identificar patrones relevantes utilizando exclusivamente variables derivadas de la estructura del cronograma, sin necesidad de incorporar información adicional relacionada con recursos, costos o productividad.

5.2 Cumplimiento de la pregunta de investigación

La pregunta de investigación planteada fue:

¿Cómo pueden los modelos de Ciencia de Datos predecir el riesgo de retraso en actividades a partir de patrones estructurales presentes en cronogramas de proyectos?

Los resultados obtenidos permiten responder afirmativamente a esta pregunta.

Los diferentes modelos evaluados lograron identificar relaciones entre las variables estructurales analizadas y la probabilidad de retraso de las actividades. En particular, el modelo XGBoost demostró que dichas variables contienen información suficiente para

generar predicciones superiores al azar y detectar una proporción significativa de actividades retrasadas.

La capacidad alcanzada por el modelo final, reflejada en un ROC-AUC de 0,659 y un Recall de 66,99 %, evidencia que la Ciencia de Datos puede utilizarse como una herramienta útil para apoyar la identificación temprana de actividades con mayor exposición al riesgo de retraso.

En consecuencia, los resultados respaldan la premisa de que los cronogramas contienen información estructural valiosa que puede ser aprovechada mediante técnicas de Machine Learning para mejorar la anticipación de posibles problemas de ejecución.

5.3 Discusión sobre las variables estructurales utilizadas

Uno de los hallazgos más relevantes de la investigación corresponde a la importancia observada en las variables estructurales seleccionadas para el análisis.

El modelo XGBoost identificó al número de predecesores como la variable más influyente dentro del proceso de clasificación. Este resultado sugiere que la cantidad de dependencias previas asociadas a una actividad constituye un factor importante para comprender su exposición al riesgo de retraso.

Desde una perspectiva de gestión de proyectos, este comportamiento resulta coherente debido a que las actividades con múltiples dependencias suelen estar más expuestas a retrasos acumulados provenientes de actividades anteriores. En consecuencia, cualquier desviación ocurrida en etapas previas puede propagarse a través de la red lógica y afectar el desempeño de actividades posteriores.

5.4 Discusión sobre los modelos predictivos

La comparación realizada entre los diferentes algoritmos permitió evaluar la capacidad de distintos enfoques de Machine Learning para identificar actividades con riesgo de retraso a partir de variables estructurales del cronograma.

Los resultados obtenidos evidenciaron diferencias importantes entre los modelos evaluados. La Regresión Logística presentó el desempeño más limitado, particularmente en términos de capacidad de detección de actividades retrasadas. Aunque alcanzó un nivel de Accuracy similar al de los demás modelos, su Recall de 3,32 % demostró que era incapaz

de identificar una proporción significativa de las actividades que efectivamente presentaron retraso.

Este comportamiento sugiere que las relaciones existentes entre las variables estructurales y la ocurrencia de retrasos no pueden ser explicadas adecuadamente mediante relaciones lineales simples. Por esta razón, el modelo presentó dificultades para capturar la complejidad del fenómeno analizado.

Por otra parte, el Árbol de Decisión y Random Forest mostraron mejoras importantes respecto a la Regresión Logística. Ambos modelos lograron incrementar significativamente la capacidad de detección de actividades retrasadas y alcanzaron valores superiores de F1 Score y ROC-AUC. Estos resultados indican que la incorporación de relaciones no lineales aporta información relevante para comprender el comportamiento de las actividades dentro del cronograma.

Sin embargo, el mejor desempeño global fue alcanzado por XGBoost. Aunque inicialmente el modelo presentó un Recall moderado utilizando el umbral tradicional de clasificación de 0,50, obtuvo el valor más alto de ROC-AUC entre todos los algoritmos evaluados. Este resultado evidencia que el modelo poseía una mayor capacidad para diferenciar entre actividades retrasadas y no retrasadas a partir de las probabilidades generadas.

Los resultados permiten concluir que los modelos basados en árboles de decisión y técnicas de ensamble son más adecuados para este tipo de problema que los enfoques lineales tradicionales. Asimismo, ponen de manifiesto la importancia de evaluar múltiples métricas de desempeño y no depender exclusivamente de Accuracy para seleccionar un modelo predictivo.

5.5 Discusión sobre la optimización del umbral de clasificación

Uno de los hallazgos más relevantes de la investigación fue la influencia que tuvo el umbral de clasificación sobre el desempeño del modelo seleccionado.

Durante la evaluación inicial, XGBoost fue analizado utilizando el umbral convencional de 0,50. Bajo esta configuración, el modelo alcanzó un Recall relativamente bajo, lo que implicaba que una parte importante de las actividades retrasadas no era identificada correctamente.

El análisis de las probabilidades generadas por el modelo mostró que una gran proporción de actividades se concentraba en valores inferiores a 0,50. Como consecuencia, numerosas actividades con riesgo real de retraso eran clasificadas como actividades sin retraso únicamente por no superar el umbral establecido.

Con el propósito de mejorar la capacidad de detección, se realizó una evaluación de distintos umbrales de clasificación. Como resultado, se seleccionó un umbral de 0,30, el cual permitió incrementar significativamente el Recall del modelo hasta alcanzar un valor de 66,99 %.

Este ajuste produjo un incremento en la cantidad de falsos positivos; sin embargo, desde una perspectiva de gestión de proyectos, esta situación resulta aceptable. El costo asociado a revisar actividades clasificadas como riesgosas suele ser considerablemente menor que el impacto generado por no detectar oportunamente actividades que terminarán afectando el cronograma.

Por lo tanto, la optimización del umbral permitió adaptar el modelo a los objetivos específicos de la investigación, priorizando la identificación temprana de actividades potencialmente problemáticas sobre la maximización de la precisión global.

Es importante señalar que la selección del umbral de clasificación se realizó mediante la evaluación de diferentes valores sobre los resultados obtenidos durante el proceso de modelado, priorizando la capacidad de detección de actividades retrasadas. Aunque este procedimiento permitió identificar una configuración alineada con los objetivos de la investigación, futuras investigaciones podrían incorporar conjuntos de validación independientes o esquemas de validación cruzada para seleccionar el umbral de manera más robusta antes de la evaluación final sobre el conjunto de prueba.

5.6 Relación de los resultados con la literatura revisada

Los hallazgos obtenidos en esta investigación son consistentes con la literatura revisada en el Capítulo 2, la cual señala que las técnicas de Machine Learning poseen una capacidad superior a los métodos estadísticos tradicionales para identificar patrones complejos presentes en datos históricos.

De manera similar a lo reportado en investigaciones previas, los modelos basados en árboles mostraron un mejor desempeño que los enfoques lineales, confirmando la existencia de relaciones no lineales entre las variables estructurales del cronograma y el riesgo de retraso.

Asimismo, los resultados respaldan la idea planteada por diversos autores respecto a que las características estructurales de las actividades contienen información valiosa para la predicción del comportamiento futuro de los proyectos. En particular, la importancia observada en el número de predecesores coincide con estudios que destacan la influencia de las dependencias lógicas sobre el desempeño temporal de las actividades.

Sin embargo, la presente investigación aporta evidencia adicional al demostrar que es posible construir modelos predictivos utilizando exclusivamente variables derivadas de la estructura del cronograma, sin recurrir a información relacionada con costos, recursos o productividad.

Este resultado refuerza la utilidad de los cronogramas como fuente de información para el desarrollo de herramientas predictivas aplicables en contextos reales de gestión de proyectos.

La duración original de las actividades ocupó el segundo lugar en importancia. Aunque esta variable no presentó una correlación lineal fuerte con el riesgo de retraso, los modelos de Machine Learning lograron identificar patrones más complejos asociados a su comportamiento.

Finalmente, el número de sucesores también aportó información relevante al modelo, aunque en una magnitud inferior respecto a las otras variables consideradas.

Estos resultados indican que el riesgo de retraso no depende únicamente de la duración de una actividad, sino también de su posición dentro de la estructura general del cronograma.

5.7 Limitaciones de la investigación

Como toda investigación aplicada, el presente estudio presenta ciertas limitaciones que deben ser consideradas al interpretar los resultados obtenidos.

En primer lugar, el análisis se desarrolló utilizando exclusivamente variables estructurales derivadas de los cronogramas. Aunque este enfoque permitió evaluar la capacidad

predictiva de la estructura de las actividades, no se incorporaron variables relacionadas con recursos, productividad, costos, condiciones operativas o factores externos que también pueden influir en la ocurrencia de retrasos.

En segundo lugar, la investigación se realizó utilizando información histórica proveniente de proyectos pertenecientes a una misma organización. Si bien el conjunto de datos analizado presenta una cantidad considerable de actividades, la capacidad de generalización de los resultados podría fortalecerse mediante la incorporación de proyectos provenientes de diferentes organizaciones o sectores económicos.

Otra limitación corresponde al desempeño predictivo alcanzado por el modelo. Aunque XGBoost logró resultados superiores a los obtenidos por los demás algoritmos evaluados, el valor de ROC-AUC obtenido indica que todavía existe una parte importante de la variabilidad del fenómeno que no puede ser explicada únicamente mediante las variables estructurales consideradas en el estudio.

Finalmente, la investigación se centró en la predicción del riesgo de retraso a nivel de actividades individuales. Por esta razón, no se analizaron los posibles efectos acumulativos de los retrasos sobre indicadores globales del proyecto, como fechas de terminación, hitos contractuales o desempeño general del cronograma.

Una limitación adicional corresponde a la estrategia de partición utilizada para entrenamiento y prueba. La división de datos se realizó mediante muestreo estratificado sobre actividades individuales, lo que implica que actividades pertenecientes a un mismo proyecto pudieron estar presentes tanto en el conjunto de entrenamiento como en el conjunto de prueba. Si bien este enfoque permitió evaluar el desempeño predictivo general de los modelos, futuras investigaciones podrían incorporar estrategias basadas en agrupación por proyecto, tales como GroupKFold o Leave-One-Project-Out, con el fin de analizar la capacidad de generalización del modelo frente a proyectos completamente no observados.

5.8 Implicaciones prácticas y académicas

Los resultados obtenidos generan implicaciones tanto para la gestión de proyectos como para futuras investigaciones relacionadas con la aplicación de técnicas de Ciencia de Datos en el análisis de cronogramas.

Desde una perspectiva práctica, la investigación demuestra que los cronogramas contienen información útil para desarrollar modelos predictivos capaces de identificar actividades con mayor probabilidad de retraso. Esto abre la posibilidad de complementar los procesos tradicionales de planificación y control mediante herramientas analíticas orientadas a la detección temprana de riesgos.

Asimismo, la integración del modelo predictivo dentro de un dashboard en Power BI constituye una demostración operativa de los resultados obtenidos, mostrando cómo las predicciones generadas por el modelo pueden transformarse en información accesible para planificadores y responsables de proyectos.

Desde una perspectiva académica, el estudio aporta evidencia sobre la utilidad de variables estructurales del cronograma para la construcción de modelos predictivos. Los resultados obtenidos contribuyen a ampliar el conocimiento existente sobre la aplicación de Machine Learning en gestión de proyectos y ofrecen una metodología reproducible para investigaciones futuras.

Adicionalmente, la comparación de distintos algoritmos bajo un mismo conjunto de datos proporciona información relevante sobre las ventajas y limitaciones de diferentes enfoques predictivos, particularmente en escenarios donde la información disponible se limita a la estructura interna del cronograma.

5.9 Síntesis del capítulo

La discusión desarrollada en este capítulo permitió interpretar los resultados obtenidos durante la investigación y analizar su significado dentro del contexto de la predicción del riesgo de retraso en actividades.

Los hallazgos confirmaron que las variables estructurales del cronograma contienen información relevante para la identificación de patrones asociados al retraso de actividades. En particular, el número de predecesores mostró la mayor contribución dentro

del modelo seleccionado, evidenciando la importancia de las relaciones de dependencia en el comportamiento de las actividades.

La comparación de modelos permitió verificar que los algoritmos basados en árboles presentan un mejor desempeño que los enfoques lineales para este tipo de problema, destacándose XGBoost como la alternativa con mayor capacidad de discriminación entre actividades retrasadas y no retrasadas.

Asimismo, se evidenció que la optimización del umbral de clasificación constituye un factor determinante para adaptar los modelos a las necesidades específicas de la gestión de proyectos, permitiendo incrementar significativamente la capacidad de detección de actividades con riesgo de retraso.

Finalmente, se identificaron las principales limitaciones del estudio y se analizaron las implicaciones prácticas y académicas derivadas de los resultados obtenidos, estableciendo una base sólida para la formulación de conclusiones y recomendaciones presentadas en el capítulo siguiente.

CAPÍTULO 6. CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

La presente investigación tuvo como objetivo predecir el riesgo de retraso en actividades utilizando técnicas de Ciencia de Datos aplicadas a datos históricos de cronogramas de proyectos. Los resultados obtenidos permiten concluir que las variables estructurales del cronograma contienen información relevante para la identificación de patrones asociados a la ocurrencia de retrasos.

La primera conclusión corresponde al cumplimiento del objetivo relacionado con la identificación de patrones estructurales. El análisis exploratorio y los modelos predictivos desarrollados permitieron evidenciar que características como la duración de las actividades, el número de predecesores y el número de sucesores presentan asociaciones con el comportamiento de las actividades dentro del cronograma. En particular, el número de predecesores se identificó como la variable con mayor influencia dentro del modelo seleccionado.

La segunda conclusión se relaciona con el desarrollo de modelos predictivos. Los resultados obtenidos sugieren que es posible construir modelos de clasificación capaces de estimar el riesgo de retraso utilizando únicamente variables derivadas de la estructura del cronograma. Esto aporta evidencia de que la información contenida en los cronogramas puede utilizarse como fuente para la generación de herramientas predictivas orientadas a la gestión de proyectos.

La tercera conclusión corresponde a la comparación de los modelos evaluados. Los algoritmos basados en árboles presentaron un desempeño superior al obtenido por la Regresión Logística, evidenciando que las relaciones entre las variables estructurales y el riesgo de retraso poseen un comportamiento no lineal. Entre los modelos evaluados, XGBoost obtuvo la mejor capacidad de discriminación, alcanzando un valor ROC-AUC de 0,659.

La cuarta conclusión se refiere a la utilidad de la optimización del umbral de clasificación. El ajuste del umbral a un valor de 0,30 permitió incrementar significativamente la capacidad de detección de actividades retrasadas, alcanzando un Recall de 66,99 % y un F1 Score de

52,05 %. Este resultado sugiere que la configuración del criterio de clasificación puede ser tan importante como la selección del algoritmo utilizado.

La quinta conclusión corresponde a la pregunta de investigación. Con base en los resultados obtenidos, se concluye que los modelos de Ciencia de Datos pueden contribuir a la predicción del riesgo de retraso en actividades mediante el análisis de patrones estructurales presentes en los cronogramas.

Aunque las variables utilizadas no explican completamente el fenómeno, sí contienen información suficiente para identificar tendencias y generar predicciones útiles para procesos de planificación y control.

Finalmente, la implementación de un dashboard en Power BI permitió ilustrar la aplicabilidad práctica del modelo desarrollado, facilitando la visualización de actividades con diferentes niveles de riesgo y contribuyendo a una gestión más proactiva basada en datos.

6.2 Recomendaciones

A partir de los resultados obtenidos, se plantean las siguientes recomendaciones para futuras investigaciones y aplicaciones prácticas.

Se recomienda incorporar variables adicionales relacionadas con recursos, productividad, especialidades involucradas, restricciones operativas y desempeño de ejecución. La inclusión de este tipo de información podría incrementar la capacidad predictiva de los modelos y mejorar los resultados obtenidos.

Asimismo, se recomienda ampliar la cantidad y diversidad de proyectos analizados, incorporando cronogramas provenientes de diferentes organizaciones o sectores industriales. Esto permitiría evaluar la capacidad de generalización de los modelos y validar los resultados en contextos distintos a los considerados en la presente investigación.

Se recomienda también explorar algoritmos adicionales de Machine Learning y técnicas más avanzadas de optimización de hiperparámetros, con el propósito de continuar mejorando el desempeño predictivo y la capacidad de discriminación de los modelos.

Desde una perspectiva práctica, se recomienda integrar el modelo desarrollado dentro de los procesos habituales de planificación y seguimiento de proyectos, permitiendo identificar actividades potencialmente críticas antes de que se materialicen los retrasos.

Además, se sugiere continuar utilizando herramientas de visualización como Power BI para facilitar la interpretación de resultados y apoyar la toma de decisiones por parte de planificadores, gerentes de proyecto y equipos de control.

Se recomienda incorporar estrategias de validación basadas en agrupación por proyecto y procedimientos formales de selección de hiperparámetros y umbrales de clasificación. Esto permitiría evaluar con mayor rigurosidad la capacidad de generalización de los modelos en proyectos no observados previamente y fortalecer la robustez metodológica de los resultados.

Finalmente, se recomienda profundizar en investigaciones dirigidas al análisis predictivo de cronogramas, considerando no solamente el retraso de actividades individuales, sino también el impacto de dichas actividades sobre hitos, rutas críticas y fechas finales de terminación de proyectos.

6.3 Síntesis final de la investigación

Los resultados obtenidos aportan evidencia de que la aplicación de técnicas de Ciencia de Datos sobre datos históricos de cronogramas constituye una alternativa viable para la identificación temprana del riesgo de retraso en actividades. A partir del análisis de 8.092 actividades históricas y la evaluación de distintos algoritmos de Machine Learning, fue posible desarrollar un modelo predictivo capaz de identificar patrones asociados a la ocurrencia de retrasos utilizando únicamente variables estructurales del cronograma.

Los resultados obtenidos evidencian el potencial de la analítica de datos para complementar los enfoques tradicionales de gestión de proyectos y contribuir al desarrollo de mecanismos de planificación más proactivos, basados en evidencia y orientados a la anticipación de riesgos.

7 REFERENCIAS

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth International Group.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. Springer.

Project Management Institute. (2021). A guide to the project management body of knowledge (PMBOK® Guide) (7th ed.). Project Management Institute.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: **Machine** learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Lishner, I., & Shtub, A. (2022). Using an Artificial Neural Network for Improving the Prediction of Project Duration. Mathematics, 10(22), 4189.

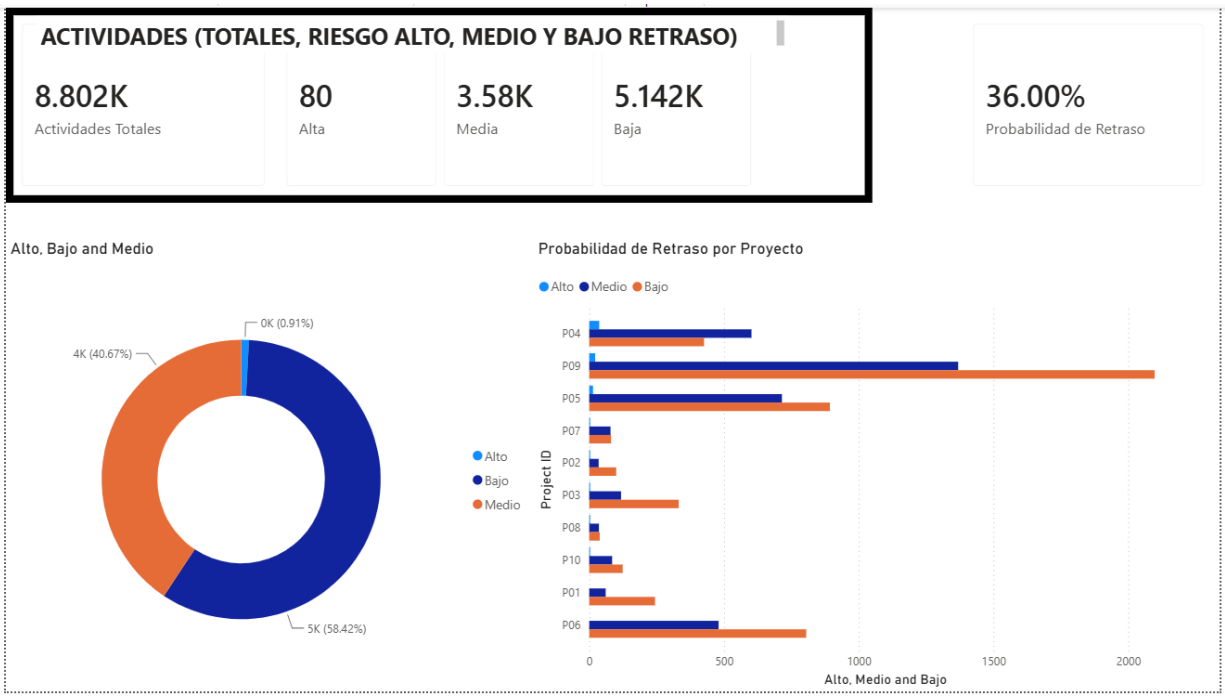
Pang, D. J. (2023). Hybrid Machine Learning Model Performance in IT Project Cost and Duration Prediction. Advances in Science, Technology and Engineering Systems Journal, 8(2), 108-115.

Kothapalli, S. (2024). Enhanced Construction Schedule Prediction Using Ensemble Machine Learning: A Comparative Case Study. Quest Journal of Architecture and Civil Engineering, 9(2), 28-36.

Restrepo Ramírez, A. F., & Rua-Machado, C. A. (2025). Predicting Delays and Cost Overruns in Construction Projects: A Machine Learning Approach. Revista EIA, 22(44)

8 ANEXOS

8.1 Anexo A. Dashboard Power BI



8.2 Anexo B. Código principal utilizado para el desarrollo del modelo predictivo

El código completo utilizado durante la investigación se incluye en el Anexo B entregado como documento complementario.

8.3 Anexo C. Data set anonimizado.