Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

A TOOL FOR HUMAN EVALUATION OF INTERPRETABILITY

Autor(a): Adrian Vargas Rangel
Tutor(a): Bojan Mihaljevic

Madrid, enero 2025

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial

Título: A TOOL FOR HUMAN EVALUATION OF INTERPRETABILITY

enero 2025

Autor(a): Adrian Vargas Rangel
Tutor(a): Bojan Mihaljevic
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

A medida que los sistemas de inteligencia artificial (IA) se integran en diversos sectores de la sociedad, la necesidad de modelos de aprendizaje automático interpretables se vuelve crucial para su aceptación y uso ético. En áreas críticas como la medicina, donde los resultados de los algoritmos utilizados para diagnósticos deben ser precisos y comprensibles, es esencial desarrollar nuevas técnicas de explicabilidad que faciliten la toma de decisiones y garanticen la fiabilidad de estas tecnologías.

Este Trabajo de Fin de Máster se inspira en la metodología del estudio *"Interpretable Decision Sets: A Joint Framework for Description and Prediction"* [1] y desarrolla un cuestionario para evaluar la interpretabilidad de modelos transparentes, como los árboles de decisión y el modelo de Interpretable Decision Sets (IDS) propuesto en dicho estudio. Para la evaluación, se emplean datos sobre el rendimiento académico en matemáticas de estudiantes, proporcionados por Paulo Cortez y disponibles en el *UCI Machine Learning Repository* [2].

El cuestionario está diseñado para extraer las reglas subyacentes de cada modelo, clasificando a los estudiantes como aprobados o no aprobados, y evaluando la capacidad de los usuarios para interpretar correctamente estas reglas, incluso en situaciones ambiguas. Esta evaluación incluye tanto la exactitud de las predicciones como la habilidad de los usuarios para identificar y comprender errores, lo cual es fundamental para fomentar la confianza y el uso efectivo de la IA en decisiones reales.

Con este enfoque, se busca contribuir al campo de la inteligencia artificial explicable proporcionando una base sólida para la construcción de herramientas de investigación que exploren cómo los humanos interpretan las decisiones de modelos transparentes.

El código fuente del análisis y la librería *IDS* personalizada utilizada en este trabajo están disponibles en el repositorio de GitHub de la herramienta web *Survey-XAI-App*: *https://github.com/adrian-vargas/survey-xai*, que incluye:

- Notebook del análisis final: *notebook_final.ipynb*.
- Archivo Excel con resultados de validación: *resultados_predicciones.xlsx*.
- Librería IDS desarrollada para este TFM: *IDS*.

El repositorio independiente de la librería *IDS* se encuentra en:
https://github.com/adrian-vargas/IDS.

Abstract

As artificial intelligence (AI) systems become increasingly integrated into various sectors of society, the need for interpretable machine learning models is crucial for their acceptance and ethical use. In critical fields such as medicine, where the results of algorithms used for diagnostics must be both accurate and understandable, it is essential to develop new explainability techniques that facilitate decision-making and ensure the reliability of these technologies.

This Master's Thesis is inspired by the methodology from the study "Interpretable Decision Sets: A Joint Framework for Description and Prediction" [1] and develops a questionnaire to assess the interpretability of transparent models, such as decision trees and the Interpretable Decision Sets (IDS) model proposed in that study. For the evaluation, data on students' academic performance in mathematics, provided by Paulo Cortez and available in the UCI Machine Learning Repository [2], were used.

The questionnaire is designed to extract the underlying rules of each model, classifying students as passing or failing, and evaluating the ability of users to correctly interpret these rules, even in ambiguous situations. This evaluation includes both the accuracy of the predictions and the ability of users to identify and understand errors, which is fundamental to fostering trust and effectively using AI in real-world decisions.

This approach aims to contribute to the field of explainable artificial intelligence by providing a solid foundation for building research tools that explore how humans interpret the decisions of transparent models.

The source code for the analysis and the custom *IDS* library used in this work are available in the GitHub repository of the web tool *Survey-XAI-App*:
https://github.com/adrian-vargas/survey-xai, which includes:

- Final analysis notebook: *notebook_final.ipynb*.
- Excel file with validation results: *resultados_predicciones.xlsx*.
- IDS library developed for this Master's Thesis: *IDS*.

The independent repository for the *IDS* library can be found at:
https://github.com/adrian-vargas/IDS.

Tabla de contenidos

Resumen **i**

Abstract **iii**

1. Introducción **1**
1.1. Motivación 1
1.2. Objetivo 2
1.3. Objetivos Específicos 2
1.4. Hipótesis 2
1.5. Estructura del Documento 2

2. Fundamentos Teóricos **5**
2.1. Introducción a la Interpretabilidad 5
2.2. Criterios de Interpretabilidad 6
2.2.1. Sparsidad 7
2.2.2. Simulabilidad 8
2.2.3. Modularidad 8
2.2.4. Parsimonía 9
2.2.5. Subgrupos: Métodos Contrastivos y Emergentes 10
2.2.6. Contexto y Audiencia 11
2.3. Modelos de Decisión Interpretables 11
2.3.1. Árboles de Decisión (DT) 11
2.3.2. Listas de Decisión 13
2.3.3. Conjuntos Interpretables de Decisión (IDS) 15
2.4. Factores que Afectan la Interpretabilidad 19
2.4.1. Transparencia del Modelo 19
2.4.2. Confianza en Visualizaciones 20
2.5. Evaluación de la Interpretabilidad: Métodos y Métricas 22
2.5.1. Necesidad de Validación Empírica 23
2.5.2. Métodos de Evaluación 23
2.6. Resumen 25

3. Estado del Arte **27**
3.1. Inteligencia Artifcial Explicable 27
3.2. XAI para Modelos de Caja Negra 28
3.3. XAI para Modelos Transparentes 29
3.4. Métricas de Evaluación de la Explicabilidad en Modelos Transparentes . 30
3.4.1. Número de Características 30

3.4.2. Complejidad de la Estructura del Modelo 30
3.5. Métricas para la Evaluación Humana de la Interpretabilidad 31
3.5.1. Métricas Cualitativas 31
3.5.2. Métricas Cuantitativas 31
3.6. Herramientas de Interpretabilidad para Modelos Transparentes 32
3.6.1. InterpretML 32
3.6.2. Yellowbrick 34
3.6.3. Anchors 35
3.6.4. DiCE: Explicaciones Contrafactuales Diversas 36
3.7. Resumen 37

4. Metodología **39**
4.1. Diseño General del Estudio 39
4.2. Preparación del Dataset 40
4.2.1. Preprocesamiento de Datos 42
4.3. Desarrollo de los Modelos 42
4.3.1. Modelo DT con Scikit-learn 43
4.3.2. Modelo DT con InterpretML 45
4.3.3. Modelo IDS 49
4.3.3.1. Implementación del Modelo 49
4.4. Evaluación Técnica de los Modelos 52
4.5. Diseño y Desarrollo del Cuestionario 53
4.5.1. Estructura final 53
4.5.2. Selección de Observaciones 54
4.5.3. Reporte Individual 55
4.5.4. Reporte Consolidado 56
4.5.5. Glosario 57
4.6. Desarrollo de la Herramienta de Interpretabilidad 60
4.6.1. Gestión de Accesos 62
4.6.2. Cuestionario 62
4.7. Pruebas de la Herramienta de Interpretabilidad 63

5. Análisis de Resultados **65**
5.1. Evaluación del Rendimiento Predictivo 65
5.2. Propiedades Estructurales 67
5.3. Cálculo de Métricas de Interpretabilidad 67
5.4. Relación Precisión-Parsimonia 68
5.5. Distribución de Probabilidades de Predicción 69
5.6. Cálculo de Interpretabilidad 70
5.7. Generación de Reportes 70
5.8. Implementación de Visualizaciones 74
5.8.1. Grafo Global del Modelo IDS 74
5.8.2. Grafo Local del Modelo IDS 74
5.8.3. Visualizaciones del Modelo DT-InterpretML 75

6. Conclusiones **77**
6.1. Logro de los Objetivos 77
6.2. Validación de Hipótesis 78
6.3. Limitaciones del Estudio 78
6.4. Trabajo Futuro 79

6.5. Reflexión Final 79

A. Métricas Complementarias de Interpretabilidad 85

B. Propiedad Complementaria: Cobertura 87

C. Relación Precisión-Cobertura 89

D. Configuraciones para la Métrica de Interpretabilidad 91

E. Plantilla del Cuestionario de Interpretabilidad 93
E.1. Pregunta 1 93
E.2. Pregunta 2 94
E.3. Pregunta 3 94
E.4. Pregunta 4 95
E.5. Pregunta 5 96
E.6. Pregunta 6 97
E.7. Pregunta 7 98
E.8. Pregunta 8 99
E.9. Pregunta 9 100
E.10Pregunta 10 101
E.11Pregunta 11 102
E.12Pregunta 12 103
E.13Pregunta 13 104
E.14Pregunta 14 105
E.15Pregunta 15 106
E.16Pregunta 16 107
E.17Pregunta 17 108
E.18Pregunta 18 109
E.19Pregunta 19 110
E.20Pregunta 20 110
E.21Pregunta 21 111

F. Resultados de los Usuarios de Prueba 113
F.1. Ambigüedad 113
F.2. Error 113
F.3. Exactitud 114
F.4. Preguntas de Seguimiento 114
F.4.1. Ambigüedad 114
F.4.2. Error 114

Capítulo 1

Introducción

En este capítulo se ofrece una visión general del contexto de la investigación. En la Sección 1.1, se aborda la motivación que ha llevado a la realización de este proyecto, explicando las razones que justifican su relevancia. La Sección 1.2 describe el objetivo principal y los objetivos específicos que se persiguen, mientras que la Sección 1.3 plantea la hipótesis que se pretende validar con este estudio. Finalmente, en la Sección 1.4 se presenta la estructura del documento como guía sobre el contenido de cada uno de los capítulos.

1.1. Motivación

Mi interés por la inteligencia artificial explicable se consolidó durante el Máster en Inteligencia Artificial, específicamente en la asignatura impartida por los profesores Bojan Mihaljevic y Esteban García Cuesta. Este campo cobra relevancia debido a la creciente preocupación en la comunidad científica y en la sociedad por la falta de transparencia en muchos sistemas de IA. En sectores críticos como la medicina, la educación y la justicia, las decisiones automatizadas pueden tener consecuencias significativas e incluso irreversibles, lo que resalta la necesidad urgente de desarrollar modelos más comprensibles, transparentes y justos [3].

La importancia de esta necesidad se evidencia en casos recientes, como el escándalo de los subsidios para el cuidado infantil en los Países Bajos en 2018. Un sistema automatizado de detección de fraudes, implementado por la Agencia de Impuestos neerlandesa, etiquetó incorrectamente a más de 26,000 familias, en su mayoría de origen migrante, como fraudulentas [4, 5]. Estas acusaciones erróneas dieron lugar a demandas de devolución de sumas de dinero que a veces alcanzaron hasta 100.000 euros [6], provocando graves consecuencias sociales y económicas, como la pérdida de empleo y vivienda, así como un aumento de problemas de salud mental entre los afectados.

Otro ejemplo que subraya la importancia de la interpretabilidad es el caso Loomis [7]. En este caso, el uso del software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) en el sistema judicial de Estados Unidos, diseñado para predecir la probabilidad de reincidencia de un acusado, fue criticado por su falta de transparencia y explicabilidad. Un estudio realizado por ProPublica en 2016 reveló que COMPAS no solo era menos preciso de lo que se afirmaba, sino que también

presentaba sesgos raciales significativos, clasificando a personas negras como de mayor riesgo de reincidencia en comparación con personas blancas con historiales similares [8].

Estos incidentes destacan la necesidad de modelos de inteligencia artificial que sean no solo técnicamente eficientes, sino también transparentes e interpretables para evitar decisiones injustas.

Basándose en el estudio “Interpretable Decision Sets (IDS)"de Lakkaraju et al. [1], esta investigación se centra en evaluar la interpretabilidad de modelos explicables desde la perspectiva de los usuarios finales.

1.2. Objetivo

El objetivo de este Trabajo de Fin de Máster (TFM) es desarrollar un cuestionario para evaluar la interpretabilidad de modelos de decisión, específicamente de los modelos *Interpretable Decision Sets* (IDS) y los árboles de decisión (DT), en el contexto de la inteligencia artificial. Este cuestionario, inspirado en el estudio “Interpretable Decision Sets: A Joint Framework for Description and Prediction” [1], será utilizado para investigar cómo diversos factores, como la estructura del modelo y la presentación visual de los resultados, influyen en la percepción de interpretabilidad y en la capacidad de los usuarios para detectar errores.

1.3. Objetivos Específicos

1. Diseñar un cuestionario para evaluar la percepción de interpretabilidad de los modelos *Interpretable Decision Sets* (IDS) y los árboles de decisión (DT).
2. Implementar el cuestionario utilizando un conjunto de datos sobre el rendimiento académico de estudiantes en matemáticas [2].
3. Analizar cómo factores como la estructura del modelo, la ambigüedad de la información y la confianza en las visualizaciones afectan la percepción de interpretabilidad.
4. Desarrollar una herramienta que sirva como base para futuros estudios en el campo de la inteligencia artificial explicable.

1.4. Hipótesis

Se plantea que los modelos IDS serán percibidos como más interpretables que los árboles de decisión, especialmente en situaciones de ambigüedad. No obstante, la confianza excesiva en las visualizaciones puede llevar a una comprensión superficial, afectando la capacidad de los usuarios para identificar errores correctamente. Por tanto, se sugiere que tanto la estructura del modelo como la presentación de resultados influyen en la percepción de interpretabilidad y precisión.

1.5. Estructura del Documento

El presente trabajo se organiza en los siguientes capítulos:

- *Capítulo 1: Introducción.* Este capítulo proporciona una introducción general al tema del TFM, incluyendo el contexto, la motivación del estudio, el objetivo principal, los objetivos específicos, la hipótesis y la estructura general del documento.
- *Capítulo 2: Fundamentos Teóricos.* Se revisan los conceptos teóricos clave relacionados con la interpretabilidad en inteligencia artificial, así como una descripción de los modelos que se evaluarán, como los *Interpretable Decision Sets* (IDS) y los *Árboles de Decisión* (DT). Se incluyen definiciones de términos, técnicas y metodologías relevantes, y se discute la importancia de la interpretabilidad en la IA.
- *Capítulo 3: Estado del Arte.* Este capítulo presenta una revisión de los estudios relevantes en el campo, destacando investigaciones previas relacionadas con la evaluación de la interpretabilidad de modelos de IA, los desafíos asociados y los enfoques utilizados en trabajos similares. Se posiciona el presente trabajo en el contexto de la literatura existente.
- *Capítulo 4: Metodología.* Se describe el diseño del cuestionario, la selección del conjunto de datos (rendimiento académico de los estudiantes en matemáticas) y el proceso de desarrollo de las herramientas utilizadas. También se incluyen los métodos de análisis de datos empleados para evaluar la interpretabilidad de los modelos.
- *Capítulo 5: Análisis de Resultados.* En este capítulo se presenta un análisis exhaustivo sobre su interpretabilidad y rendimiento predictivo de los modelos evaluados. Se consideran aspectos como el rendimiento técnico, las propiedades estructurales, las métricas de interpretabilidad y la relación precisión-parsimonia. También se describen las visualizaciones y reportes generados por la herramienta desarrollada.
- *Capítulo 6: Conclusiones.* Este capítulo resume las conclusiones obtenidas del estudio. Se reflexiona sobre el cumplimiento de los objetivos y la validación de las hipótesis planteadas. Además, se presentan las limitaciones del trabajo y se proponen líneas de investigación futura para mejorar y ampliar los resultados obtenidos.

Capítulo 2

Fundamentos Teóricos

Este capítulo presenta los conceptos clave sobre la interpretabilidad en inteligencia artificial y su importancia en algunos sectores críticos. El capítulo se divide en:

- *Sección 2.1:* Introducción a la interpretabilidad y sus enfoques principales (*global* y *local*).
- *Sección 2.2:* Criterios para mejorar la interpretabilidad, como la *sparsidad*, *simulabilidad*, *modularidad* y *parsimonia*.
- *Sección 2.3:* Modelos de decisión interpretables: árboles de decisión (DT), listas de decisión y conjuntos interpretables de decisión (IDS).
- *Sección 2.4:* Factores que afectan la interpretabilidad: transparencia del modelo y visualizaciones.
- *Sección 2.5:* Métodos de evaluación de interpretabilidad mediante métricas cuantitativas y estudios empíricos.

2.1. Introducción a la Interpretabilidad

La interpretabilidad en el campo de la inteligencia artificial (IA) se define como la capacidad de explicar o presentar los resultados de un sistema de aprendizaje automático de manera comprensible para los seres humanos. Es fundamental para garantizar que se cumplan criterios como la equidad, la privacidad, la fiabilidad, la robustez, la causalidad, la usabilidad y la confianza. De este modo, permite a los usuarios entender cómo funcionan los algoritmos y verificar si estos cumplen con los objetivos esperados [9, 10].

A diferencia de la explicabilidad, que se enfoca en proporcionar explicaciones posteriores al desarrollo de los modelos (post-hoc), la interpretabilidad basada en modelos implica un diseño intencional desde la etapa de modelado para que sean comprensibles para los usuarios. Esto puede lograrse mediante la construcción de modelos intrínsecamente interpretables, como los árboles de decisión, que ofrecen claridad sobre las relaciones aprendidas a partir de los datos. Sin embargo, este enfoque presenta el reto de equilibrar la simplicidad del modelo con su precisión predictiva, además de lidiar con posibles sesgos en los datos o en los métodos explicativos utilizados [11, 12].

Existen dos enfoques principales para abordar la interpretabilidad:

- *Interpretabilidad local:* Se centra en explicar la predicción del modelo para una instancia específica. Responde preguntas como: "¿Por qué se tomó esta decisión para este dato?". Técnicas como LIME (Local Interpretable Model-agnostic Explanations) muestran qué características del dato son más relevantes para una predicción individual [9, 13]. Es útil en situaciones donde se necesita entender decisiones individuales, como en diagnósticos médicos o aprobaciones de crédito.
- *Interpretabilidad global:* Busca comprender el comportamiento general del modelo en todo el conjunto de datos. Responde a preguntas como: "¿Cómo toma decisiones el modelo en general?" o "¿Qué patrones ha aprendido el modelo?". Métodos como los árboles de decisión y los modelos aditivos generalizados (GAMs) ofrecen una visión global, permitiendo observar cómo las predicciones varían según las características de los datos [9, 12]. Es crucial en aplicaciones que requieren entender las reglas generales del modelo, como en políticas públicas o investigaciones científicas.

Ambos enfoques son complementarios y se aplican en diferentes contextos. Mientras que la *interpretabilidad local* facilita la comprensión de decisiones individuales, la *interpretabilidad global* proporciona una visión más amplia del funcionamiento del modelo, lo cual es esencial para su validación y aceptación en aplicaciones críticas [13, 9, 12].

En este TFM, la interpretabilidad se analizará en el contexto de modelos de decisión interpretables, específicamente en los *Interpretable Decision Sets* (IDS) y los árboles de decisión (DT). Se explorará cómo la estructura del modelo, la presentación de resultados y otros factores influyen en la percepción de interpretabilidad y en la capacidad de los usuarios para detectar errores.

2.2. Criterios de Interpretabilidad

Esta sección se basa en los trabajos de Murdoch et al. (2019) [11] y Rudin (2019) [14], quienes destacan, desde diferentes perspectivas, la importancia de utilizar modelos de decisión que sean inherentemente interpretables en aplicaciones donde la transparencia es crucial, como la medicina, la justicia o la toma de decisiones críticas en entornos regulatorios.

Los modelos de caja negra, utilizados en decisiones de alto riesgo, pueden tener consecuencias negativas significativas debido a la falta de transparencia y a la dificultad para interpretar sus resultados. Por ello, ambos autores recomiendan el uso de modelos interpretables que proporcionen explicaciones fieles y comprensibles, en lugar de depender de explicaciones post hoc.

Según estos estudios, es más efectivo diseñar modelos que sean interpretables desde el principio, ya que esto evita las complicaciones asociadas con la explicación de modelos complejos. Para alcanzar esta interpretabilidad, se pueden utilizar varios enfoques intrínsecos, tales como la sparsidad, la simulabilidad, y la modularidad. A continuación, se exploran estos tres enfoques y sus beneficios en el diseño de modelos interpretables.

2.2.1. Sparsidad

La sparsidad implica reducir el número de parámetros no nulos en un modelo, lo que facilita la comprensión del papel específico de cada variable en las predicciones. Este enfoque es particularmente útil en casos donde se sabe que la relación subyacente depende de un conjunto limitado de señales significativas. Por ejemplo, los modelos lineales, como la regresión LASSO (Least Absolute Shrinkage and Selection Operator), o métodos como el sparse coding, aplican técnicas de penalización para mantener la simplicidad del modelo sin comprometer significativamente la precisión predictiva [15]. Esta reducción en complejidad no solo mejora la interpretabilidad, sino que también puede ayudar a evitar el sobreajuste al destacar características verdaderamente relevantes.

La regresión LASSO se define mediante la siguiente ecuación [15]:

$$\min_{\theta} \left(\frac{1}{2} \|Z - \Phi\theta\|_2^2 + \lambda \|\theta\|_1 \right), \tag{2.1}$$

donde Z es el vector de las salidas observadas, Φ es la matriz de regresores (características), θ es el vector de parámetros desconocidos del modelo, $\|\cdot\|_2$ denota la norma L_2, y $\|\cdot\|_1$ denota la norma L_1.

La sparsidad se promueve a través del término de penalización $\lambda\|\theta\|_1$, que reduce algunos de los coeficientes θ a cero. Este enfoque favorece modelos más simples y parsimoniosos, mejorando la interpretabilidad al centrarse únicamente en las características más relevantes.

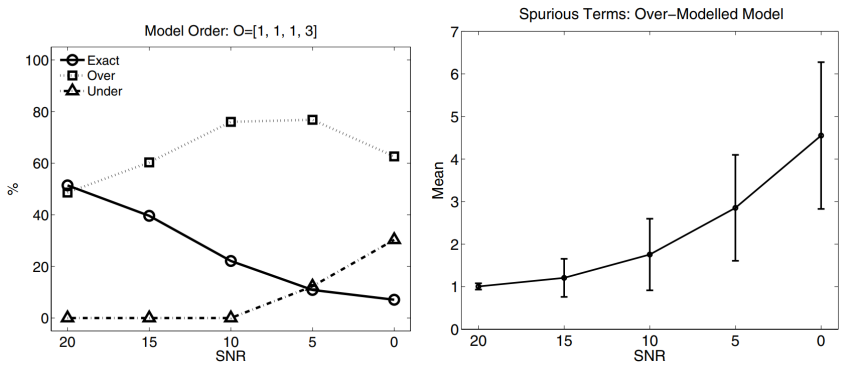


Figura 2.1: Relación entre sparsidad y selección de modelos utilizando LASSO. El gráfico de la izquierda muestra la tasa de selección de modelos exactos (círculos), sobre-modelados (cuadrados) y sub-modelados (triángulos) en función del nivel de ruido en la señal (SNR). Un modelo exacto tiene solo las variables relevantes, mientras que un sobre-modelado incluye variables adicionales innecesarias (términos espurios) y un sub-modelado omite variables relevantes. El gráfico de la derecha muestra la media y desviación estándar del número de términos espurios seleccionados en modelos sobre-modelados, destacando cómo LASSO minimiza la inclusión de estas variables irrelevantes, promoviendo así la sparsidad y la interpretabilidad del modelo. Adaptado de [15].

2.2.2. Simulabilidad

Por otro lado, la simulabilidad se refiere a la capacidad de un modelo para ser reproducido y comprendido fácilmente por un ser humano. Modelos como los árboles de decisión y las listas de reglas "si-entonces" ejemplifican esta característica, ya que permiten seguir paso a paso el proceso de toma de decisiones del modelo. Este enfoque es especialmente valioso en contextos donde las decisiones deben ser comprensibles para personas no expertas, como en aplicaciones médicas, donde tanto los pacientes como los profesionales de la salud necesitan entender las recomendaciones del modelo.

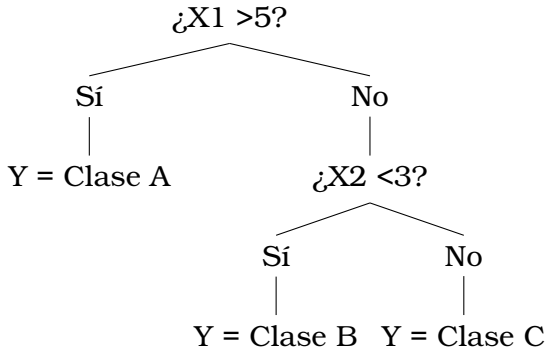


Figura 2.2: Árbol de decisión basado en [11] que ejemplifica la simulabilidad al dividir el espacio de características en regiones de decisión claras mediante reglas "si-entonces". Cada nodo representa una pregunta sobre las características (X_1 y X_2), y las ramas llevan a decisiones simples ("Sí" o "No"), facilitando que cualquier usuario pueda seguir y reproducir el proceso de toma de decisiones del modelo.

2.2.3. Modularidad

La modularidad es un enfoque clave en modelos interpretables, ya que permite descomponer el modelo en partes significativas que se pueden interpretar de forma independiente. Este enfoque es particularmente útil en modelos complejos, donde los subcomponentes o módulos tienen un significado interpretativo claro.

Por ejemplo, los Modelos Aditivos Generalizados (GAMs)[16], como el Explainable Boosting Machine (EBM) [11], restringen las relaciones entre las variables a una forma aditiva, lo que facilita la interpretación de cada término individual. Esto significa que cada función en el modelo representa el efecto de un predictor específico en la predicción final, y estos efectos se pueden analizar de manera aislada para comprender su contribución individual.

La ecuación de un GAM se puede expresar como:

$$g(\mu) = \beta + f_1(x_1) + \cdots + f_m(x_m) \tag{2.2}$$

donde $g(\mu)$ es la función de enlace, β es el término independiente, y $f_i(x_i)$ son funciones suaves que representan el efecto de cada predictor x_i.

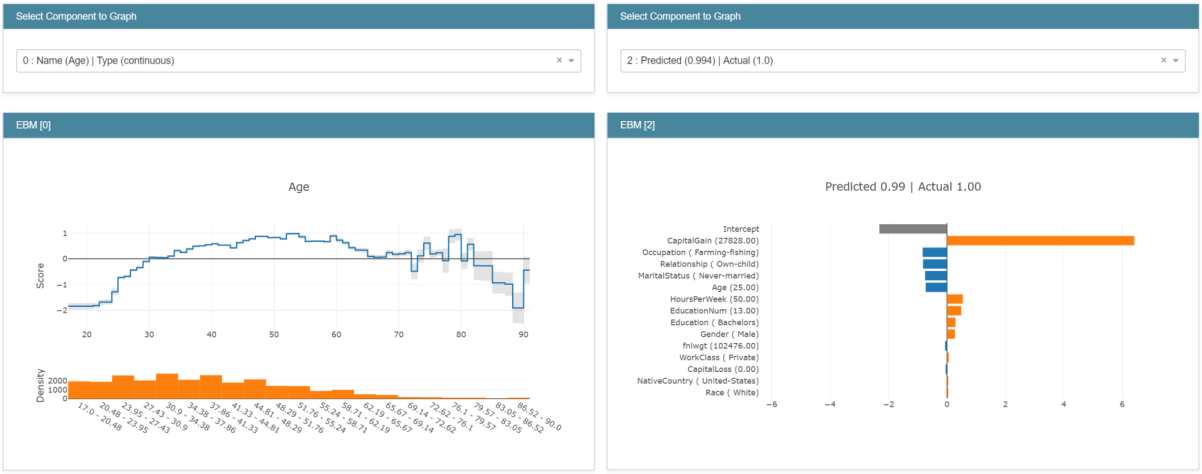


Figura 2.3: Ejemplo de modularidad en un modelo aditivo como el EBM. Izquierda: La función f_{Age} muestra cómo la característica 'Edad' afecta la predicción final. Derecha: Descomposición de una predicción individual, donde se destaca cómo cada característica contribuye de manera independiente, con 'CapitalGain' dominando la predicción. Adaptado de [11].

2.2.4. Parsimonía

La parsimonia se refiere a la simplicidad de un modelo al utilizar el mínimo número de parámetros o términos necesarios para capturar la dinámica de un sistema. Esto facilita la interpretabilidad y la generalización del modelo, ya que reduce la complejidad y ayuda a evitar sobreajustes. Según Kutz y Brunton (2022), promover la parsimonia en el aprendizaje automático, especialmente en modelos informados por la física, resulta en modelos más interpretables y físicamente coherentes, permitiendo una mejor generalización a nuevos escenarios.

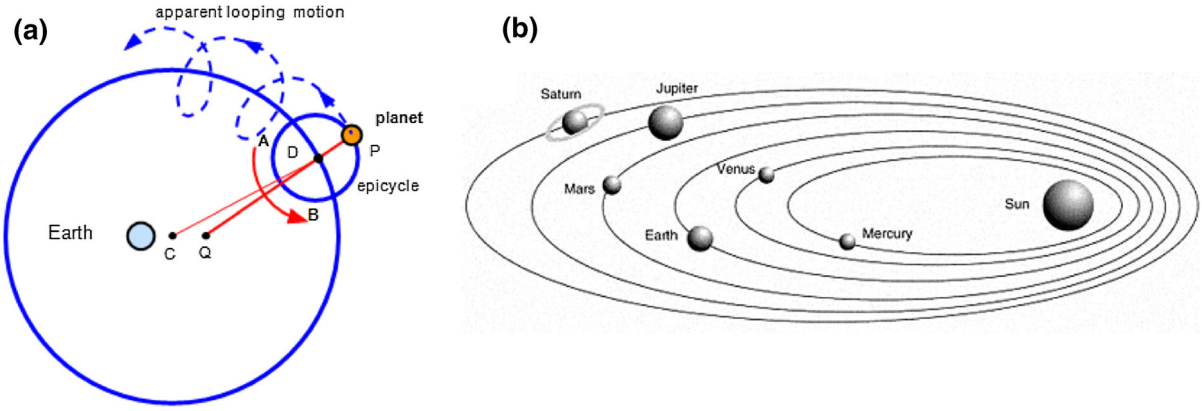


Figura 2.4: Comparación entre dos modelos astronómicos que ilustran el principio de parsimonía: (a) El modelo ptolemaico describe el movimiento planetario con epiciclos y deferentes, donde el ángulo azimutal del planeta se calcula mediante $\alpha(t) = \Omega t - \sin^{-1}\left(\frac{CE}{R}\sin(\Omega t)\right)$ [17], lo que implica una mayor complejidad debido a la necesidad de múltiples parámetros; (b) El modelo heliocéntrico de Copérnico, refinado por Newton, utiliza una ley de gravitación universal más simple, $F = G\frac{m_1 m_2}{r^2}$, que requiere menos supuestos y es más parsimonioso. Imagen adaptada de [18].

2.2.5. Subgrupos: Métodos Contrastivos y Emergentes

Además de los enfoques intrínsecos como la sparsidad, la simulabilidad, la modularidad y la parsimonia, los métodos basados en reglas también emplean técnicas específicas para mejorar la interpretabilidad mediante la identificación de patrones diferenciados en los datos. Un enfoque destacado es el *descubrimiento de subgrupos*, que busca encontrar segmentos de datos que sean relevantes o interesantes con respecto a una variable de interés.

A diferencia de métodos como las Máquinas de Soporte Vectorial o Redes Neuronales, que se enfocan en maximizar la precisión de la clasificación de ejemplos individuales, los métodos basados en reglas, como el descubrimiento de subgrupos, se centran en caracterizar las clases a través de sus relaciones con otras entidades presentes en los datos [19]. Estos métodos no solo se preocupan por la precisión de la predicción, sino también por la claridad y comprensibilidad de las relaciones aprendidas, lo cual es esencial para la interpretabilidad [20].

El *descubrimiento de subgrupos*, un concepto ampliamente utilizado en minería de datos, consiste en identificar segmentos de la población que sean estadísticamente relevantes según ciertos criterios. Estos segmentos se definen mediante reglas del tipo "si-condición(es)-entonces-clase", que permiten describir de forma comprensible cómo se agrupan los datos respecto a una propiedad específica [21, 20]. Este paradigma incluye tres enfoques principales:

1. *Descubrimiento de subgrupos*: Este método busca identificar subgrupos dentro del conjunto de datos que tengan una alta probabilidad de cumplir con una determinada característica o condición de interés. Por ejemplo, en un estudio médico, puede identificar pacientes con síntomas similares que tienen una alta probabilidad de padecer una enfermedad específica.

2. *Conjunto de contraste*: Este enfoque se centra en encontrar pares de atributos y valores que sean únicos para cada grupo o clase. Su objetivo es maximizar la diferenciación entre clases, destacando atributos específicos que caracterizan exclusivamente a cada clase. Por ejemplo, en los árboles de decisión, se busca que cada nodo defina claramente las fronteras entre las diferentes clases (ver Figura 2.2).

3. *Patrón emergente*: Se enfoca en extraer subgrupos en los que las frecuencias relativas de la variable objetivo varían de manera significativa entre los distintos grupos. Utiliza listas de decisión y métodos centrados en la "cobertura" de reglas para identificar los atributos que son más representativos o influyentes para una clase específica. Este enfoque es útil para descubrir reglas o patrones que puedan ser menos evidentes pero importantes para la comprensión del modelo (ver Figura 2.3).

En general, estos algoritmos utilizan heurísticas para aproximar un conjunto óptimo de reglas, lo que significa que introducen ciertos supuestos sobre la estructura de los datos para hacer el problema manejable en tiempo polinomial. Estas técnicas ayudan a construir modelos que no solo sean precisos, sino también comprensibles para los usuarios finales, permitiendo una mejor toma de decisiones basada en las interpretaciones claras proporcionadas por las reglas generadas.

2.2.6. Contexto y Audiencia

Finalmente, la elección del enfoque interpretativo depende en gran medida del contexto del problema y de la audiencia objetivo. En aplicaciones donde la precisión predictiva es menos crítica que la interpretabilidad, como en auditorías de modelos para asegurar la equidad, se prefieren modelos más simples y transparentes. En cambio, en situaciones que exigen alta precisión, se pueden considerar métodos post hoc para interpretar modelos más complejos, asegurando así un equilibrio adecuado entre interpretabilidad y precisión.

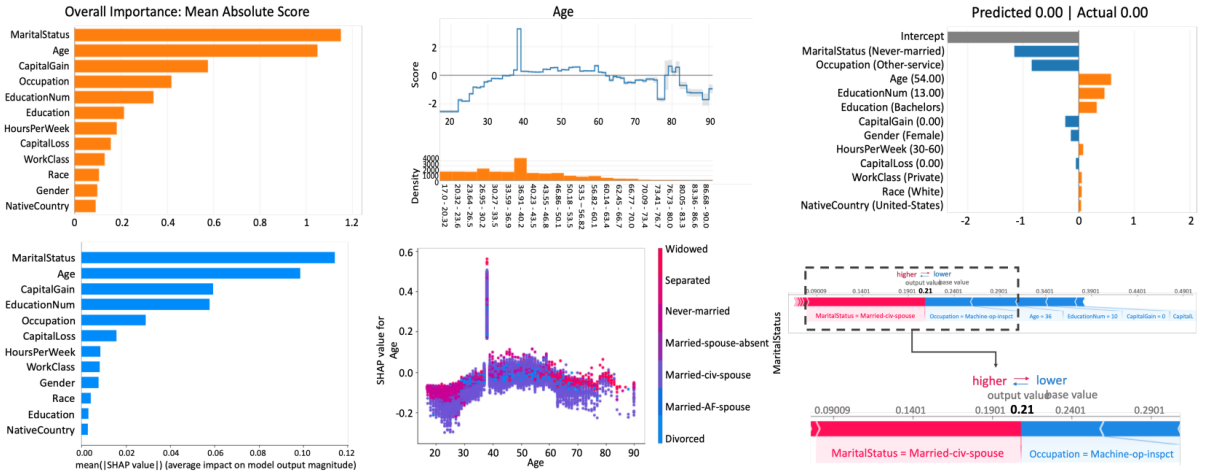


Figura 2.5: Las visualizaciones de GAM (arriba) y SHAP (abajo), adaptadas de [12] y generadas por InterpretML [22], muestran cómo diferentes enfoques interpretativos se ajustan a diversas audiencias y contextos. Las explicaciones globales (columna izquierda) ayudan a científicos de datos a identificar las variables más influyentes en el modelo; los gráficos de componentes o de dependencia (columna central) son útiles para que analistas de negocio interpreten el impacto de factores específicos, como edad o ingreso, en la puntuación crediticia; y las explicaciones locales (columna derecha) son cruciales en medicina personalizada para justificar decisiones como la recomendación de un tratamiento específico.

2.3. Modelos de Decisión Interpretables

A continuación, se describen tres enfoques teóricos relevantes para los modelos de decisión interpretables: Árboles de Decisión (DT), Listas de Decisión (DL) y Conjuntos de Decisiones Interpretables (IDS).

2.3.1. Árboles de Decisión (DT)

Esta sección se basa en los trabajos de Mienye et al. (2024) [23] y Duda et al. (2000) [19], quienes proporcionan un análisis exhaustivo de los conceptos, algoritmos y aplicaciones de los Árboles de Decisión (DT).

Los DT son modelos de aprendizaje supervisado utilizados tanto para tareas de clasificación como de regresión. Un árbol de decisión se construye dividiendo iterativamente un conjunto de datos en subconjuntos más pequeños basados en reglas de

decisión derivadas de las características de los datos. Cada nodo interno del árbol representa una característica o atributo, mientras que cada rama representa una regla de decisión y cada hoja final una decisión o resultado.

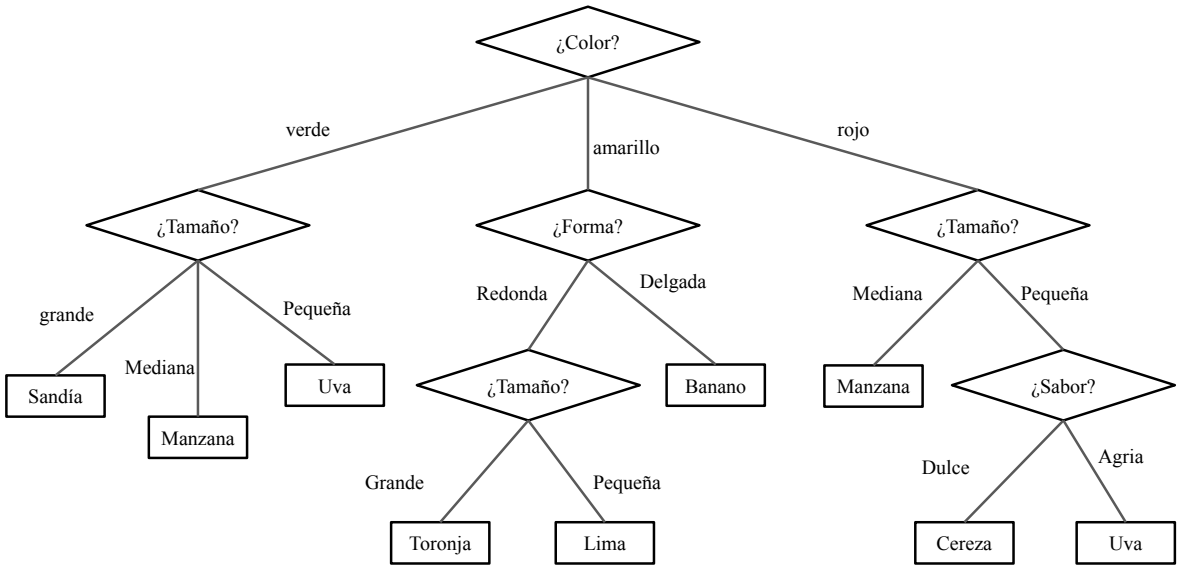


Figura 2.6: Ilustración de un árbol de decisión basada en técnicas de clasificación no paramétricas. Adaptada de [19], Capítulo 4. Los árboles de decisión son fáciles de interpretar cuando son pequeños, pero pueden volverse complejos y difíciles de validar a medida que crecen [24].

si color = verde **y** tamaño = grande **entonces** Sandía
si color = verde **y** tamaño = mediana **entonces** Manzana
si color = verde **y** tamaño = pequeña **entonces** Uva
si color = amarillo **y** forma = redonda **y** tamaño = grande **entonces** Toronja
si color = amarillo **y** forma = redonda **y** tamaño = pequeña **entonces** Lima
si color = amarillo **y** forma = delgada **entonces** Banano
si color = rojo **y** tamaño = mediana **entonces** Manzana
si color = rojo **y** tamaño = pequeña **y** sabor = dulce **entonces** Cereza
si color = rojo **y** tamaño = pequeña **y** sabor = agrio **entonces** Uva

Figura 2.7: Reglas de decisión basadas en el árbol de decisión de la Figura 2.6.

El entrenamiento de un árbol es un proceso recursivo que comienza con todos los datos en la raíz y divide sucesivamente los nodos según la mejor característica, repitiendo el proceso en cada nodo hijo [19]. Uno de los algoritmos más utilizados para construir árboles de decisión es el ID3, que selecciona la característica que proporciona la mayor ganancia de información. A continuación se presenta el algoritmo:

Algorithm 1: Algoritmo ID3 para Árbol de Decisión

Input: Conjunto de datos de entrenamiento $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
Output: Árbol de decisión T

Función `ID3(`D`)`**:**
  if D *está vacío* **then**
    return *un nodo terminal con clase por defecto* $c_{default}$
  if *todas las instancias en* D *tienen la misma etiqueta de clase* y **then**
    return *un nodo terminal con clase* y
  if *el conjunto de atributos* J *está vacío* **then**
    return *un nodo terminal con la clase más frecuente en* D
  Seleccionar el atributo f que mejor divide los datos usando ganancia de información;
  Crear un nodo de decisión para f;
  for *cada valor posible* b_i *de* f **do**
    Crear una rama para b_i;
    Sea D_i el subconjunto de D donde $x_i = b_i$;
    Recursivamente construir el subárbol para D_i;
    Adjuntar el subárbol a la rama para b_i;
  return *el nodo de decisión*

Uno de los criterios más comunes para dividir los nodos es la ganancia de información, definida como:

$$\text{Ganancia de Información} = I(p) - \sum_i \frac{|p_i|}{|p|} I(p_i), \tag{2.3}$$

donde $I(p)$ es la entropía del conjunto de datos original, y $I(p_i)$ es la entropía del subconjunto resultante después de la división.

Otro criterio utilizado es el índice de Gini:

$$\text{Índice de Gini} = 1 - \sum_i p_i^2, \tag{2.4}$$

donde p_i representa la proporción de observaciones de la clase i en el nodo.

2.3.2. Listas de Decisión

Las listas de decisión son una representación para funciones Booleanas construidas como secuencias ordenadas de predicados lógicos del tipo "si [condición(es)] entonces [clase]" seguida de otras condiciones adicionales (e.g., "más si [condición(es)] entonces [clase]"). Las condiciones corresponden a predicados sobre las características del problema de clasificación (los *features*) [24, 25, 26].

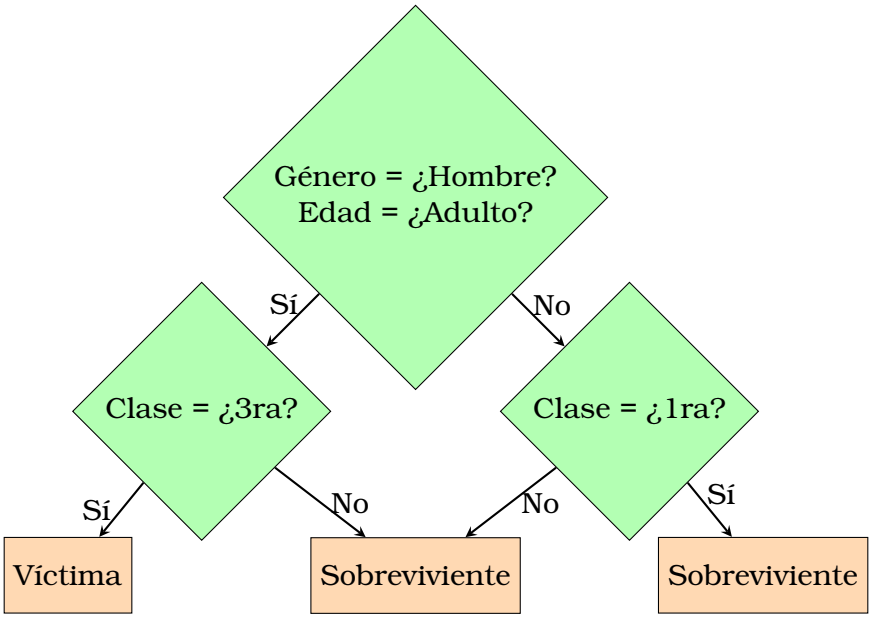


Figura 2.8: Ejemplo de lista de decisión. Adaptado de [25]. Se muestra entrenado con el conjunto de datos del Titanic

si género = hombre **y** edad = adulto **entonces** víctima
si no, si clase = 3ra **entonces** víctima
si no, si clase = 1ra **entonces** sobreviviente
si no, sobreviviente

Figura 2.9: Reglas de decisión basadas en el árbol de decisión de la Figura 2.6.

A diferencia de otros métodos de aprendizaje automático como Máquinas de Soporte Vectorial, Bosques Aleatorios o Redes Neuronales, las listas de decisión ofrecen una explicación natural y sencilla para cada predicción, permitiendo que un experto entienda el mecanismo de decisión y razone sobre el resultado. Esto se considera una ventaja significativa de las listas de decisión [25].

Además, las listas de decisión también ofrecen ventajas durante el proceso de entrenamiento cuando se comparan con otros métodos interpretables, como los árboles de decisión. Mientras que estos métodos utilizan algoritmos avaros, lo cual permite tiempos de entrenamiento cortos pero a costa de una reducción en el desempeño, las listas de decisión construyen modelos secuenciales más robustos, manteniendo una mayor capacidad explicativa [24].

Gracias a su naturaleza "emergente", las listas de decisión se construyen mediante reglas que particionan el espacio de características de manera más significativa. Por ejemplo, una lista de decisión de tamaño k (k-DL) utiliza reglas en forma de cláusulas conjuntas, donde cada regla r_i puede representarse matemáticamente como:

$$r_i : \bigwedge_{j=1}^{m_i} (x_j = a_{ij}) \rightarrow y = c_i,$$

donde x_j son los atributos, a_{ij} son los valores específicos de dichos atributos, y c_i

es la clase asignada. Este enfoque garantiza la tratabilidad computacional del problema, explorando un mayor número de combinaciones posibles sin comprometer la eficiencia [25].

En particular, basándonos en el trabajo de Rivest (1987), las listas de decisión son polinómicamente aprendibles. Esto significa que pueden ser aprendidas de manera eficiente usando algoritmos de aprendizaje, como se define en el marco teórico de Valiant (1984). Rivest demuestra que las listas de decisión de tamaño k (k-DL) pueden ser identificadas en tiempo polinómico mediante un algoritmo codicioso que selecciona en cada paso la regla que maximiza la ganancia de información:

$$\Delta I = I(S) - I(S|A),$$

donde $I(S)$ es la entropía del conjunto de datos S y $I(S|A)$ es la entropía después de dividir S usando el atributo A que mejor explica los datos restantes [26].

El algoritmo encuentra iterativamente reglas que cubren el mayor número de ejemplos de una misma clase, organizándolas en una lista secuencial. Cada regla se aplica en orden hasta que todos los ejemplos han sido clasificados.

Algorithm 2: Algoritmo para Listas de Decisión

Input: Conjunto de datos de entrenamiento $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
Output: Lista de decisión L

Función `DecisionList(`D`)`**:**
    $L \leftarrow \emptyset$;
    while D *no está vacío* **do**
        Encontrar la regla $r = (c, v)$ que cubra más ejemplos en D;
        $L \leftarrow L \cup \{r\}$;
        Eliminar de D los ejemplos cubiertos por r;
    return L

2.3.3. Conjuntos Interpretables de Decisión (IDS)

Esta sección se basa en el trabajo de Lakkaraju et al. (2016) [1], quienes proponen un marco para construir modelos predictivos que son altamente precisos y al mismo tiempo altamente interpretables. Un conjunto de decisión es un conjunto de predicados lógicos asociados a una etiqueta de clase. Las variables de los predicados corresponden a los atributos del problema de clasificación (los *features*). El predicado sigue la forma "si [condición(es)] entonces [clase]". Cada predicado está compuesto por ítems (predicados más pequeños) unidos por conjunción. Se dice que el conjunto asigna una clase a un ejemplo si el predicado es verdadero. En caso de empates (varios predicados son verdaderos, pero asignan una clase diferente), se utiliza alguna regla de desempate, generalmente *ad-hoc* para cada aplicación, como dar prioridad a la regla más específica o a la primera regla que se cumpla. La figura 2.10 muestra un ejemplo de conjunto de decisión.

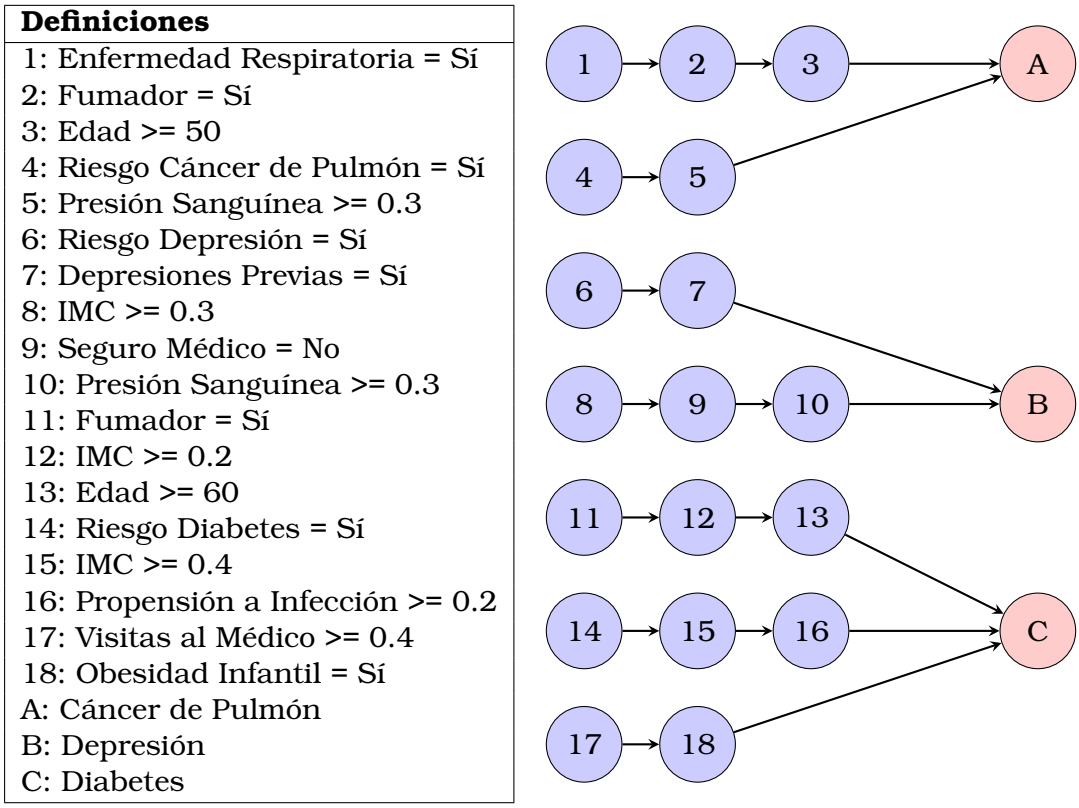


Figura 2.10: Ejemplo de un conjunto interpretable de decisión. Adaptado de [1].

si enfermedad respiratoria = sí **y** fumador = sí **y** edad $\geq$ 50 **entonces** cáncer de pulmón
si riesgo cáncer de pulmón = sí **y** presión sanguínea $\geq$ 0.3 **entonces** cáncer de pulmón
si riesgo depresión = sí **y** depresiones previas = sí **entonces** depresión
si IMC $\geq$ 0.3 **y** seguro médico = no **y** presión sanguínea $\geq$ 0.3 **entonces** depresión
si fumador = sí **y** IMC $\geq$ 0.2 **y** edad $\geq$ 60 **entonces** diabetes
si riesgo diabetes = sí **y** IMC $\geq$ 0.4 **y** propensión a infección $\geq$ 0.2 **entonces** diabetes
si visitas al médico $\geq$ 0.4 **y** obesidad infantil = sí **entonces** diabetes

Figura 2.11: Reglas de decisión basadas en el conjunto interpretable de decisión de la Figura 2.10.

Al igual que las listas de decisión, los conjuntos de decisión ofrecen ventajas sobre otros métodos de aprendizaje automático debido a su sencillez de inferencia e interpretabilidad [27, 24]. Como se mencionó anteriormente, la interpretabilidad se refiere a la facilidad con que un experto humano puede entender y razonar sobre el mecanismo utilizado por el modelo para asignar una clase [14]. Sin embargo, a diferencia de las listas de decisión, que por su naturaleza anidada pueden requerir razonamiento sobre múltiples características y recordar decisiones previas, los conjuntos de decisión solo requieren evaluar predicados simples de forma independiente, lo que los hace aún más interpretables.

Dicha "facilidad de interpretación" es difícil de cuantificar. Aunque la estructura de los conjuntos de decisión se considera interpretable por construcción, es posible construir conjuntos tan grandes y con predicados tan largos que se vuelven difíciles de manejar para un humano, de forma similar a lo que sucede con árboles de

decisión muy grandes. Para abordar este problema, Lakkaraju et al. proponen un marco para construir modelos predictivos que optimizan tanto la precisión como la interpretabilidad. Este marco define una serie de métricas que cuantifican propiedades inherentes a cualquier conjunto de decisión interpretable, tales como:

- *Tamaño del conjunto:* Esta métrica se refiere al número total de reglas en el conjunto de decisión, denotado como $|R|$. Un conjunto más pequeño ($|R|$) tiende a ser más fácil de interpretar, ya que requiere evaluar menos reglas para comprender el modelo completo:

$$|R| \quad \text{(Número total de reglas en el conjunto de decisión).} \tag{2.5}$$

- *Longitud de los predicados:* Mide el número de condiciones dentro de cada regla ïf-then". Si consideramos una regla $r \in R$, la longitud de r, denotada como $\text{len}(r)$, se define como el número de condiciones lógicas en esa regla. Predicados más cortos ($\text{len}(r)$ más bajos) simplifican el proceso de interpretación, ya que reducen la complejidad cognitiva al momento de entender cómo se toma una decisión:

$$\text{len}(r) \quad \text{(Número de condiciones lógicas dentro de una regla).} \tag{2.6}$$

- *Solapamiento entre reglas:* Evalúa la cantidad de solapamiento entre reglas. Sea R_i y R_j dos reglas distintas en el conjunto R. La superposición entre estas reglas puede definirse como:

$$\text{Overlap}(R_i, R_j) = \frac{|R_i \cap R_j|}{\text{mín}(|R_i|, |R_j|)}, \tag{2.7}$$

  donde $|R_i|$ y $|R_j|$ representan el número de instancias cubiertas por las reglas R_i y R_j, respectivamente. Minimizar la superposición es crucial para asegurar que las reglas no entren en conflicto y que cada regla cubra una parte distinta del espacio de datos, facilitando la claridad y precisión de la interpretación.

Estas métricas son integradas en una 'función objetivo' dentro del algoritmo de construcción de IDS. El objetivo del algoritmo no es solo maximizar la precisión predictiva, sino también encontrar un equilibrio que minimice la complejidad del modelo, resultando en conjuntos de decisión más simples y comprensibles. La función objetivo pondera cada una de estas métricas para guiar el proceso de selección de reglas durante el entrenamiento del modelo. Así, se priorizan reglas que ofrecen un buen rendimiento predictivo, pero que al mismo tiempo mantienen la simplicidad y claridad necesarias para ser interpretadas fácilmente por expertos humanos.

Este enfoque garantiza que el modelo resultante no solo sea preciso, sino también interpretable, haciendo que las decisiones derivadas del modelo sean más transparentes y confiables.

El algoritmo de construcción de IDS, denominado *Smooth Local Search* (SLS), optimiza la precisión del modelo y la interpretabilidad de las reglas generadas al buscar el mejor conjunto de reglas bajo ciertas restricciones. El algoritmo funciona de la siguiente manera: minimiza el solapamiento entre las reglas y asegura que cada regla cubra la mayor cantidad de datos posible sin aumentar significativamente la complejidad del modelo. A diferencia de otros métodos de búsqueda, SLS explora de manera eficiente el espacio de soluciones utilizando una búsqueda local suave que evita quedarse atrapado en óptimos locales al suavizar la función objetivo.

El procedimiento del algoritmo SLS se puede describir en los siguientes pasos:

- *Inicialización:* Se inicia con un conjunto vacío de reglas, $R = \emptyset$.
- *Generación de Candidatos:* Se generan candidatos de reglas para cada clase utilizando el conjunto de datos de entrenamiento D. Estas reglas son evaluadas en términos de precisión y cobertura, definidos como:

$$\text{Precisión}(r) = \frac{\text{Número de ejemplos correctamente clasificados por la regla } r}{\text{Número total de ejemplos cubiertos por la regla } r} \tag{2.8}$$

$$\text{Cobertura}(r) = \frac{\text{Número de ejemplos de la clase que cubre la regla } r}{\text{Número total de ejemplos de esa clase en el conjunto de datos } D} \tag{2.9}$$

- *Evaluación y Selección de Reglas:* Se selecciona la regla candidata que maximiza la función objetivo:

$$\text{Función Objetivo}(r) = \alpha \cdot \text{Precisión}(r) + \beta \cdot \text{Cobertura}(r) - \gamma \cdot \text{Solapamiento}(r, R), \tag{2.10}$$

  donde α, β, γ son coeficientes de ponderación que ajustan la importancia relativa de cada término.

- *Verificación de Interpretabilidad:* Antes de añadir una regla al conjunto R, se verifica si cumple con el umbral de interpretabilidad definido por el parámetro θ:

$$\text{Complejidad}(r) \leq \theta. \tag{2.11}$$

- *Optimización de la Complejidad del Modelo:* Después de añadir todas las reglas que cumplen con los criterios anteriores, el conjunto de reglas R es ordenado para optimizar su aplicabilidad y minimizar la complejidad del modelo.

El uso del algoritmo SLS permite construir un conjunto de reglas interpretables de alta calidad que no solo logra una precisión predictiva comparable a otros modelos más complejos, sino que también facilita la comprensión por parte de usuarios no expertos. A continuación, se presenta el algoritmo en pseudocódigo:

Algorithm 3: Algoritmo SLS para Conjuntos Interpretables de Decisión

Input : Conjunto de datos de entrenamiento D, umbral de interpretabilidad θ
Output: Conjunto de reglas R

Función SLS(D, θ)**:**
  Inicializar el conjunto de reglas $R \leftarrow \emptyset$;
  while *exista alguna clase sin cubrir en* D **do**
    Generar candidatos de reglas para cada clase utilizando el conjunto de datos D;
    Evaluar cada regla candidata basada en precisión y cobertura;
    Seleccionar la regla que maximiza la precisión y cobertura y minimiza el solapamiento con reglas existentes en R;
    if *la regla cumple con el umbral de interpretabilidad* θ **then**
      Añadir la regla seleccionada a R;
    end
  end
  Ordenar el conjunto de reglas R para optimizar la aplicabilidad y minimizar la complejidad del modelo;
  return R

2.4. Factores que Afectan la Interpretabilidad

La interpretabilidad de un modelo de aprendizaje automático depende de dos factores principales: la transparencia del modelo y la presentación de los resultados (visualizaciones).

2.4.1. Transparencia del Modelo

La transparencia en los modelos de aprendizaje automático se refiere a la facilidad con la que se pueden entender los mecanismos internos del modelo y cómo estos conducen a una predicción específica. Un modelo se considera transparente cuando su funcionamiento puede ser descrito y comprendido por los usuarios sin necesidad de conocimientos técnicos avanzados [14].

Existen diferentes niveles de transparencia que influyen en la interpretabilidad de un modelo:

- *Transparencia simulable:* Implica que una persona pueda simular mentalmente las operaciones del modelo. Por ejemplo, los árboles de decisión pequeños o los conjuntos de reglas simples, como los *Interpretable Decision Sets* (IDS), permiten seguir cada paso de su cálculo en un tiempo razonable [12]. (Véase Figura 2.6).
- *Transparencia descriptiva:* Se refiere a la capacidad de explicar cómo el modelo toma decisiones en términos comprensibles para el usuario, descomponiendo el modelo en reglas más pequeñas y manejables. Esto es particularmente útil en los IDS, donde las reglas se presentan de manera independiente y clara [1]. (Véase Figura 2.10).
- *Transparencia algorítmica:* Se centra en la comprensión de los algoritmos que rigen el comportamiento del modelo y cómo estos afectan sus resultados. Modelos como los árboles de decisión (DT), las listas de decisiones (DL) y los IDS son

considerados algorítmicamente transparentes porque se basan en estructuras claramente definidas [28]. (Véase Figura 2.8).

La falta de transparencia puede generar problemas significativos en la interpretabilidad. Los modelos de caja negra, como las redes neuronales profundas, suelen ser complejos y opacos, dificultando su comprensión y aceptación en aplicaciones críticas [3]. En contraste, los modelos más simples y explicables, como los árboles de decisión y los IDS, proporcionan una mayor transparencia, lo que es esencial para garantizar decisiones confiables y justificables.

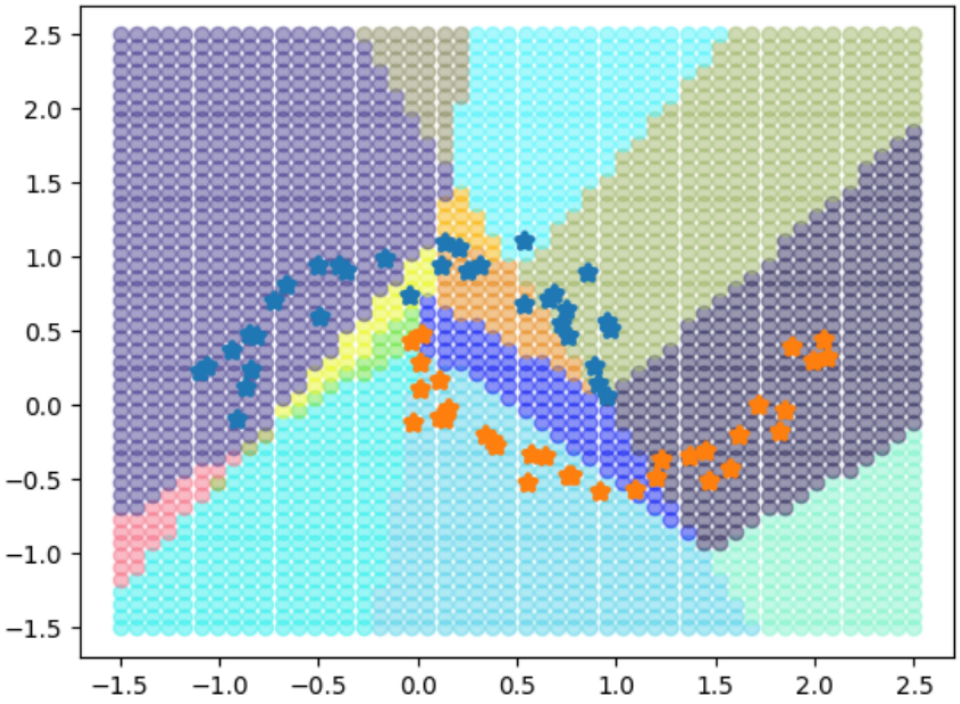


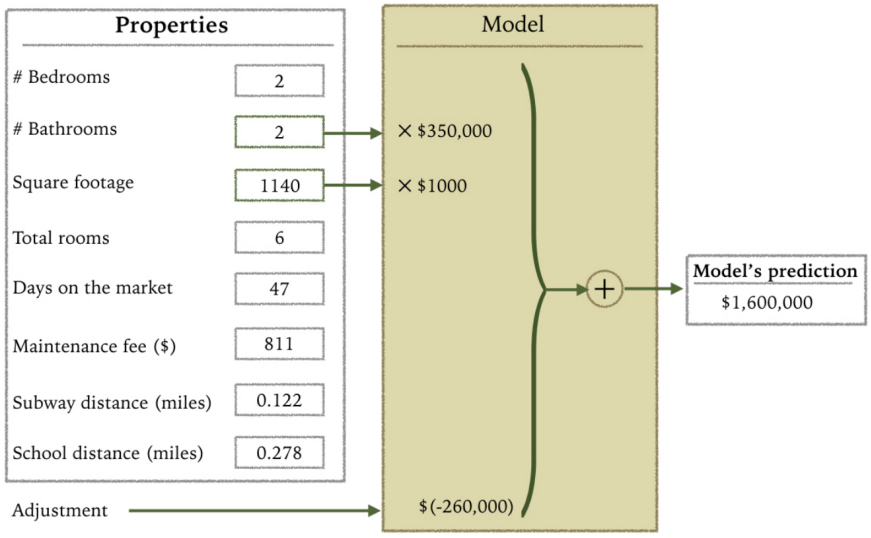
Figura 2.12: En esta gráfica de [29], se muestra cómo un árbol de decisión clasifica un conjunto de datos dividiendo el espacio de características en regiones distintas, representadas por diferentes colores. Esto ilustra cómo los árboles de decisión pueden mejorar la interpretabilidad de modelos complejos, como las redes neuronales profundas, al simplificar sus decisiones en reglas comprensibles. Los puntos indican las muestras de datos, mientras que las fronteras de color reflejan los límites de decisión del modelo, facilitando la comprensión de sus predicciones.

2.4.2. Confianza en Visualizaciones

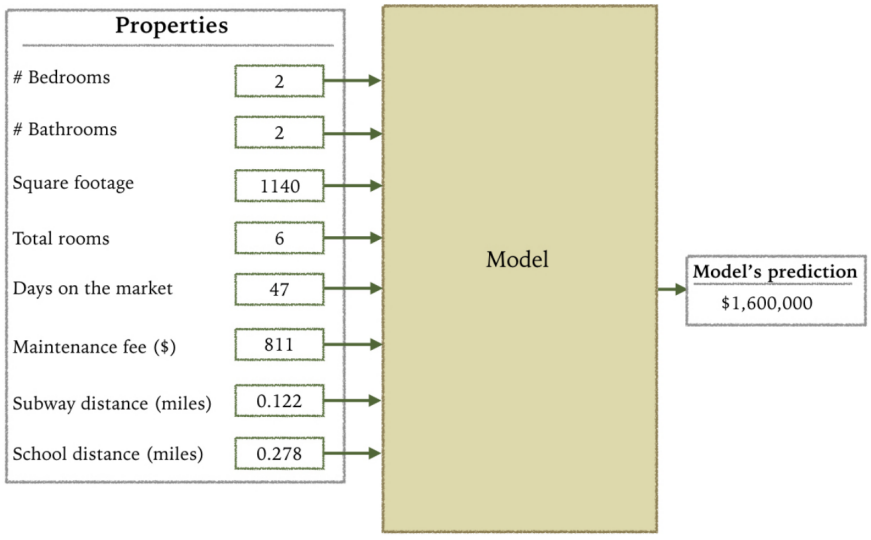
La visualización de datos juega un papel fundamental en la forma en que los usuarios interpretan y confían en los resultados de los modelos de aprendizaje automático. Estas visualizaciones pueden hacer que los modelos complejos sean más comprensibles al proporcionar una representación gráfica de los datos y las decisiones del modelo. Sin embargo, esta simplificación visual también conlleva ciertos riesgos que pueden afectar significativamente la confianza del usuario [30, 12].

Por un lado, las visualizaciones eficaces pueden mejorar la transparencia y facilitar la comprensión de las predicciones del modelo. Diagramas de árboles de decisión, gráficos de reglas y otros tipos de visualizaciones pueden ayudar a los usuarios a

seguir el proceso de toma de decisiones de un modelo, aumentando así su confianza en la precisión y confiabilidad de las predicciones. (Véase Figura 2.13).



(a) Modelo Transparente (CLEAR-2): Muestra los cálculos internos, lo que puede aumentar la confianza del usuario, pero también sobrecargar de información.

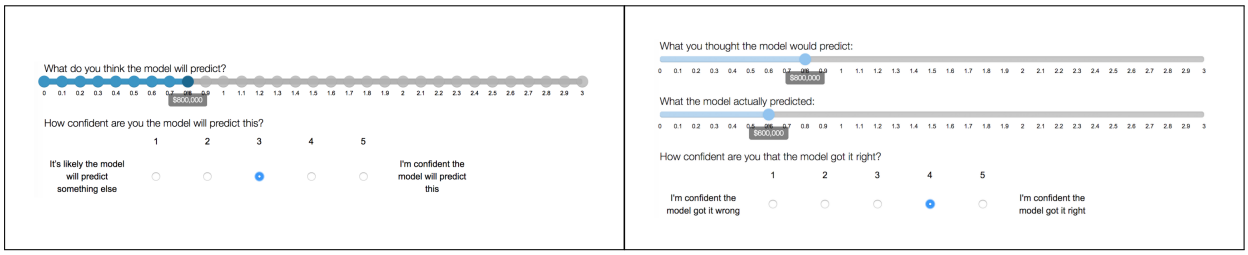


(b) Modelo de Caja Negra (BB-8): Oculta los cálculos internos, lo que reduce la transparencia, pero evita la sobrecarga cognitiva.

Figura 2.13: Adaptación de las condiciones experimentales de Poursabzi et al. [30], que muestran cómo la presentación del modelo afecta la confianza del usuario.

Por otro lado, la confianza en las visualizaciones puede llevar a una comprensión superficial o a una confianza excesiva en los resultados. Investigaciones han demostrado que los usuarios tienden a confiar más en modelos cuyos resultados están respaldados por visualizaciones atractivas, incluso si no comprenden completamente los algoritmos subyacentes [12]. Por ejemplo, en la Figura 2.14, se ilustra cómo la presentación de los resultados puede influir en la percepción de los usuarios. Este

fenómeno puede llevar a que los usuarios acepten las decisiones de los modelos sin cuestionarlas o sin detectar posibles errores, simplemente porque la visualización es persuasiva.



(a) Paso 1: Los participantes adivinan lo que el modelo predecirá y expresan su confianza en esa suposición.

(b) Paso 2: Los participantes ajustan su confianza en la predicción del modelo después de conocer el resultado.

Figura 2.14: Adaptación de las fases del experimento de Poursabzi et al. [30].

Además, las visualizaciones pueden ocultar la complejidad o ambigüedad de los datos, lo que resulta en una representación simplificada que no refleja completamente las incertidumbres o limitaciones del modelo. Como se observa en la Figura 2.15, es crucial que las visualizaciones se diseñen con cuidado para evitar inducir una falsa sensación de confianza o comprensión.

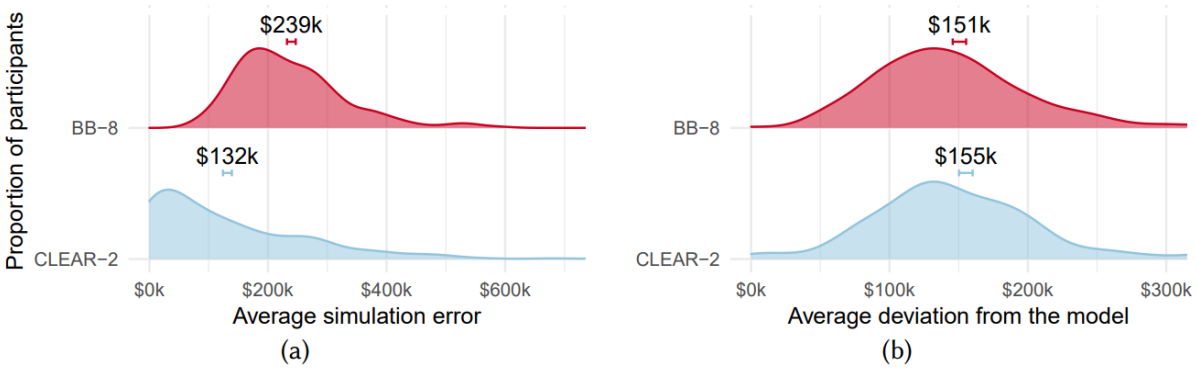


Figura 2.15: Resultados del Experimento 1 que comparan la percepción de los usuarios sobre dos tipos de modelos: un modelo de caja negra (BB-8) y un modelo transparente (CLEAR-2). En (a), se muestra el error promedio de simulación, donde los participantes tienden a percibir mayores errores en el modelo de caja negra en comparación con el modelo transparente. En (b), se presenta la desviación promedio de las predicciones del modelo respecto a los valores reales, indicando cómo los participantes evalúan la precisión del modelo según su nivel de transparencia. Adaptado de Poursabzi et al. [30].

2.5. Evaluación de la Interpretabilidad: Métodos y Métricas

Para garantizar que los usuarios finales comprendan cómo funcionan los modelos de IA, es crucial llevar a cabo estudios empíricos que evalúen su interpretabilidad [9].

2.5.1. Necesidad de Validación Empírica

La interpretabilidad no es un concepto absoluto; varía según el contexto, el modelo y las expectativas de los usuarios finales [31]. Por ello, no se puede asumir que un modelo es interpretativo sin estudios empíricos que lo validen, los cuales permiten evaluar cómo los usuarios interactúan con los modelos, comprenden sus decisiones y cómo esta comprensión afecta su confianza y capacidad para detectar errores o sesgos [12].

Estos estudios también identifican las características del modelo más relevantes para los usuarios y cómo estas influyen en su toma de decisiones, proporcionando información valiosa para adaptar las explicaciones del modelo a diferentes necesidades [32]. Las metodologías empíricas, como experimentos con usuarios, encuestas, entrevistas o evaluación de tareas, permiten medir la capacidad de los usuarios para entender y utilizar las salidas del modelo de manera efectiva, desarrollando así herramientas adaptadas a diversos contextos [9].

2.5.2. Métodos de Evaluación

Existen varios métodos para evaluar la interpretabilidad de los modelos de aprendizaje automático. Algunos de los enfoques más comunes incluyen:

- *Estudios de Usuario:* involucran experimentos con usuarios reales para evaluar cómo entienden y perciben la interpretabilidad de un modelo. Por ejemplo, se puede medir el tiempo que los usuarios tardan en comprender una predicción o su precisión al identificar errores en las predicciones del modelo. Este tipo de estudios es particularmente útil en entornos donde la decisión basada en el modelo tiene un impacto crítico, como en diagnósticos médicos [12].
- *Métricas Cuantitativas:* estas métricas miden características específicas de los modelos que se consideran proxies de interpretabilidad, tales como la simplicidad, la consistencia y la cobertura de las reglas en los modelos basados en reglas como los *Interpretable Decision Sets* (IDS) y los árboles de decisión (DT) [13]. Ejemplos incluyen el número de nodos en un árbol de decisión o la longitud de una regla en un conjunto de reglas. Las métricas cuantitativas permiten una evaluación objetiva de la interpretabilidad, aunque no siempre capturan completamente la percepción del usuario.
- *Evaluación Basada en Tareas:* evalúa cómo los usuarios utilizan el modelo para completar tareas específicas. Por ejemplo, se podría evaluar si los usuarios pueden hacer predicciones más precisas con la ayuda del modelo interpretativo o si pueden identificar correctamente los errores cuando se les presenta una visualización del modelo. Este enfoque es útil para validar la efectividad práctica de un modelo en escenarios del mundo real [9].
- *Encuestas y Cuestionarios:* herramientas como cuestionarios estandarizados pueden medir la percepción de los usuarios sobre la transparencia, la facilidad de comprensión y la confianza en el modelo [32]. Los cuestionarios pueden incluir preguntas sobre cuán claro es el modelo, cuán fácil es de entender, y cuánta confianza inspira en sus decisiones [12]. Estas herramientas son esenciales para capturar las percepciones subjetivas de los usuarios y ajustar los modelos interpretativos a sus necesidades específicas.

Combinar estos métodos proporciona una visión más completa y precisa de la interpretabilidad de los modelos de IA, adaptada a las necesidades y expectativas de diferentes usuarios y aplicaciones. Una evaluación integral de la interpretabilidad debe tener en cuenta tanto las métricas cuantitativas como los estudios cualitativos, para capturar tanto la objetividad de las características del modelo como las percepciones subjetivas de los usuarios [9, 12].

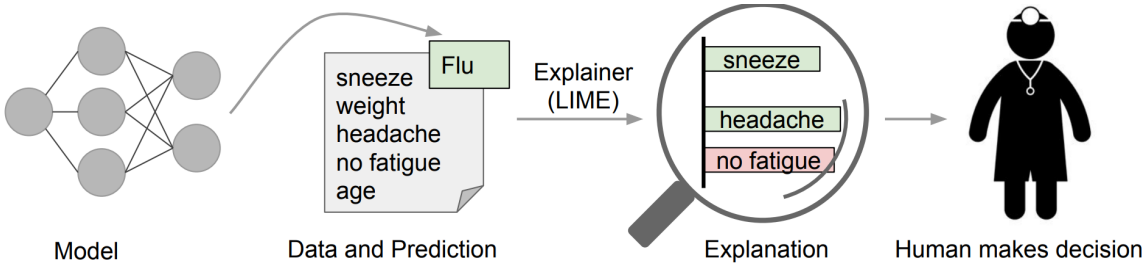


Figura 2.16: Explicación de predicciones individuales usando LIME. Un modelo predice que un paciente tiene gripe, y LIME resalta los síntomas en la historia del paciente que llevaron a la predicción. “Sneeze” (estornudo) y “headache” (dolor de cabeza) contribuyen a la predicción de “flu” (gripe), mientras que “no fatigue” (sin fatiga) es evidencia en contra de ella. Con esta información, un médico puede tomar una decisión informada sobre si confiar o no en la predicción del modelo.

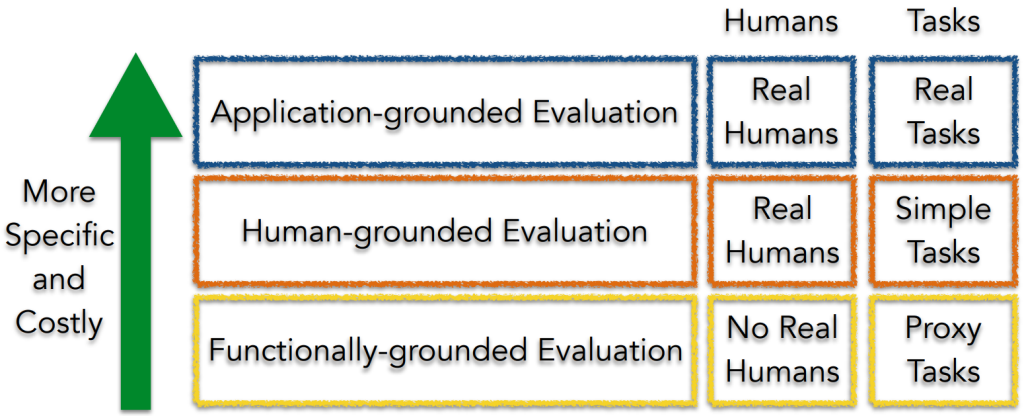


Figura 2.17: Clasificación de métodos de evaluación de la interpretabilidad, según Doshi-Velez y Kim (2017), en tres categorías: evaluación basada en aplicaciones, evaluación basada en humanos y evaluación basada en funciones. Cada categoría varía en términos de especificidad y costo, siendo las evaluaciones basadas en aplicaciones las más específicas y costosas, ya que implican tareas reales y usuarios humanos reales, mientras que las evaluaciones basadas en funciones son las menos costosas y específicas, al no requerir la participación de usuarios reales y emplear tareas proxy. Imagen tomada de [9]

Método	Modelo	Interpretabilidad	Usabilidad	Aplicación
Explicaciones de Saliencia	Caja Negra	Local	Media	Diagnóstico Médico
Gráficos de Dependencia Parcial	Caja Blanca/Negra	Global	Alta	Análisis Económico
Modelos de Reglas	Caja Blanca	Local/Global	Alta	Aplicaciones Regulatorias

Cuadro 2.1: Comparación simplificada de métodos de interpretabilidad de modelos de aprendizaje automático. La tabla clasifica tres métodos comunes de interpretabilidad: Explicaciones de Saliencia, Gráficos de Dependencia Parcial y Modelos de Reglas, según el tipo de modelo al que se aplican (caja negra o blanca), el nivel de interpretabilidad que proporcionan (local o global), su usabilidad y sus aplicaciones típicas. Esta clasificación se basa en estudios previos [12, 9, 13] que discuten la efectividad de cada método en diferentes contextos, permitiendo seleccionar el enfoque más adecuado según las necesidades de los usuarios finales y el tipo de modelo utilizado.

2.6. Resumen

En este capítulo, se han abordado los conceptos clave relacionados con la interpretabilidad en inteligencia artificial, destacando su importancia en aplicaciones críticas donde la transparencia y la confianza en los modelos son esenciales. Se han explorado enfoques intrínsecos para mejorar la interpretabilidad, como la *sparsidad*, la *simulabilidad*, la *modularidad* y la *parsimonia*, cada uno facilitando una mejor comprensión del proceso de toma de decisiones de los modelos.

También se han examinado métodos basados en reglas, como el *descubrimiento de subgrupos*, los *conjuntos de contraste* y los *patrones emergentes*, que identifican patrones diferenciados en los datos y proporcionan explicaciones claras de las decisiones del modelo, fundamentales para caracterizar las clases de manera comprensible para los usuarios finales.

Estos fundamentos teóricos proporcionan un marco sólido para el diseño del cuestionario de evaluación de interpretabilidad de este TFM.

Capítulo 3

Estado del Arte

En este capítulo se revisan las principales técnicas y herramientas en el campo de la Inteligencia Artificial Explicable (*XAI*). Este capítulo se divide en las siguientes secciones:

- *Sección 3.1*: Concepto, objetivos y relevancia de la explicabilidad en IA.
- *Sección 3.2*: Técnicas post-hoc como modelos subrogados, LIME y SHAP.
- *Sección 3.3*: Modelos intrínsecamente interpretables y optimización para la interpretabilidad.
- *Sección 3.4*: Evaluación mediante sparsidad y complejidad estructural.
- *Sección 3.5*: Métricas cualitativas y cuantitativas enfocadas en el usuario.
- *Sección 3.6*: Descripción de *InterpretML*, *Yellowbrick*, *Anchors* y *DiCE*.
- *Sección 3.7*: Síntesis de los enfoques clave revisados.

Esta revisión proporciona el marco teórico necesario para evaluar la interpretabilidad de los modelos en este TFM.

3.1. Inteligencia Artifcial Explicable

La Inteligencia Artificial Explicable (XAI, por sus siglas en inglés de *eXplainable Artificial Intelligence*) se refiere a un conjunto de métodos y técnicas diseñados para hacer que los modelos de aprendizaje automático sean comprensibles e interpretables para los usuarios humanos. El objetivo es permitir a los usuarios confiar y gestionar de manera efectiva los sistemas basados en aprendizaje automático [10].

En los últimos años, la comunidad científica se ha interesado en desarrollar estos métodos debido a la creciente utilización de la inteligencia artificial, lo que incrementa la necesidad de asegurar su uso ético y seguro. En este contexto, XAI se enfoca en cuatro objetivos principales [33]:

- *Confianza y aceptación*. Para que los sistemas de IA sean ampliamente aceptados, es fundamental que los usuarios confíen en sus decisiones. Explicar cómo y por qué un modelo toma una decisión específica es crucial para construir esta confianza.

- *Transparencia*. En muchas aplicaciones, especialmente en dominios sensibles como la medicina, la justicia y las finanzas, es esencial que los modelos de IA sean transparentes. Esto facilita no solo la detección y corrección de errores, sino también asegura que las decisiones sean justas y no discriminatorias.
- *Cumplimiento normativo*. Diversas regulaciones, como el Reglamento General de Protección de Datos (GDPR) en Europa, exigen que las decisiones automatizadas sean explicables. Esto garantiza que los usuarios tengan derecho a una explicación clara y comprensible de cómo se toman las decisiones que les afectan.
- *Mejora del modelo*. Comprender cómo funciona un modelo permite a los desarrolladores identificar áreas de mejora y ajustar los modelos para obtener mayor rendimiento y precisión.

Modelo	**Enfoque**	**Características**
Caja Negra	Modelos Subrogados	Uso de modelos más simples para aproximar las predicciones
	Métricas del Efecto	Evaluación del impacto de variables (locales y globales)
	Explicabilidad basada en Ejemplos	Uso de ejemplos específicos para explicar predicciones
Transparentes	Intrínsecamente Interpretables	Diseñados para ser comprensibles desde su construcción
	Optimización para Interpretabilidad	Ajustados específicamente para mejorar la claridad y explicación

Cuadro 3.1: Comparación de Técnicas en XAI Según el Tipo de Modelo.

Los métodos y técnicas desarrollados en XAI se pueden agrupar en dos categorías principales según el tipo de modelo: modelos de caja negra y modelos transparentes [33]. Estas categorías se subdividen en técnicas específicas, como se muestra en el Cuadro 3.1.

3.2. XAI para Modelos de Caja Negra

Aunque este trabajo no se enfoca en métodos aplicables a modelos de caja negra, es importante mencionarlos para brindar un panorama completo de la Inteligencia Artificial Explicable (XAI). En general, un *modelo de caja negra* es aquel cuyo funcionamiento interno es desconocido o no es interpretable. Esto puede deberse a que su acceso está restringido (por propiedad intelectual) o porque su mecanismo de predicción es inherentemente complejo, como en el caso de las redes neuronales [33, 14].

Existen tres tipos principales de técnicas para abordar la explicabilidad en modelos de caja negra [33]:

- *Modelos subrogados*: Se refiere a la creación de un modelo más sencillo e interpretable que imita las predicciones del modelo de caja negra. Este modelo subrogado puede aproximar el comportamiento del modelo original de manera *local* (para un subconjunto específico de datos de entrada) o *global* (para el conjunto completo de datos).

- *Explicabilidad basada en ejemplos*: Busca explicar una predicción específica proporcionando ejemplos similares y contraejemplos. Esto ayuda al usuario a entender cómo diferentes características influyen en una predicción en particular.
- *Métricas del efecto de las variables sobre la predicción*: Estas métricas buscan cuantificar la influencia de cada variable de entrada en las predicciones del modelo. Algunas técnicas destacadas en esta categoría son:
  - *LIME (Local Interpretable Model-agnostic Explanations)*: Genera explicaciones locales al identificar cuáles características influyen en una predicción específica y su efecto (positivo o negativo) [32].
  - *SHAP (SHapley Additive exPlanations)*: Calcula la contribución de cada característica a la predicción utilizando conceptos de la Teoría de Juegos [34].
  - *Eliminación de características*: Consiste en eliminar una característica del modelo (generalmente, reentrenando el modelo) para observar cómo cambian las predicciones sin dicha característica [33].
  - *Ocultamiento de características*: Similar a la eliminación, pero se ocultan parcial o totalmente algunas características, utilizado principalmente en redes neuronales convolucionales para entender qué partes de los datos son más relevantes [33].

3.3. XAI para Modelos Transparentes

Los modelos transparentes son aquellos que no solo exponen el mecanismo mediante el cual generan las predicciones, sino que este proceso es fácilmente comprensible para un usuario humano. Un ejemplo clásico es la regresión lineal: es fácil entender cómo este modelo genera una predicción (mediante multiplicaciones por coeficientes y una suma), y también es fácil interpretar el significado de estos coeficientes (el valor de cada coeficiente indica el efecto de la variable correspondiente sobre la predicción).

En principio, se presume que los modelos transparentes son interpretables, es decir, que un humano puede entender el significado de los elementos del proceso de decisión y sus relaciones, como en el ejemplo de la regresión lineal. Sin embargo, en la práctica, un modelo transparente puede ser tan complejo que sobrepase la capacidad cognitiva de una persona. Por ejemplo, una regresión lineal con mil variables resulta difícil de interpretar, ya que es complejo entender el impacto de tantos factores en una predicción.

Para mejorar la interpretabilidad de estos modelos, se pueden aplicar técnicas de regularización como la *Lasso*, que promueven la *sparsidad* al reducir el número de variables que tienen un impacto significativo en las predicciones. Esta técnica puede observarse en la Figura 2.1, donde se ilustra cómo Lasso minimiza la inclusión de variables irrelevantes, mejorando así la simplicidad y la claridad del modelo.

De esta forma, se facilita que el usuario enfoque su atención solo en las características más importantes, lo que mejora la comprensión general del modelo.

3.4. Métricas de Evaluación de la Explicabilidad en Modelos Transparentes

Esta sección discute brevemente métricas cuantitativas propuestas en la literatura para evaluar y mejorar la explicabilidad de modelos transparentes. Aunque estas métricas se justifican intuitivamente, es importante destacar que su efectividad puede depender del contexto específico del modelo y del usuario, como se ha señalado en los fundamentos teóricos (ver Figuras 2.1, 2.2, y 2.4).

3.4.1. Número de Características

El *número de características* utilizadas por un modelo es una de las métricas más comúnmente aplicadas para medir la explicabilidad [35, 1, 30]. Este concepto se relaciona estrechamente con la *sparsidad*, ilustrada en la Figura 2.1, que busca reducir el número de parámetros no nulos en un modelo para facilitar la comprensión del papel específico de cada variable en las predicciones. En modelos como las regresiones lineales (y otros modelos aditivos, como los *Generalized Additive Models*), es posible promover la *sparsidad* introduciendo un término de regularización en la función de pérdida, tal como se explicó en los fundamentos teóricos.

Esta idea también se extiende a otros tipos de modelos, como los árboles de decisión. Por ejemplo, [36, 35] proponen el uso de técnicas de regularización para favorecer árboles de decisión *esparsos* (del inglés *sparse*), utilizando programación dinámica y teoremas que garantizan un desempeño mínimo. Este enfoque promueve la simplicidad y la *parsimonia*, representada en la Figura 2.4, utilizando la menor cantidad de variables posible sin comprometer el rendimiento.

En el caso de conjuntos de decisión, [1] proponen tres métricas: una para penalizar la función de pérdida basada en el número de reglas individuales en el conjunto de decisiones, otra para penalizar la longitud de las reglas, y una última para maximizar la cobertura de una regla (es decir, aplicarla al mayor número posible de datos). Este enfoque de minimización también se alinea con los principios de *parsimonia* y *modularidad* (ver Figura 2.4), al reducir la complejidad del modelo mientras se mantiene la efectividad de la predicción.

3.4.2. Complejidad de la Estructura del Modelo

La *complejidad de la estructura* de un modelo transparente también afecta su interpretabilidad, como se discutió en los fundamentos teóricos en relación con la *simulabilidad* (ver Figura 2.2). Diferentes configuraciones estructurales pueden hacer que un modelo sea más o menos fácil de explicar.

Por ejemplo, en los árboles de decisión, tanto el número de ramas por nodo como la profundidad del árbol influyen en su interpretabilidad. Un árbol con muchas ramas o de gran profundidad puede ser difícil de entender, contraviniendo el principio de *simulabilidad* (ver Figura 2.2), que sugiere que los modelos deben ser fáciles de reproducir y comprender por humanos. Para mitigar esto, [35] recomienda el uso de árboles binarios (cada nodo tiene solo dos ramas), y la programación dinámica para limitar la profundidad del árbol, controlando así la complejidad del modelo.

Para los conjuntos de decisión, [1] sugiere el uso de listas simples de reglas, sin

anidación, ya que se considera que son más fáciles de entender que las secuencias de reglas anidadas. Esta preferencia por la simplicidad y la claridad se alinea con la noción de *parsimonia* (ver Figura 2.4), donde se favorecen las soluciones más simples y directas.

3.5. Métricas para la Evaluación Humana de la Interpretabilidad

Esta sección presenta las métricas propuestas en la literatura para la evaluación humana de la interpretabilidad. Dichas métricas son esenciales para determinar si los usuarios pueden entender un modelo de aprendizaje automático, especialmente cuando se trata de modelos transparentes, como los discutidos en la Figura 2.3 sobre la modularidad y en la Figura 2.2 sobre la simulabilidad. Estas métricas se dividen en dos categorías principales: cualitativas y cuantitativas, dependiendo de su enfoque para medir el entendimiento.

3.5.1. Métricas Cualitativas

Las métricas cualitativas buscan evaluar la capacidad del usuario para comprender el modelo mediante la recopilación de datos no numéricos. Generalmente, estos datos se obtienen a través de entrevistas semi-estructuradas o preguntas contextualizadas [12]. Tal como se ilustra en la Figura 2.2, los modelos que favorecen una estructura simple, como los árboles de decisión, facilitan la comprensión del proceso de decisión por parte del usuario. En este contexto, se puede emplear una metodología en la que el usuario tenga acceso al modelo o a visualizaciones relacionadas, y un investigador formule preguntas, asigne tareas o simplemente escuche al usuario para determinar su grado de entendimiento.

Por ejemplo, [12] realizó entrevistas semi-estructuradas para identificar las principales dificultades que enfrentan los científicos de datos al interpretar modelos complejos. Basándose en estos resultados, se diseñaron entrevistas adicionales, utilizando el mismo enfoque mostrado en la Figura 2.5, para determinar cómo los diferentes grupos de usuarios perciben la interpretabilidad de los modelos en función de su contexto y necesidades específicas.

3.5.2. Métricas Cuantitativas

Las métricas cuantitativas, a diferencia de las cualitativas, se enfocan en la recopilación de datos numéricos que puedan ser analizados de manera estadística. Similar a cómo se evaluó la interpretabilidad mediante la reducción de características en la Figura 2.1, estas métricas intentan medir la capacidad del usuario para comprender un modelo a través de la exactitud de sus predicciones o la eficiencia con la que completan tareas basadas en el modelo.

La métrica más común es la exactitud o el error de desviación [1, 30], que consiste en solicitar al usuario que realice predicciones utilizando el modelo, como se describe en la Figura 2.3, donde los Modelos Aditivos Generalizados (GAMs) descomponen las predicciones en componentes interpretables. Esta métrica permite determinar si un usuario es capaz de seguir el razonamiento del modelo y entender cómo se llega a una predicción específica.

Adicionalmente, se emplean otras métricas como el *error de simulación* [30], en las que se muestra al usuario una serie de ejemplos y se le pide realizar predicciones para evaluar su capacidad de aprendizaje sin ayuda del modelo. De manera complementaria, el uso de cuestionarios como el NASA-TLX (*Task Load Index*) [37], que puede ayudar a cuantificar la carga cognitiva que experimentan los usuarios al interactuar con el modelo, proporcionando una medida indirecta de la interpretabilidad.

En conjunto, estas métricas ofrecen una visión integral de cómo los usuarios perciben la interpretabilidad de los modelos de aprendizaje automático, conectando las técnicas teóricas presentadas en la sección de fundamentos con la evaluación práctica de la experiencia del usuario.

3.6. Herramientas de Interpretabilidad para Modelos Transparentes

A diferencia de los modelos de caja negra, que requieren explicaciones post-hoc, los modelos transparentes permiten interpretar directamente sus decisiones debido a su estructura lógica. Sin embargo, aunque estos modelos, como los árboles de decisión (DT) y los Interpretable Decision Sets (IDS), ya son comprensibles, la claridad en la presentación y las visualizaciones influyen en cómo los usuarios perciben su interpretabilidad [33, 14]. La manera en que se presentan los resultados puede facilitar o dificultar la identificación de patrones, la detección de errores y la confianza en las decisiones.

Para abordar este reto, se han desarrollado herramientas que facilitan tanto el análisis de los modelos transparentes como la exploración de cómo las visualizaciones y la estructura del modelo afectan la experiencia del usuario. Estas herramientas combinan enfoques visuales y técnicos que mejoran la comprensión y fomentan la confianza en las predicciones.

A continuación, se describen herramientas clave en el campo de la inteligencia artificial explicable que facilitan el análisis y optimización de la interpretabilidad de modelos transparentes. Estas herramientas permiten evaluar cómo la presentación visual y la estructura de un modelo influyen en la percepción de interpretabilidad y la confianza del usuario en los sistemas de IA. Además, proporcionan una base sólida para estudiar los modelos transparentes de este trabajo, aplicando metodologías prácticas y enfoques visuales que mejoran tanto la claridad como la comprensión de los modelos de decisión.

3.6.1. InterpretML

InterpretML, propuesto por Nori et al. [22], es un marco de código abierto diseñado para proporcionar interpretabilidad tanto en modelos de caja blanca como en técnicas de explicabilidad post-hoc para modelos de caja negra. Este marco unifica diversas técnicas y permite a los investigadores y profesionales comparar diferentes enfoques de interpretabilidad a través de una API integrada y visualizaciones interactivas.

InterpretML ofrece dos tipos principales de interpretabilidad:

- Modelos de caja blanca: Estos modelos son transparentes desde su construcción, permitiendo que los usuarios comprendan directamente cómo cada ca-

racterística influye en las predicciones. Ejemplos incluyen los modelos aditivos generalizados (GAM) y los modelos lineales. En la Figura 2.3 se muestra un ejemplo de cómo la característica 'Edad' afecta la predicción en un modelo EBM, proporcionando una visualización clara de las relaciones entre características y resultados.

- Explicabilidad para modelos de caja negra: Utilizando técnicas post-hoc como dependencia parcial, LIME y SHAP, InterpretML proporciona interpretaciones locales y globales de modelos más complejos, sin necesidad de conocer su estructura interna. La Figura 2.5 compara visualizaciones generadas por GAM (arriba) y SHAP (abajo), mostrando cómo diferentes enfoques interpretativos pueden ajustarse a audiencias y contextos específicos.

Una de las contribuciones clave de InterpretML es el Explainable Boosting Machine (EBM), un modelo de caja blanca basado en un modelo aditivo generalizado (GAM) que utiliza técnicas modernas como bagging y boosting para ajustar cada característica iterativamente, reduciendo la correlación entre ellas. Esto permite que el EBM logre niveles de precisión comparables a los de modelos de caja negra como los bosques aleatorios o XGBoost, pero manteniendo la interpretabilidad total.

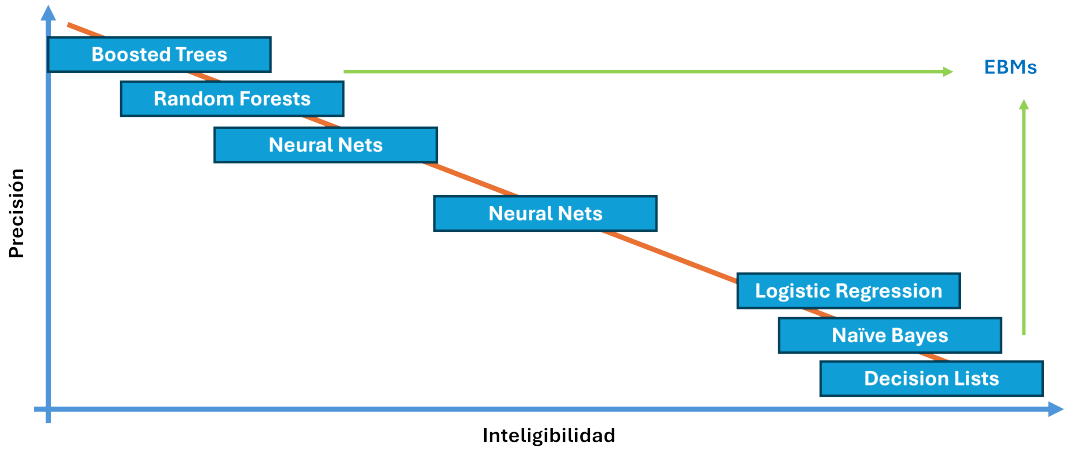


Figura 3.1: Gráfico comparativo de precisión e inteligibilidad de modelos de aprendizaje automático, inspirado en un gráfico presentado por [38].

El EBM es modular y permite visualizar cómo cada característica influye en las predicciones, facilitando la comprensión para los usuarios mediante gráficos claros, como los mostrados en la Figura 2.3, que ilustran la descomposición de una predicción individual y cómo cada característica contribuye de forma independiente.

InterpretML presenta varias ventajas:

- Facilidad de comparación: La API unificada facilita la comparación de diferentes algoritmos de interpretabilidad, con integración sencilla a frameworks como Scikit-learn.
- Interoperabilidad: Es compatible con herramientas como Jupyter Notebook y plotly, permitiendo una interacción intuitiva con los modelos y sus explicaciones.
- Visualización interactiva: InterpretML ofrece un panel de control interactivo que

permite explorar visualmente las explicaciones generadas, ayudando a los usuarios a entender fácilmente las contribuciones de cada característica.

InterpretML es reconocido por su capacidad para combinar precisión y transparencia. Esto lo convierte en una herramienta relevante para aplicaciones que requieren decisiones responsables.

3.6.2. Yellowbrick

Yellowbrick es una biblioteca de visualización diseñada para facilitar la evaluación e interpretación de modelos de aprendizaje automático creados con Scikit-learn. Esta herramienta ofrece una amplia gama de visualizaciones que permiten a los usuarios entender mejor el comportamiento de los modelos, ayudando en la toma de decisiones y mejorando la interpretabilidad de los mismos [39].

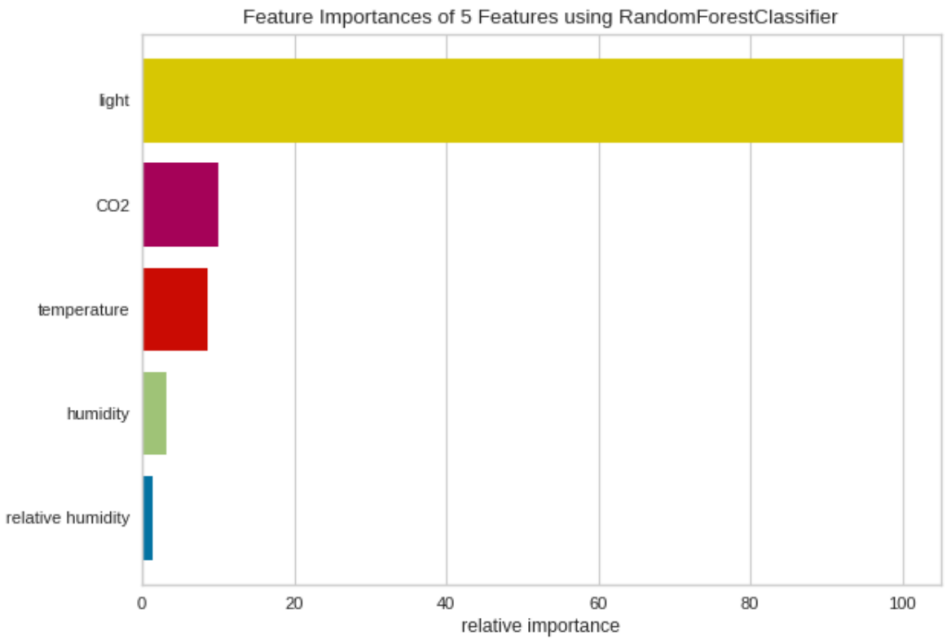


Figura 3.2: Visualización de la importancia de características generada por Yellowbrick utilizando un modelo *RandomForestClassifier*. El gráfico muestra la influencia relativa de cinco características sobre las predicciones del modelo. La característica 'light' tiene la mayor importancia relativa, seguida de 'CO2' y 'temperature', lo que indica su fuerte contribución a las decisiones del modelo [39].

Yellowbrick se enfoca en generar gráficos que muestran cómo las características y el rendimiento de los modelos afectan los resultados, lo cual es esencial para los modelos transparentes como los Árboles de Decisión (DT) y los Conjuntos de Decisión Interpretables (IDS) utilizados en este trabajo. Entre sus funcionalidades más destacadas se incluyen:

- *Visualización de la importancia de características*: Yellowbrick permite representar gráficamente la importancia relativa de las características dentro de un modelo, como se muestra en la Figura 3.2. En este TFM, esta herramienta es fundamental para comprender el impacto de las distintas características en las

predicciones de los modelos de decisión. En particular, se ha empleado para identificar las características más influyentes dentro del dataset de matemáticas utilizado en este proyecto.

- *Curvas de validación y curvas de aprendizaje*: Estas curvas permiten observar cómo el rendimiento del modelo varía según los valores de los hiperparámetros y cómo mejora a medida que recibe más datos de entrenamiento. Este tipo de visualización es útil para diagnosticar problemas de sobreajuste o subajuste, los cuales pueden afectar la interpretabilidad del modelo.
- *Matrices de confusión y reportes de clasificación*: Facilitan la visualización de cómo el modelo clasifica correctamente e incorrectamente las instancias, lo que ayuda a comprender mejor el comportamiento del modelo en cada clase.
- *Curvas ROC y AUC*: Permiten evaluar el rendimiento de los modelos de clasificación al mostrar el compromiso entre la tasa de verdaderos positivos y la tasa de falsos positivos en diferentes umbrales de decisión. Son especialmente útiles para evaluar la capacidad de los modelos para manejar errores de clasificación.

Yellowbrick ofrece ventajas significativas, entre las que destaca su adecuada integración con scikit-learn. Esto permite a los usuarios generar visualizaciones sin una configuración adicional compleja, y proporciona una interfaz fácil de usar que se adapta tanto a principiantes como a expertos, facilitando así la evaluación visual de los modelos.

3.6.3. Anchors

Anchors, propuestos por Ribeiro, Singh y Guestrin [40], son reglas locales que explican predicciones individuales de un modelo a través de condiciones suficientes. Estas reglas aseguran que si las condiciones de un *anchor* se cumplen, la predicción del modelo será la misma con alta probabilidad. Los *anchors* son especialmente útiles en modelos de caja negra o en tareas donde la interpretabilidad es fundamental. Aunque en este trabajo no implementamos *Anchors*, su relevancia para el campo de la IA explicable es notable, ya que proporcionan explicaciones simples y de alta precisión que son fácilmente comprensibles por los usuarios.

En la siguiente tabla se muestra un ejemplo adaptado del artículo original, donde se utiliza *Anchors* para etiquetar la parte del discurso de la palabra “play” en diferentes contextos.

Instancia	**Condición**	**Predicción**
I want to play(V) ball.	La palabra previa es **PARTICLE**	play es **VERBO**.
I went to a play(N) yesterday.	La palabra previa es **DETERMINANTE**	play es **SUSTANTIVO**.
I play(V) ball on Mondays.	La palabra previa es **PRONOMBRE**	play es **VERBO**.

Cuadro 3.2: Ejemplo de Anchors para la etiqueta de parte del discurso de la palabra “play” (adaptado de Ribeiro, Singh y Guestrin [40]).

En este ejemplo, el modelo predice si la palabra “play” es un verbo o un sustantivo dependiendo de la palabra que la precede, generando reglas locales que son fáciles de comprender y aplicar. Estos *anchors* proporcionan una explicación clara sobre

el comportamiento del modelo en casos específicos, lo que mejora la confianza y la interpretabilidad para los usuarios finales.

3.6.4. DiCE: Explicaciones Contrafactuales Diversas

DiCE (*Diverse Counterfactual Explanations*) es un marco propuesto por Mothilal, Sharma y Tan [41] para generar explicaciones contrafactuales que ayuden a los usuarios a comprender el comportamiento de modelos de aprendizaje automático. DiCE se centra en proporcionar múltiples explicaciones contrafactuales diversas, que exploran varias formas en las que se podría modificar una instancia para obtener una predicción diferente. Este enfoque permite a los usuarios entender qué cambios mínimos en los atributos de entrada podrían alterar la decisión del modelo, haciendo que este sea más interpretable.

Una *explicación contrafactual* responde a la pregunta: "¿Qué habría que cambiar en esta instancia para obtener un resultado diferente?". DiCE es capaz de generar varias explicaciones contrafactuales que muestran diferentes caminos posibles para lograr un cambio en la predicción. Esto es particularmente útil cuando hay varias combinaciones de características que pueden influir en el resultado, ya que proporciona al usuario un conjunto diverso de escenarios.

DiCE se basa en los siguientes principios:

- *Diversidad*: DiCE no genera una única explicación contrafactual, sino varias. Cada una ofrece una manera diferente de alterar la predicción de un modelo. Esta diversidad es útil para evitar que el usuario se enfoque solo en una explicación que podría no ser la mejor o más factible.
- *Flexibilidad*: Los usuarios pueden especificar qué características se pueden cambiar y cuáles no. Esto es útil en casos donde ciertos atributos son fijos (por ejemplo, la edad de una persona) y no pueden ser modificados en las explicaciones contrafactuales.
- *Compatibilidad con modelos de caja negra*: DiCE es un enfoque agnóstico al modelo, lo que significa que puede ser aplicado a una variedad de modelos, incluidos los modelos de caja negra.
- *Optimización*: DiCE utiliza un algoritmo de optimización para generar explicaciones que minimizan el número de cambios necesarios en la instancia original. De esta forma, se realizan interpretaciones más realistas y fáciles de entender.

DiCE emplea un enfoque basado en la búsqueda de ejemplos contrafactuales cercanos que resulten en un cambio en la predicción del modelo. Dados los inputs de una instancia x y una predicción del modelo $f(x)$, DiCE busca generar nuevos ejemplos x' que pertenezcan a una clase diferente y que estén lo más cerca posible de la instancia original. Para lograr esto, DiCE resuelve el siguiente problema de optimización:

$$\min_{x'} d(x, x') \quad \text{sujeto a} \quad f(x') \neq f(x)$$

donde $d(x, x')$ es una métrica de distancia que asegura que los ejemplos contrafactuales sean similares a la instancia original. El objetivo es generar instancias con-

trafactuales x' que difieran lo menos posible de x y que den lugar a una predicción diferente por parte del modelo f.

Como ejemplo, consideremos un modelo de predicción de aprobación de préstamos. Si el modelo predice que un cliente no será aprobado para un préstamo, DiCE puede generar múltiples explicaciones contrafactuales mostrando qué atributos del cliente deben cambiarse para que la solicitud sea aprobada. Por ejemplo, una explicación podría sugerir que aumentando los ingresos anuales y reduciendo la deuda actual, el cliente podría obtener una aprobación. Otra explicación podría sugerir que reducir el número de créditos activos mejoraría las probabilidades de aprobación.

Atributo Original	**Atributo Modificado (Contrafactual 1)**	**Atributo Modificado (Contrafactual 2)**
Ingresos anuales: $35,000	Ingresos anuales: $50,000	Ingresos anuales: $35,000
Deuda actual: $10,000	Deuda actual: $5,000	Deuda actual: $10,000
Créditos activos: 3	Créditos activos: 3	Créditos activos: 1

Cuadro 3.3: Ejemplo de explicaciones contrafactuales generadas por DiCE para un modelo de aprobación de préstamos. Se muestran múltiples combinaciones de atributos que podrían llevar a una predicción diferente.

Aunque DiCE es eficaz para generar explicaciones contrafactuales diversas, no garantiza que todas las explicaciones sean factibles o realistas en un contexto del mundo real. Por ejemplo, en el caso de la predicción de préstamos, aumentar los ingresos anuales puede no ser una opción viable a corto plazo para muchos individuos. Por lo tanto, es crucial que los usuarios de DiCE evalúen la plausibilidad de las explicaciones generadas.

DiCE es una herramienta poderosa para la IA explicable, permitiendo a los usuarios explorar cómo pequeños cambios en las entradas pueden alterar los resultados del modelo, y proporcionando una visión más completa del comportamiento del modelo a través de explicaciones diversas.

3.7. Resumen

En este estado del arte se han revisado las principales herramientas y enfoques de XAI, con especial énfasis en los modelos transparentes. También, se han explorado herramientas clave como InterpretML y Yellowbrick, que permiten visualizar y analizar la interpretabilidad de estos modelos, facilitando su comprensión por parte de los usuarios.

Asimismo, se han discutido métodos como Anchors, que proporciona reglas locales explicativas, y DiCE, que genera explicaciones contrafactuales diversas, destacando cómo estas técnicas ayudan a mejorar la transparencia y confianza en las predicciones de los modelos.

Este capítulo proporciona los enfoques más relevantes para medir cómo los usuarios perciben la interpretabilidad de los modelos utilizados.

Capítulo 4

Metodología

En este capítulo se describe la metodología utilizada para evaluar la interpretabilidad de dos modelos de decisión transparentes: Árboles de Decisión (*Decision Trees*, *DT*) e Interpretable Decision Sets (*IDS*). El objetivo principal es analizar cómo los usuarios interpretan las reglas generadas por estos modelos y en qué medida dichas reglas son útiles para comprender las predicciones y detectar posibles errores. Para ello, se diseñó y desarrolló una herramienta web personalizada que implementa un cuestionario para capturar la percepción de los usuarios sobre la claridad y utilidad de las explicaciones.

4.1. Diseño General del Estudio

La metodología de este trabajo sigue un flujo de ocho etapas que integran análisis técnicos y subjetivos para evaluar la interpretabilidad de modelos explicables en un contexto educativo.

Figura 4.1: Flujo metodológico seguido en este TFM.

A continuación, se describen las etapas desarrolladas:

1. *Preparación del Dataset [4.2]:* Se utilizó el *Student Performance Dataset* para realizar tareas de preprocesamiento y análisis exploratorio. Se identificaron características relevantes mediante técnicas como la matriz de correlación y modelos DT iniciales, seleccionando seis variables clave para los análisis posteriores.

2. *Desarrollo de los Modelos [4.3]:* Se implementaron dos modelos *DT* utilizando las herramientas *Scikit-learn* e *InterpretML*, y un modelo *IDS* basado en programación lineal entera mixta (MILP). Se generaron reglas y predicciones para ambos enfoques, priorizando claridad y simplicidad.

3. *Evaluación Técnica de los Modelos [4.4]:* Se analizaron las métricas de rendimiento predictivo, como precisión, *recall*, *F1-score* y *accuracy*, así como propiedades

estructurales (parsimonia, cobertura y solapamiento). Estos análisis proporcionaron una base cuantitativa para evaluar la interpretabilidad técnica.

4. *Diseño y Desarrollo del Cuestionario [4.5]:* A partir de los resultados de los modelos con las instancias de prueba, se seleccionaron las observaciones y se formularon preguntas de acuerdo con las categorías *Exactitud*, *Ambigüedad* y *Error* para capturar la percepción de los usuarios sobre la interpretabilidad de los modelos. Adicionamlmente, se generaron visualizaciones para complementar el cuestionario, se añadieron preguntas de seguimiento y se planteó la estructura del reporte para su posterior análisis.

5. *Desarrollo de la Herramienta de Interpretabilidad [4.6]:* Se diseñó una herramienta web personalizada para implementar el cuestionario, registrar las respuestas de los usuarios y medir los tiempos de participación. También se automatizó la generación del reporte de la evaluación de la interpretabilidad de los modelos.

6. *Pruebas de la Herramienta de Interpretabilidad [4.7]:* Se realizaron pruebas internas de la herramienta para validar su funcionalidad y desempeño. Estas pruebas incluyeron la captura de respuestas de usuarios ficticios, la generación de reportes en formato Excel y la verificación de los tiempos de respuesta registrados en la base de datos.

7. *Análisis de Resultados [5]:* Los datos obtenidos durante las pruebas se analizaron cuantitativa y cualitativamente. Se compararon métricas de interpretabilidad, tiempos de respuesta y preferencias de los usuarios para identificar los aspectos destacados de cada modelo.

8. *Conclusiones [6]:* Finalmente, se validaron las hipótesis relacionadas con la percepción de interpretabilidad y se propusieron recomendaciones para futuros desarrollos de herramientas destinadas a la evaluación de modelos transparentes en inteligencia artificial.

4.2. Preparación del Dataset

El conjunto de datos utilizado en este trabajo es el dataset de Matemáticas del *Student Performance Dataset* [42], el cual contiene información académica, social y demográfica de estudiantes. Este dataset fue seleccionado debido a su riqueza en características que permiten evaluar modelos interpretables en un problema de clasificación binaria, alineado con los objetivos de este trabajo. En total, se dispone de 30 características tras excluir las variables *G1* y *G2*, las cuales están correlacionadas con la calificación final (*G3*), utilizada como variable objetivo.

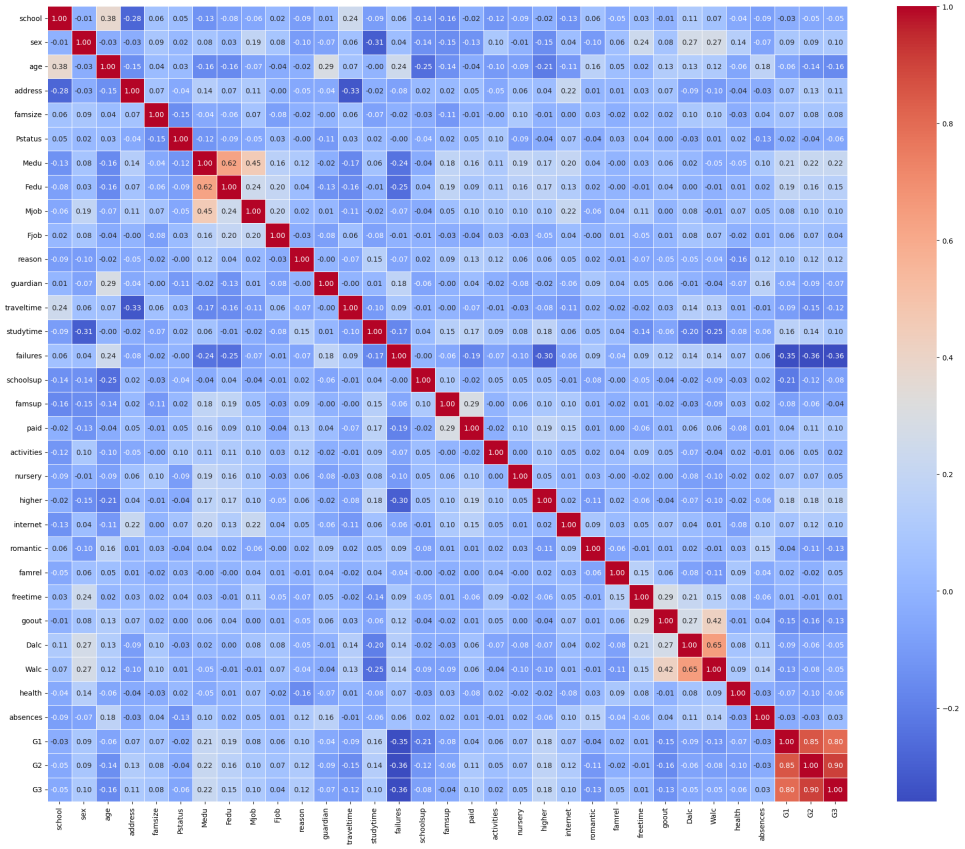
Se realizó un análisis exploratorio para entender la estructura del dataset y las relaciones entre las características. La matriz de correlación (Figura 4.2b) reveló los siguientes hallazgos principales:

- *Relaciones clave:* Características como *studytime* y *failures* presentaron correlaciones significativas con *G3*, justificando su inclusión en los modelos finales.

- *Distribuciones desequilibradas:* Variables como *reason_reputation* y *guardian* mostraron distribuciones heterogéneas, consideradas durante el preprocesamiento.

#	Atributo	Descripción
1	school	Escuela
2	sex	Sexo del estudiante
3	age	Edad del estudiante
4	address	Tipo de dirección
5	famsize	Tamaño de la familia
6	Pstatus	Estado civil de los padres
7	Medu	Nivel educativo de la madre
8	Fedu	Nivel educativo del padre
9	Mjob	Trabajo de la madre
10	Fjob	Trabajo del padre
11	reason	Razón para elegir la escuela
12	guardian	Tutor del estudiante
13	traveltime	Tiempo de viaje a la escuela
14	studytime	Tiempo de estudio semanal
15	failures	Número de suspensos
16	schoolsup	Apoyo educativo extra
17	famsup	Apoyo educativo familiar
18	paid	Clases extra pagadas
19	activities	Actividades extracurriculares
20	nursery	Asistió a preescolar
21	higher	Desea educación superior
22	internet	Acceso a Internet en casa
23	romantic	Tiene relación romántica
24	famrel	Calidad de relaciones familiares
25	freetime	Tiempo libre después de la escuela
26	goout	Salir con amigos
27	Dalc	Consumo de alcohol entre semana
28	Walc	Consumo de alcohol en fin de semana
29	health	Estado de salud actual
30	absences	Número de ausencias escolares

(a) Características del conjunto de datos.



(b) Matriz de correlación del *Student Performance Dataset*.

Figura 4.2: Análisis exploratorio del conjunto de datos.

4.2.1. Preprocesamiento de Datos

El preprocesamiento de los datos se estructuró en las siguientes etapas, aplicadas uniformemente a todos los modelos para garantizar una evaluación justa y consistente:

- *Transformación de la variable objetivo:* Para abordar el problema como una tarea de clasificación binaria, la variable *G3* fue recodificada como binaria: calificaciones mayores o iguales a 10 se etiquetaron como *Aprobado* (1), y las menores como *Reprobado* (0).
- *Codificación de variables categóricas:* Las variables categóricas fueron transformadas mediante *one-hot encoding* para generar indicadores binarios compatibles con los algoritmos de aprendizaje automático.
- *Manejo de valores faltantes:* Los valores faltantes fueron imputados utilizando la mediana de cada característica seleccionada para preservar la integridad de los datos y evitando la pérdida de información relevante.
- *Selección de características:* Se identificaron seis características clave: *absences*, *goout*, *studytime*, *reason_reputation*, *failures* y *Fedu*. Estas características fueron seleccionadas mediante los modelos *DT* iniciales y la matriz de correlación.
- *Verificación de columnas booleanas:* Las columnas con valores booleanos fueron verificadas y convertidas explícitamente a enteros para asegurar consistencia en su representación.
- *División del conjunto de datos:* El conjunto de datos se dividió en entrenamiento (70 %) y prueba (30 %) preservando la proporción original de las clases para garantizar una representación fiel en ambos subconjuntos.
- *Balanceo de clases:* Para corregir el desbalance en las clases, se aplicó *SMOTE* al conjunto de entrenamiento para generar instancias sintéticas de la clase minoritaria. Posteriormente, las columnas booleanas afectadas fueron reconvertidas a valores binarios consistentes.

Como resultado, se obtuvo un conjunto de datos balanceado y preprocesado, con 384 instancias en ambos casos: utilizando 39 características con todas las variables disponibles y 6 características tras la selección de las más relevantes.

4.3. Desarrollo de los Modelos

Se entrenaron los siguientes modelos utilizando el mismo conjunto de datos preprocesado. En los modelos iniciales se utilizaron todas las características, mientras que en los modelos finales, se utilizaron las seis características clave seleccionadas.

- *DT-Scikit-learn:* Un modelo básico que permitió identificar la importancia relativa de las características y extraer reglas interpretables desde los nodos del árbol.
- *DT-InterpretML:* Enfocado en interpretabilidad, este modelo generó explicaciones globales y locales sobre el impacto de las características en las predicciones.
- *IDS:* Modelo basado en reglas interpretables construido con una librería personalizada desarrollada para este estudio, lo que limitó su inclusión en los modelos iniciales debido a su diseño específico para las características seleccionadas.

Estos modelos permitieron evaluar la influencia de las características seleccionadas y comparar tanto el rendimiento como la interpretabilidad. A continuación, se detallan los resultados de esta comparación y el proceso de desarrollo de cada modelo.

4.3.1. Modelo DT con Scikit-learn

El Árbol de Decisión implementado con *Scikit-learn* se evaluó en dos configuraciones principales: un modelo inicial que utilizó todas las características disponibles y otro que empleó únicamente las seis características seleccionadas. Ambos modelos compartieron los mismos hiperparámetros, asegurando condiciones comparables para la evaluación.

En el modelo inicial, las características más relevantes en orden de importancia fueron: *absences*, *failures*, *Fedu*, *Walc*, *goout*, *Fjob_other*, *reason_reputation*, *free_time*, *studytime*, *Mjob_services*, *Fjob_services* y *Medu*. Aunque estas características dominan las decisiones del modelo, muchas otras características tuvieron una influencia marginal o nula, aumentando la complejidad sin aportar beneficios significativos.

Hiperparámetro	**Valor**
criterion	gini
max_depth	4
min_samples_leaf	5
min_samples_split	2
ccp_alpha	0.0

Cuadro 4.1: Hiperparámetros clave utilizados en el modelo *DT-Scikit-learn*.

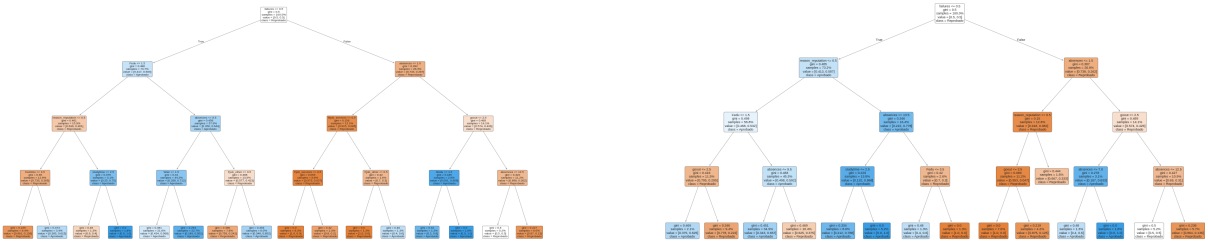
El uso de los mismos hiperparámetros en ambos modelos garantiza una comparación justa, enfocada en evaluar el impacto de las características seleccionadas sobre la estructura y el rendimiento del árbol. Los hiperparámetros se eligieron siguiendo valores estándar recomendados en la literatura para mantener un equilibrio entre simplicidad y rendimiento.

Al emplear las seis características seleccionadas, el árbol generado fue más compacto, mejorando la interpretabilidad sin comprometer el rendimiento predictivo, como se detalla en la Tabla 4.2.

	Modelo Inicial			**Modelo Final**			
Clase	**Precisión**	**Recall**	**F1-score**	**Precisión**	**Recall**	**F1-score**	**Soporte**
Reprobado	0.58	0.41	0.48	0.58	0.48	0.52	46
Aprobado	0.69	0.81	0.74	0.70	0.78	0.74	73
Accuracy		0.66			0.66		119
Macro avg	0.63	0.61	0.61	0.64	0.63	0.63	119
Weighted avg	0.64	0.66	0.64	0.66	0.66	0.66	119

Cuadro 4.2: Desempeño entre los modelos DT-Scikit-learn inicial y final.

En la Figura 4.3, se presentan los árboles generados por ambos modelos:

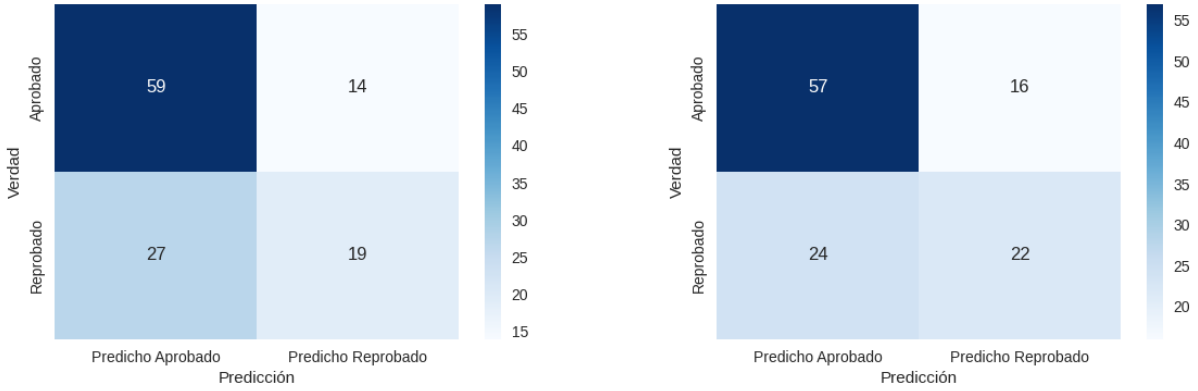


(a) Modelo inicial (b) Modelo final

Figura 4.3: Comparación de árboles de decisión.

- *(a) Modelo inicial:* El árbol presenta una mayor complejidad, con 16 reglas y una longitud promedio de 4.00, lo que dificulta su interpretación.
- *(b) Modelo final:* El árbol es más compacto, con 15 reglas y una longitud promedio de 3.93, lo que mejora ligeramente su interpretabilidad y reduce la carga cognitiva para los usuarios.

En la Figura 4.4, se compara el rendimiento a través de matrices de confusión:



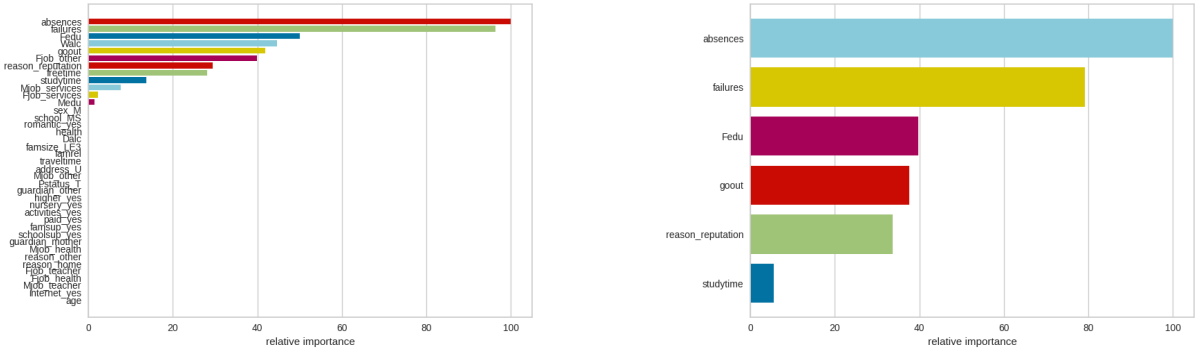
(a) Modelo inicial (b) Modelo final

Figura 4.4: Comparación de matrices de confusión.

- *(a) Modelo inicial:* Registra 14 falsos negativos en la clase *Aprobado* y 27 falsos positivos. Esto refleja una sensibilidad moderada y un nivel considerable de errores en la clasificación de estudiantes reprobados.
- *(b) Modelo final:* Incrementa los falsos negativos a 16, pero reduce los falsos positivos a 24. Esto indica una mejor precisión para identificar a los estudiantes reprobados a costa de una ligera disminución en la sensibilidad de la clase *Aprobado*.

El modelo final mejora la precisión para clasificar estudiantes reprobados, aunque compromete ligeramente la capacidad para identificar correctamente a los estudiantes aprobados.

Finalmente, la Figura 4.5 muestra la importancia relativa de las características:



(a) Modelo inicial (b) Modelo final

Figura 4.5: Comparación de importancia de características.

- *(a) Modelo inicial:* De las 39 características, las 12 más relevantes incluyen *absences*, *failures* y *Fedu*, pero la cantidad de variables menos influyentes añade complejidad innecesaria.
- *(b) Modelo final:* Conserva *absences* y *failures* como las más relevantes, lo que reduce la complejidad sin comprometer el rendimiento.

El modelo de DT implementado con *Scikit-learn* generó un total de 15 reglas. La Figura 4.6 muestra la precisión individual de cada regla. Se observa que las reglas más específicas, es decir, aquellas con múltiples condiciones combinadas, suelen tener una mayor precisión. En contraste, las reglas más generales, con menos condiciones, tienden a ser menos precisas debido a su mayor nivel de generalización.

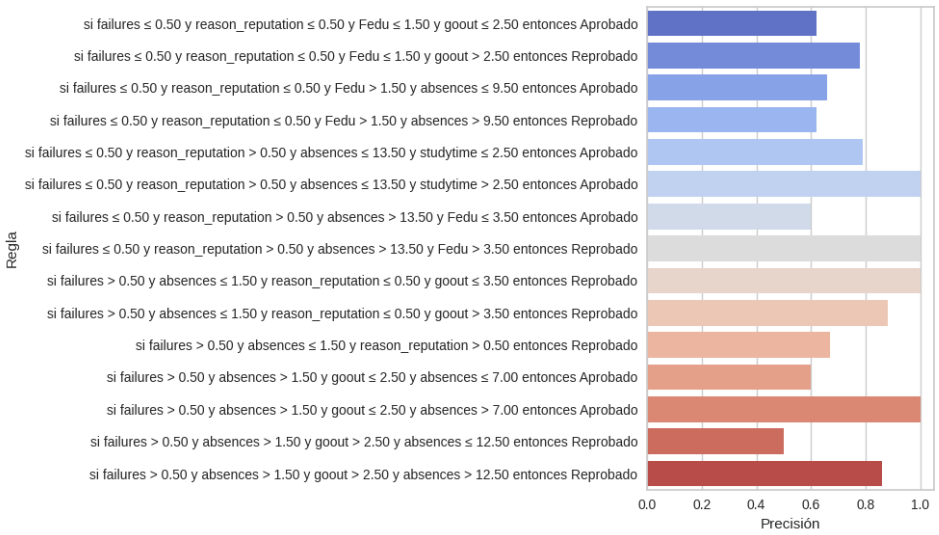


Figura 4.6: Precisión de reglas en el modelo DT-Scikit-learn.

4.3.2. Modelo DT con InterpretML

Al igual que el modelo *DT-Scikit-learn*, el modelo DT desarrollado con *InterpretML* incluyó un modelo inicial que utilizó todas las características y un modelo final entrenado únicamente con las seis características seleccionadas. Durante el análisis inicial con *InterpretML*, se identificó un problema en las visualizaciones de los árbo-

les, donde los valores del número de observaciones (#Obs) por nodo eran incorrectos debido a que la herramienta genera probabilidades en lugar de proporciones reales. Para solucionarlo, se accedió al modelo subyacente de *Scikit-learn* mediante la función *.model()*, lo que permitió generar grafos más precisos del árbol.

Hiperparámetro	**Valor**
criterion	gini
max_depth	3
min_samples_leaf	5
min_samples_split	2
ccp_alpha	0.0

Cuadro 4.3: Hiperparámetros clave del modelo *DT-InterpretML*.

En *DT-InterpretML*, los hiperparámetros mostrados en la Tabla 4.3, se seleccionaron para balancear interpretabilidad y rendimiento predictivo. Limitar la profundidad del árbol (*max_depth*) mejora la comprensibilidad de las reglas generadas, mientras que establecer un número mínimo de muestras por hoja (*min_samples_leaf*) contribuye a reducir el riesgo de sobreajuste.

	Modelo Inicial			**Modelo Final**			**Soporte**
Clase	**Precisión**	**Recall**	**F1-score**	**Precisión**	**Recall**	**F1-score**	
Reprobado	0.54	0.57	0.55	0.58	0.46	0.51	46
Aprobado	0.72	0.70	0.71	0.70	0.79	0.74	73
Accuracy		0.65			0.66		119
Macro avg	0.63	0.63	0.63	0.64	0.63	0.63	119
Weighted avg	0.65	0.65	0.65	0.65	0.66	0.65	119

Cuadro 4.4: Desempeño entre los modelos DT-InterpretML inicial y el modelo final.

La Tabla 4.4 muestra que el modelo final, al reducir la complejidad del árbol, mantuvo un desempeño predictivo consistente. En particular, el *recall* de la clase *Aprobado* mejoró, lo que indica una mayor sensibilidad al identificar observaciones positivas.

En las Figuras 4.7 y 4.8, se presentan los árboles generados por ambos modelos:

Figura 4.7: Árbol de decisión del modelo inicial generado con *InterpretML*.

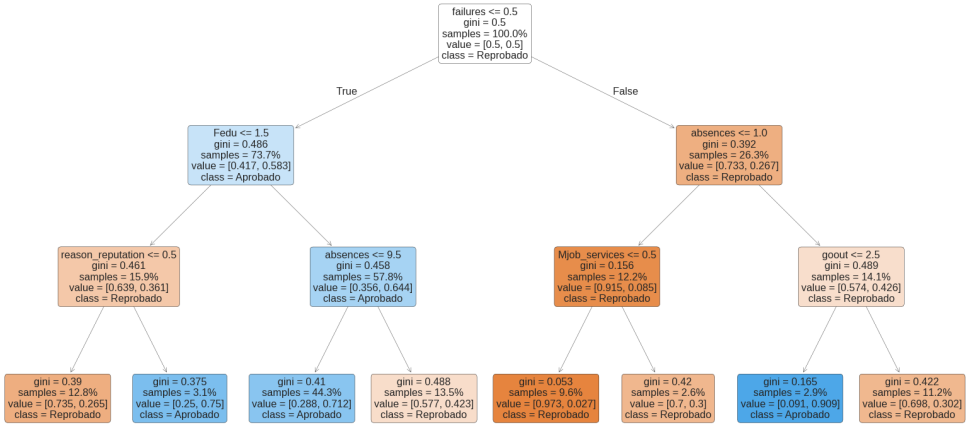
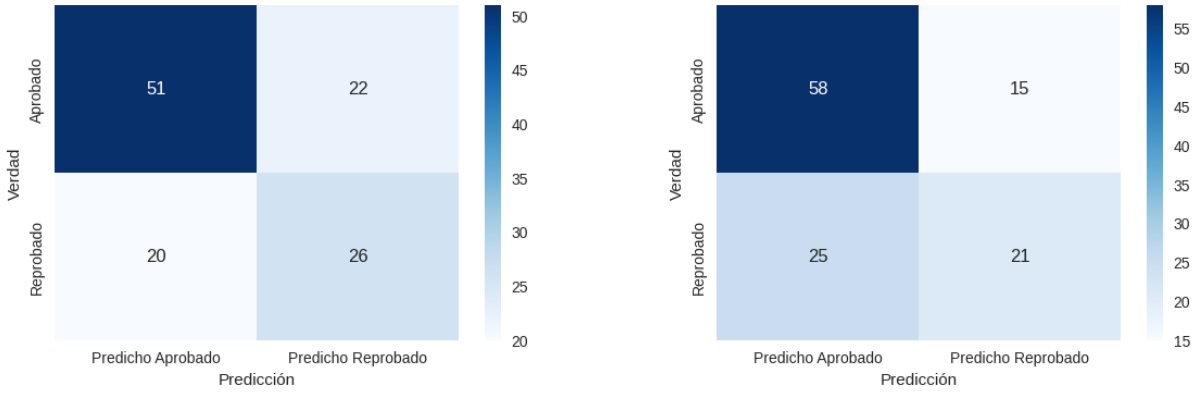


Figura 4.8: Árbol de decisión subyacente del modelo final generado con *Scikit-learn*.

- *(a) Modelo inicial:* El árbol incluye características menos relevantes, lo que aumenta la longitud promedio de las reglas y dificulta su interpretación.
- *(b) Modelo final:* El árbol es más compacto, con reglas más cortas y una estructura simplificada, lo que mejora su interpretabilidad.

El grafo subyacente del modelo, extraído de *scikit-learn*, ofrece información más precisa sobre las métricas de cada nodo, como el número real de observaciones y las proporciones exactas de clases. Esto facilita una interpretación más detallada y confiable del árbol. En contraste, el grafo nativo de *InterpretML* utiliza aproximaciones basadas en probabilidades que pueden generar confusión en el análisis de las decisiones del modelo.

En la Figura 4.9, se presentan las matrices de confusión de ambos modelos:



(a) Modelo inicial.

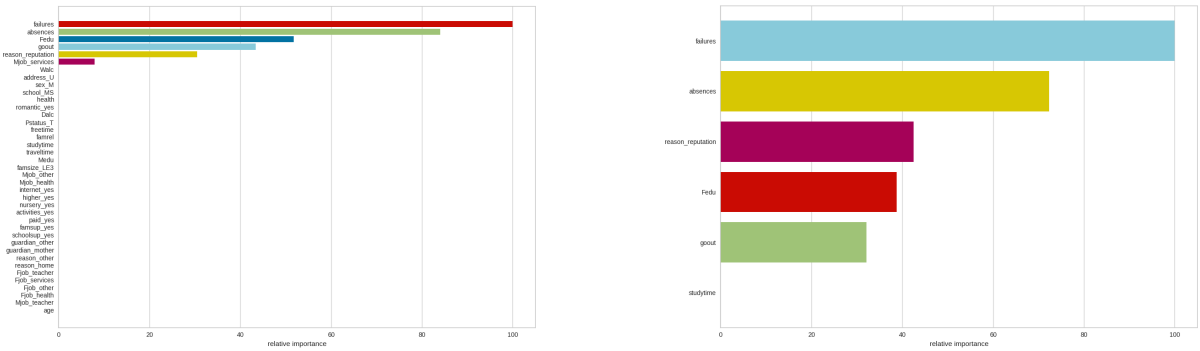
(b) Modelo final.

Figura 4.9: Comparación de matrices de confusión.

- *(a) Modelo inicial:* Presenta 22 falsos negativos en la clase *Aprobado*, indicando una menor sensibilidad. Además, refleja un nivel de error significativo al clasificar a 20 estudiantes reprobados como aprobados.
- *(b) Modelo final:* Reduce los falsos negativos a 15, mejorando la sensibilidad en la clase *Aprobado*, pero aumenta los falsos positivos a 25, lo que afecta el desempeño en la clase *Reprobado*.

El modelo final mejora la sensibilidad para identificar estudiantes aprobados, aunque incrementa los errores en la clase *Reprobado*, reflejando un cambio en el balance entre ambas clases.

Finalmente, la Figura 4.10 ilustra la importancia de las características en ambos modelos:



(a) Modelo inicial.

(b) Modelo final.

Figura 4.10: Comparación de la importancia de características en ambos modelos.

- *Modelo inicial:* Características como *failures*, *absences*, *Fedu* y *reason_reputation* dominaron las decisiones del modelo. Sin embargo, la inclusión de variables menos relevantes incrementó la complejidad estructural, dificultando la interpretación y análisis.
- *Modelo final:* Prioriza características clave como *failures* y *absences*, reduciendo la complejidad del árbol sin comprometer el rendimiento predictivo. Esto se traduce en una mayor interpretabilidad y reglas más manejables.

El modelo DT implementado con *InterpretML* generó un conjunto de reglas más compacto que el de *Scikit-learn*, con 12 reglas en total. La Figura 4.11 muestra que las reglas con condiciones intermedias mantienen un nivel de precisión estable. Esto sugiere que el modelo logra un equilibrio adecuado entre generalización y especificidad.

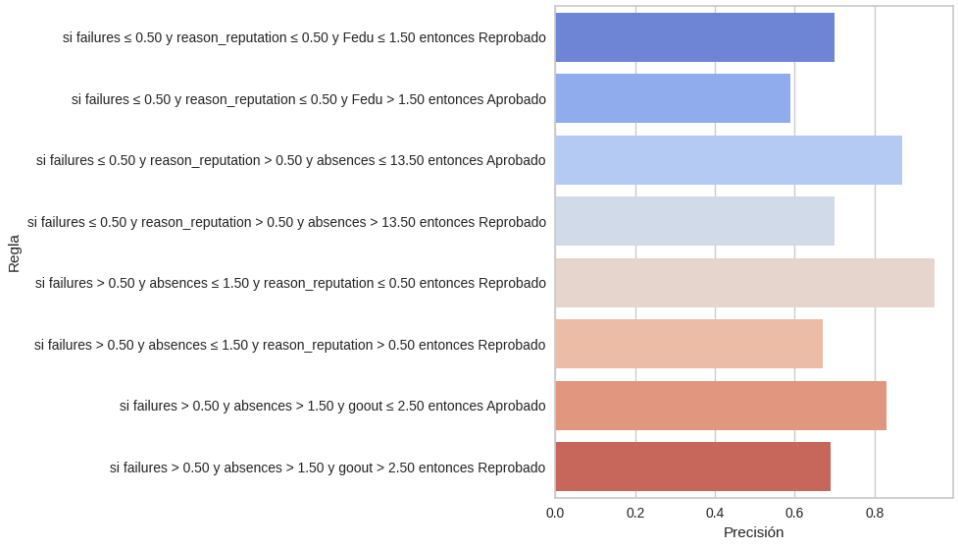


Figura 4.11: Precisión de reglas en el modelo DT-InterpretML.

4.3.3. Modelo IDS

El modelo IDS, basado en el trabajo de Lakkaraju et al. [1], busca equilibrar precisión e interpretabilidad mediante un conjunto de reglas disjuntas que asignan clases a las instancias de forma transparente. Dado que no existen implementaciones estándar en paquetes disponibles, se desarrolló una librería personalizada en Python, fundamentada en los repositorios *pyIDS* [43] e *Interpretable Decision Sets* [44]. Este enfoque fue adaptado específicamente a las necesidades del presente estudio.

En este trabajo, se utilizó programación lineal entera mixta (MILP) para construir el conjunto de reglas. A diferencia del algoritmo *Smooth Local Search* (SLS) propuesto en [1], MILP ofrece soluciones óptimas y un control matemático más preciso, lo cual es ideal para el tamaño moderado del dataset utilizado en este estudio. La formulación de MILP retoma las métricas de interpretabilidad descritas en los fundamentos teóricos, incluyendo el tamaño del conjunto ($|R|$, Ecuación 2.5), la longitud promedio de reglas ($\text{len}(r)$, Ecuación 2.6) y el solapamiento entre reglas ($\text{Overlap}(R_i, R_j)$, Ecuación 2.7). Estas métricas, junto con las restricciones que definen la cobertura mínima, la longitud máxima de las reglas y la penalización por solapamiento, se implementaron utilizando la herramienta *PuLP* [45], una librería de Python para la resolución de problemas de optimización lineal y entera mixta. Las formulaciones detalladas se encuentran en las Ecuaciones 2.9-2.11.

El modelo selecciona un subconjunto de reglas $R = \{r_1, r_2, \ldots, r_n\}$, donde cada regla r_i toma la forma:

$$r_i : \{\text{condición}_{i1}, \text{condición}_{i2}, \ldots, \text{condición}_{ik}\} \rightarrow c_i,$$

y $c_i \in \{0, 1\}$ corresponde a la clase predicha (0: Reprobado, 1: Aprobado).

La implementación incluyó:

- *Generación de reglas candidatas:* Identificación de combinaciones frecuentes de características en los datos.
- *Optimización mediante MILP:* Selección de reglas óptimas utilizando la función objetivo descrita en la Ecuación 2.10.
- *Visualización:* Generación de grafos globales y locales que resumen las reglas seleccionadas.

El uso de MILP permitió adaptar y extender el enfoque de [1], priorizando la construcción de un modelo transparente ajustado a las necesidades del presente trabajo, con potencial para maximizar la interpretabilidad.

4.3.3.1. Implementación del Modelo

La configuración de los hiperparámetros del modelo IDS fue seleccionada mediante análisis experimental, considerando las características del dataset y los objetivos del modelo. Los valores ajustados logran un balance entre interpretabilidad, simplicidad y precisión, como se detalla a continuación:

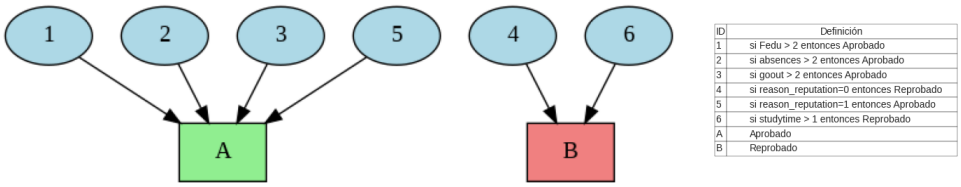
- λ_1 (*Penalización por tamaño*): Controla el número total de reglas en el modelo. Un valor bajo (0.01) asegura suficiente flexibilidad para capturar la complejidad del problema sin redundancia.

- λ_2 (*Penalización por longitud*): Penaliza reglas largas, favoreciendo reglas más simples y cognitivamente manejables (0.01).
- λ_3 (*Penalización por errores*): Priorizando precisión, este valor (1.0) penaliza errores de clasificación, asegurando reglas efectivas.
- λ_4 (*Penalización por solapamiento*): Reduce conflictos entre reglas asignando un valor de 1.0, fomentando disyunción entre ellas.
- *min_support* (*Soporte mínimo*): Un valor de 0.10 limita el análisis a patrones relevantes en el dataset.
- *min_confidence* (*Confianza mínima*): Un umbral de 0.50 garantiza que las reglas seleccionadas sean al menos un 50 % precisas.
- *max_rule_length* (*Longitud máxima de reglas*): Restringida a 3 condiciones por regla, asegura reglas simples y fáciles de interpretar.

Cuadro 4.5: Hiperparámetros utilizados para el modelo IDS.

Hiperparámetro	**Valor**
λ_1	0.01
λ_2	0.01
λ_3	1.0
λ_4	1.0
min_support	0.10
min_confidence	0.50
max_rule_length	3

El grafo global generado por el modelo IDS, mostrado en la Figura 4.12, representa la relación entre las reglas seleccionadas y las predicciones realizadas. Los nodos azules corresponden a las reglas, mientras que los nodos verdes y rojos representan las clases *Aprobado* y *Reprobado*, respectivamente.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si goout > 2 entonces Aprobado
4	si reason_reputation=0 entonces Reprobado
5	si reason_reputation=1 entonces Aprobado
6	si studytime > 1 entonces Reprobado
A	Aprobado
B	Reprobado

Figura 4.12: Grafo de reglas del modelo IDS con tabla de nodos.

El conjunto de reglas seleccionado por el modelo IDS refleja las características esperadas tras la configuración de los hiperparámetros y el análisis experimental. A continuación, se resumen las propiedades clave derivadas del modelo entrenado:

- *Tamaño del modelo (size)*: 6 reglas seleccionadas.
- *Longitud promedio (avg_length)*: 1 condición por regla.
- *Cobertura total (cover)*: Número de instancias del conjunto de entrenamiento balanceado (X_train_smote) cubiertas por al menos una regla (384 de 384).

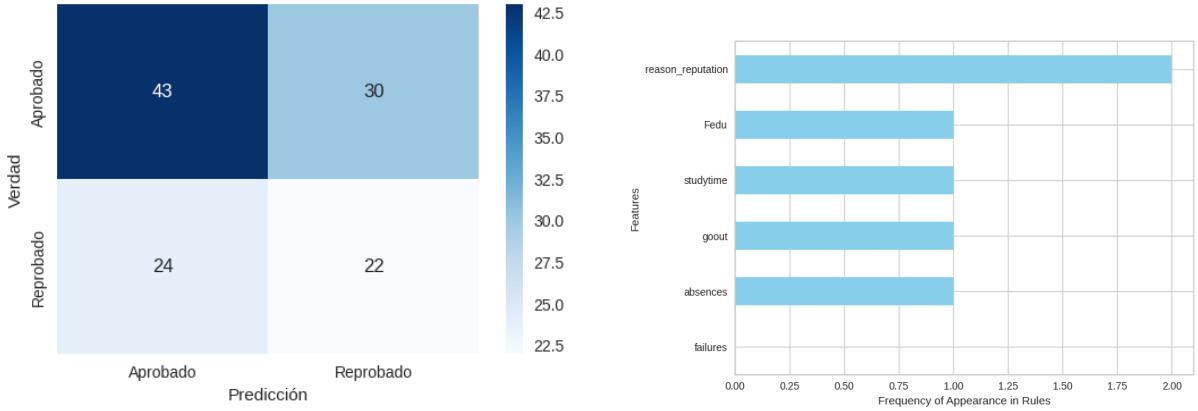
- *Solapamiento (overlap)*: Número de instancias cubiertas por más de una regla (373 de 384).

La Tabla 4.6 resume las métricas de rendimiento del modelo IDS. Con una precisión global (*Accuracy*) del 55 %, el modelo muestra un mejor desempeño clasificando a estudiantes aprobados, alcanzando un *Recall* del 59 %. Sin embargo, su capacidad para identificar estudiantes reprobados es menor, con un *Recall* del 48 %. Estos resultados indican que, si bien el modelo es más efectivo para la clase *Aprobado*, podría beneficiarse de mejoras en la detección de la clase *Reprobado*.

Clase	**Precisión**	**Recall**	**F1-Score**	**Soporte**
Reprobado	0.42	0.48	0.45	46
Aprobado	0.64	0.59	0.61	73
Accuracy			0.55	
Macro Avg	0.53	0.53	0.53	119
Weighted Avg	0.56	0.55	0.55	119

Cuadro 4.6: Métricas de rendimiento del modelo IDS.

La Figura 4.13 incluye (a) la matriz de confusión y (b) la importancia de características, que complementan el análisis del modelo IDS:



(a) Matriz de confusión.

(b) Importancia de características.

Figura 4.13: Desempeño del modelo IDS: (a) matriz de confusión que ilustra la precisión y errores de clasificación, y (b) importancia de las características utilizadas para generar las reglas del modelo.

En (a), se observa que el modelo clasifica correctamente 43 estudiantes *Aprobados* y 22 *Reprobados*, pero también presenta 30 falsos positivos y 24 falsos negativos, lo que indica un sesgo hacia la clase *Aprobado*. En (b), se evidencia que *reason_reputation* es la característica más relevante, apareciendo en dos reglas distintas, mientras que *Fedu*, *studytime*, *goout* y *absences*, también tienen un impacto significativo.

El modelo IDS se caracteriza por generar reglas más interpretables y compactas, con un total de 6 reglas. Como se observa en la Figura 4.14, las reglas principales presentan una precisión alta, lo que refuerza su objetivo de interpretabilidad

sin sacrificar significativamente el rendimiento. Esto es evidente en reglas como *si reason_reputation = 1 entonces Aprobado*, que destaca por su claridad y alta precisión.

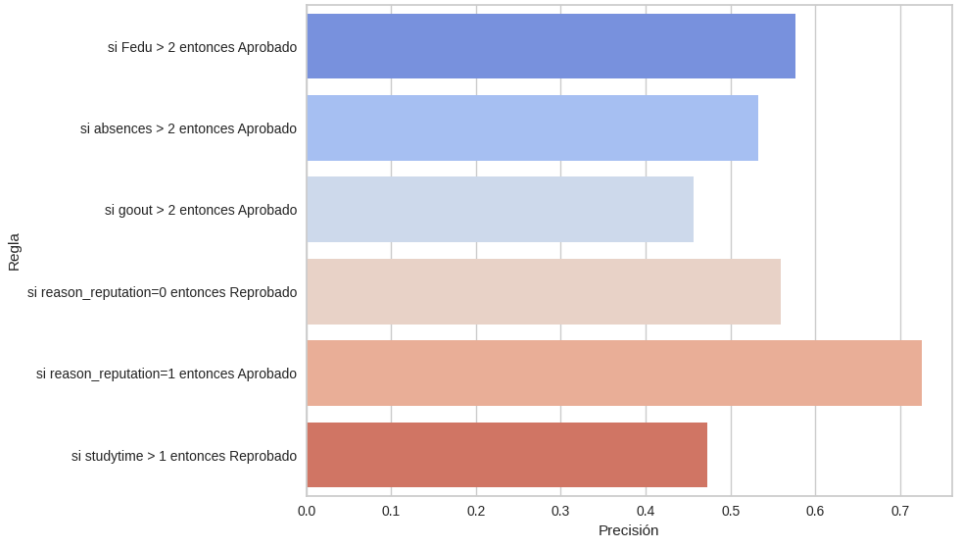


Figura 4.14: Precisión de reglas en el modelo IDS.

4.4. Evaluación Técnica de los Modelos

La interpretabilidad, como se define en la Sección 2.1, se refiere a la capacidad de un modelo para explicar de manera comprensible sus predicciones. En este trabajo, se emplea una métrica compuesta que combina dimensiones descritas en los fundamentos teóricos (Sección 2.3.3) para evaluar los modelos. Estas dimensiones son:

- **Parsimonia ($1/L$)**: Inversa de la longitud promedio de las reglas generadas (L), donde reglas más cortas favorecen la interpretabilidad.
- **Cobertura ($|C_R|/|D|$)**: Proporción de instancias explicadas por al menos una regla, reflejando la capacidad del modelo para generalizar.
- **Solapamiento ($|O_R|/|D|$)**: Proporción de instancias cubiertas por múltiples reglas, lo que puede dificultar la interpretación.
- **Precisión (P)**: Rendimiento predictivo del modelo en el conjunto de prueba, asegurando su utilidad práctica.

Estas dimensiones se combinan en la siguiente fórmula de interpretabilidad:

$$I = \alpha \cdot \mathrm{P} + \beta \cdot \frac{1}{L} + \gamma \cdot \frac{|C_R|}{|D|} - \delta \cdot \frac{|O_R|}{|D|}, \tag{4.1}$$

donde los coeficientes α, β, γ y δ ponderan la relevancia de cada dimensión. En este estudio, se asignaron los valores $\alpha = 0.3$, $\beta = 0.4$, $\gamma = 0.2$ y $\delta = 0.1$, priorizando la simplicidad y claridad interpretativa en línea con los objetivos del trabajo.

Pruebas comparativas con distintas configuraciones de pesos evaluaron el impacto en las puntuaciones de interpretabilidad para los modelos *DT-Scikit-learn*, *DT-*

InterpretML e *IDS*. Los resultados, resumidos en la Tabla 4.7, muestran que la configuración seleccionada (Configuración 1) ofrece un balance adecuado entre las dimensiones evaluadas.

Cuadro 4.7: Configuraciones de pesos y resultados de interpretabilidad.

Configuración	α	β	γ	δ	Scikit-learn	InterpretML	IDS
1	0.3	0.4	0.2	0.1	0.4802	0.4992	0.4667
2	0.4	0.3	0.2	0.1	0.5264	0.5405	0.4714
3	0.2	0.5	0.2	0.1	0.4341	0.4578	0.4621
4	0.3	0.3	0.3	0.1	0.5600	0.5742	0.5167
5	0.1	0.4	0.3	0.2	0.4475	0.4664	0.3604

Para analizar la interpretabilidad de los modelos IDS y DT-InterpretML, se diseñaron visualizaciones personalizadas que muestran sus estructuras globales y locales. Los detalles y resultados de estas evaluaciones se presentan en la Sección 5.6.

4.5. Diseño y Desarrollo del Cuestionario

Para evitar sobrecarga cognitiva en los participantes, se tomaron las siguientes decisiones para simplificar el diseño del cuestionario:

- *Preguntas de seguimiento:* Limitadas únicamente a las categorías *Ambigüedad* y *Error* para analizar el nivel de confianza de los usuarios en sus respuestas.
- *Exclusión del modelo DT-Sklearn:* Eliminado debido a su redundancia con *DT-InterpretML*, lo que permitió mantener el enfoque en herramientas más avanzadas de interpretabilidad.
- *Diversidad en los ejemplos:* Se seleccionaron observaciones representativas que incluyen valores extremos y combinaciones conflictivas, con el objetivo de explorar diferentes escenarios interpretativos.

4.5.1. Estructura final

El cuestionario final se compone de 21 preguntas principales, organizadas en las siguientes categorías:

- *Exactitud (6 preguntas):* Diseñadas para evaluar si los usuarios comprenden y verifican correctamente las predicciones de los modelos, basándose en reglas y visualizaciones.
- *Ambigüedad (6 preguntas):* Orientadas a analizar la claridad de las predicciones en casos donde las reglas presentan ambigüedad.
- *Error (6 preguntas):* Enfocadas en medir la capacidad de los usuarios para identificar predicciones incorrectas.
- *Preferencias de Visualización (2 preguntas):* Destinadas a comparar qué modelo facilita mejor la comprensión de las predicciones.

- *Pregunta Descriptiva (1 pregunta):* Busca recoger opiniones subjetivas sobre la utilidad de las reglas y grafos en el proceso de interpretación de predicciones.

4.5.2. Selección de Observaciones

El proceso de selección y clasificación de observaciones se realizó utilizando el conjunto de datos de prueba (*X_test*) y sus etiquetas reales (*y_test_labels*), junto con las predicciones generadas por los modelos *DT-InterpretML* e *IDS*.

Los pasos principales de este proceso fueron:

1. *Creación del DataFrame:* Se construyó un *DataFrame* con las siguientes columnas:
   - *id*: Identificador único de cada observación.
   - *pregunta*: Índice que vincula cada observación con una pregunta específica del cuestionario.
   - Características seleccionadas (*absences, goout, studytime, reason_reputation, failures, Fedu*): Representan las variables más relevantes.
   - *real*: Clase real de la observación (*Aprobado*/*Reprobado*).
   - *interpretml* y *ids*: Predicciones generadas por los modelos *DT-InterpretML* e *IDS*, respectivamente.

   La Tabla 4.8 presenta las observaciones seleccionadas para el cuestionario:

id	pregunta	absences	goout	studytime	reason_reputation	failures	Fedu	real	interpretml	ids	categoría
8	234	18	2	2	0	0	1	Reprobado	Reprobado	Reprobado	Exactitud
12	209	6	2	3	1	0	3	Reprobado	Aprobado	Aprobado	Error
26	113	10	2	1	0	0	2	Aprobado	Aprobado	Reprobado	Ambigüedad
42	145	0	2	2	0	0	1	Aprobado	Reprobado	Reprobado	Error
52	56	0	2	2	1	0	3	Aprobado	Aprobado	Aprobado	Exactitud
63	157	6	5	1	0	3	1	Aprobado	Reprobado	Aprobado	Ambigüedad
106	350	8	4	2	0	3	1	Reprobado	Reprobado	Reprobado	Exactitud
114	79	12	3	2	0	0	4	Reprobado	Aprobado	Aprobado	Error
115	281	19	4	1	0	1	2	Aprobado	Reprobado	Aprobado	Ambigüedad

Cuadro 4.8: Subconjunto de observaciones seleccionadas para el cuestionario.

2. *Clasificación en Categorías:* Cada observación fue clasificada en una de las siguientes categorías, en función de las predicciones realizadas por los modelos:
   - *Exactitud:* Ambas predicciones coinciden con la clase real:

$$y\_real = y\_interpretml = y\_ids$$

   - *Error:* Las predicciones coinciden entre sí, pero no con la clase real:

$$y\_real \neq y\_interpretml = y\_ids$$

   - *Ambigüedad:* Las predicciones de los dos modelos no coinciden:

$$y\_interpretml \neq y\_ids$$

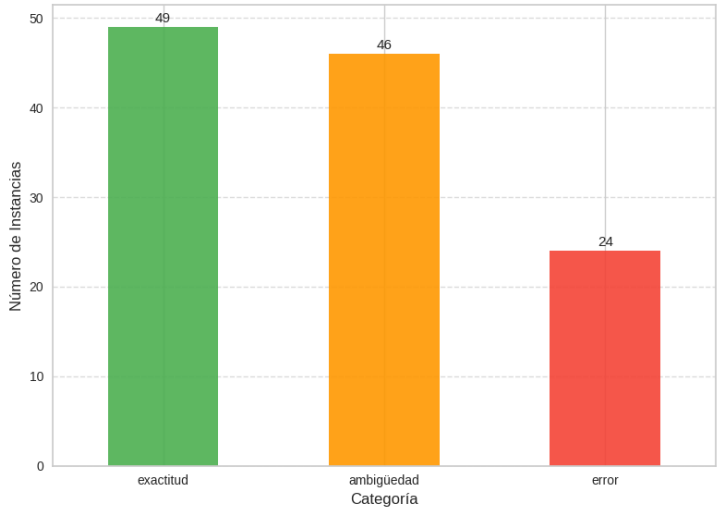


Figura 4.15: Distribución de instancias por categoría en conjunto de pruebas (*X_test*): exactitud, ambigüedad y error.

El análisis categórico revela que mientras *DT-InterpretML* y *IDS* coinciden en gran parte de las predicciones correctas, las discrepancias en la categoría de ambigüedad destacan áreas donde los modelos divergen y proporciona información relevante sobre su consistencia y capacidad de generalización.

4.5.3. Reporte Individual

Este reporte, generado para cada usuario, contiene los siguientes elementos:

- α (*user_id*): Identificador único del usuario.
- β (*question*): Número de la pregunta en el cuestionario.
- γ (*category*): Categoría principal de la pregunta.
- δ (*sub_category*): Subcategoría de la pregunta.
- ϵ (*absences*): Número de ausencias del estudiante.
- ζ (*goout*): Tiempo que pasa el estudiante socializando.
- η (*studytime*): Horas dedicadas al estudio semanal.
- θ (*reason_reputation*): Razón principal para elegir la escuela.
- ι (*failures*): Número de materias reprobadas.
- κ (*Fedu*): Nivel educativo del padre.
- λ (*real_prediction*): Clase real asignada a la observación.
- μ (*prediction_model_ids*): Predicción realizada por el modelo IDS.
- ν (*prediction_model_dt*): Predicción realizada por el modelo DT-InterpretML.
- ξ (*user_answer*): Respuesta seleccionada por el usuario.
- o (*follow_up_question*): Pregunta de seguimiento asociada.

- π (*follow_up_answer*): Respuesta a la pregunta de seguimiento.
- τ (*response_time_seconds*): Tiempo empleado en responder (en segundos).

α	β	γ	δ	ϵ	ζ	η	θ	ι	κ	λ	μ	ν	ξ	o	π	τ
1	1	1	1	18	2	2	0	0	1	0	0	0	1	0	0	51.8
1	1	1	1	18	2	2	0	0	1	0	0	0	0	0	0	12.8
1	3	1	2	0	2	2	1	0	3	1	1	1	1	0	0	7.3
1	4	1	2	0	2	2	1	0	3	1	1	1	0	0	0	12.4
1	5	1	3	8	4	2	0	3	1	0	0	0	1	0	0	12.4
1	6	1	3	8	4	2	0	3	1	0	0	0	0	0	0	2.7
1	7	2	1	10	2	1	0	0	2	1	0	1	1	1	3	12.4
1	8	2	1	10	2	1	0	0	2	1	0	1	0	1	1	11.0
1	9	2	2	6	5	1	0	3	1	1	1	0	1	1	2	9.0
1	10	2	2	6	5	1	0	3	1	1	1	0	0	1	3	14.5
1	11	2	3	19	4	1	0	1	2	1	1	0	1	1	1	11.3
1	12	2	3	19	4	1	0	1	2	1	1	0	0	1	2	9.0
1	13	3	1	6	2	3	1	0	3	0	1	1	2	1	3	10.3
1	14	3	1	6	2	3	1	0	3	0	1	1	3	1	1	9.5
1	15	3	2	0	2	2	0	0	1	1	0	0	2	1	2	9.8
1	16	3	2	0	2	2	0	0	1	1	0	0	3	1	3	11.4
1	17	3	3	12	3	2	0	0	4	0	1	1	2	1	1	69.3
1	18	3	3	12	3	2	0	0	4	0	1	1	3	1	2	72.9
1	19	4	-	-	-	-	-	-	-	-	-	-	5	0	0	4.0
1	20	4	-	-	-	-	-	-	-	-	-	-	4	0	0	2.9
1	21	5	-	-	-	-	-	-	-	-	-	-	6	0	0	6.8

Cuadro 4.9: Diseño del reporte individual User_1 implementado en la herramienta web para generar el reporte de cada usuario.

4.5.4. Reporte Consolidado

Este reporte integra las respuestas de todos los usuarios. Sus elementos son:

- Q *(question):* Número de la pregunta en el cuestionario.
- A *(aprobado):* Número de veces que la respuesta fue *aprobado*.
- R *(reprobado):* Número de veces que la respuesta fue *reprobado*.
- A_m *(aprobado_mucho):* Número de respuestas con *aprobado-mucho*.
- A_p *(aprobado_poco):* Número de respuestas con *aprobado-poco*.
- A_n *(aprobado_nada):* Número de respuestas con *aprobado-nada*.
- R_m *(reprobado_mucho):* Número de respuestas con *reprobado-mucho*.
- R_p *(reprobado_poco):* Número de respuestas con *reprobado-poco*.
- R_n *(reprobado_nada):* Número de respuestas con *reprobado-nada*.
- C_m *(correcto_mucho):* Número de respuestas con *correcto-mucho*.
- C_p *(correcto_poco):* Número de respuestas con *correcto-poco*.

- C_n *(correcto_nada):* Número de respuestas con *correcto-nada*.
- I_m *(incorrecto_mucho):* Número de respuestas con *incorrecto-mucho*.
- I_p *(incorrecto_poco):* Número de respuestas con *incorrecto-poco*.
- I_n *(incorrecto_nada):* Número de respuestas con *incorrecto-nada*.
- C *(correcto):* Número total de respuestas con correcto.
- I *(incorrecto):* Número total de respuestas con con incorrecto.
- IDS *(ids):* Número de veces que el modelo IDS fue seleccionado.
- DT *(dt):* Número de veces que el modelo DT-InterpretML fue seleccionado.
- T *(avg_time_(s)):* Tiempo promedio (en segundos) que los usuarios tardaron en responder la pregunta.

Q	A	R	A_m	A_p	A_n	R_m	R_p	R_n	C_m	C_p	C_n	I_m	I_p	I_n	C	I	IDS	DT	T
1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21.1
2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.7
3	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.1
4	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.6
5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5.7
6	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.5
7	1	2	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	8.9
8	1	2	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	13.8
9	2	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	12.4
10	2	1	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	12.6
11	1	2	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	9.9
12	1	2	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	7.8
13	0	0	0	0	0	0	0	0	1	1	0	0	0	1	2	1	0	0	10.0
14	0	0	0	0	0	0	0	0	0	2	0	0	0	1	2	1	0	0	7.3
15	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	2	0	0	10.9
16	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	2	0	0	8.3
17	0	0	0	0	0	0	0	0	1	0	1	0	1	0	2	1	0	0	29.7
18	0	0	0	0	0	0	0	0	2	0	0	0	1	0	2	1	0	0	29.6
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	7.7
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	3.0

Cuadro 4.10: Diseño del reporte consolidado del cuestionario implementado en la herramienta web. Disponible en la hoja de Excel *report_generated*.

4.5.5. Glosario

A continuación, se presenta un ejemplo de la estructura del glosario (*glossary*) generado junto con los reportes:

- *user_id*: Identificador único asignado automáticamente por la herramienta web a cada usuario que completa el cuestionario. Este identificador permite rastrear y analizar las respuestas mientras se mantiene el anonimato. Para los usuarios de prueba, los valores registrados son:
  - *1*: *243afd66-9eae-4396-a355-fc77886fc28c*
  - *2*: *5e76796d-8605-4cc8-9dce-ce25a5fe81ef*

- *3*: *e90a4052-4e16-4d85-ad1b-5ef227c35538*

- *question*: Identifica las preguntas del cuestionario organizadas según su propósito y enfoque. En cada par de preguntas del rango 1 a 18, se evalúa la misma observación: la primera pregunta aplicada al modelo *DT-InterpretML*, y la segunda al modelo *IDS*. A continuación, se presentan los índices de las preguntas del cuestionario:
  - *1–6* (preguntas de exactitud): Se evalúa si el usuario realiza la misma predicción que los modelos: *Basándote en las reglas proporcionadas, selecciona si la predicción correcta para esta observación es 'Aprobado' o 'Reprobado'*
    - *1–2*: Se proporcionan sólo las reglas aplicables.
    - *3–4*: Preguntas basadas en reglas y grafos globales.
    - *5–6*: Preguntas basadas en reglas y grafos locales.
  - *7–12* (preguntas de ambigüedad): Se evalúa la predicción del usuario en casos en los que los modelos se contradicen: *Basándote en las reglas proporcionadas, selecciona si la predicción correcta para esta observación es 'Aprobado' o 'Reprobado'. Si al analizar las reglas no encuentras una respuesta evidente, considera que esto puede reflejar ambigüedad en las reglas.*
    - *7–8*: Se proporcionan sólo las reglas aplicables.
    - *9–10*: Preguntas basadas en reglas y grafos globales.
    - *11–12*: Preguntas basadas en reglas y grafos locales.
  - *13–18* (preguntas de error): Se evalúa si el usuario es capaz de detectar cuando ambos modelos se equivocan: *Basándote en la regla resaltada y los valores de la observación, califica si la predicción del modelo es adecuada según las reglas proporcionadas. Si encuentras inconsistencias entre la predicción y las reglas, considera esto al tomar tu decisión.*
    - *13–14*: Se proporcionan sólo las reglas aplicables.
    - *15–16*: Preguntas basadas en reglas y grafos globales.
    - *17–18*: Preguntas basadas en reglas y grafos locales.
  - *19*: Pregunta comparativa sobre la utilidad de los grafos: *¿Cuál de los siguientes grafos encontraste más útil para entender el funcionamiento del modelo?*
  - *20*: Pregunta comparativa sobre cuál modelo facilita más la comprensión: *¿Qué modelo te facilitó comprender la predicción y analizar posibles errores?*
  - *21*: Pregunta abierta sobre la importancia de visualizar grafos junto con predicciones: *¿Crees que la visualización del grafo y sus reglas debería siempre acompañar las predicciones para mejorar la comprensión?*
- *user_answer*: Representa la respuesta seleccionada por los participantes según el rango de preguntas:
  - *0*: *Reprobado* (preguntas 1–12).
  - *1*: *Aprobado* (preguntas 1–12).

  - *2*: *Correcto* (preguntas 13–18).
  - *3*: *Incorrecto* (preguntas 13–18).
  - *4*: *DT-InterpretML* (preguntas 19–20).
  - *5*: *IDS* (preguntas 19–20).
  - *6*: Respuesta en texto libre (pregunta 21).
- *category*: Clasifica cada pregunta en una de las siguientes categorías:
  - *1*: Exactitud (ambos modelos predicen correctamente).
  - *2*: Ambigüedad (ambos modelos se contradicen).
  - *3*: Error (ambos modelos predicen incorrectamente).
  - *4*: Preferencias de Visualización (entre ambos modelos).
  - *5*: Pregunta Descriptiva (opiniones subjetivas sobre las visualizaciones).
- *sub_category*: Especifica el nivel de detalle de las preguntas dentro de cada categoría:
  - *1*: Basadas en reglas individuales.
  - *2*: Basadas en grafos globales.
  - *3*: Basadas en grafos locales.
- *class*: Indica la clase real de cada observación:
  - *1*: Aprobado.
  - *0*: Reprobado.
- *follow_up_question*: Preguntas de seguimiento incluidas en las categorías de *Ambigüedad* y *Error* para medir el nivel de confianza de los participantes:
  - *1*: *¿Qué tan seguro(a) estás de tu respuesta?*
- *follow_up_answer*: Representa el nivel de confianza expresado por los participantes en las preguntas de seguimiento:
  - *1*: Nada.
  - *2*: Poco.
  - *3*: Mucho.
- *user_id_<id_de_usuario>*: Identificador único para cada usuario. En usuarios de prueba, incluye las respuestas de texto libre para verificar el funcionamiento del sistema:
  - *user_id_1*: Respuesta: *test_1*.
  - *user_id_2*: Respuesta: *test_2*.
  - *user_id_3*: Respuesta: *test_3*.

4.6. Desarrollo de la Herramienta de Interpretabilidad

La herramienta web desarrollada para el cuestionario, denominada *Survey-XAI-App* fue diseñada para evaluar la interpretabilidad de modelos transparentes mediante la interacción directa con los participantes.

La aplicación está construida con HTML, CSS y JavaScript en el frontend, y Flask como framework en el backend. Se utilizó MongoDB como base de datos y Railway para el alojamiento de la aplicación. A continuación, se detallan sus características principales:

- *Página de introducción:* Proporciona una visión general del cuestionario, describiendo su propósito, los modelos evaluados y las condiciones de privacidad asociadas.

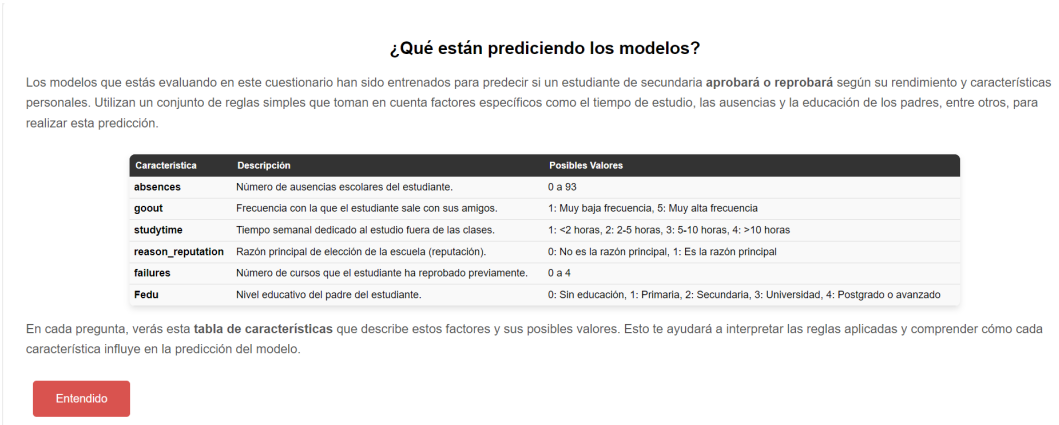


Figura 4.16: Introducción al cuestionario.

- *Explicación de los modelos:* Incluye descripciones sobre las reglas y estructuras globales y locales generadas por los modelos.

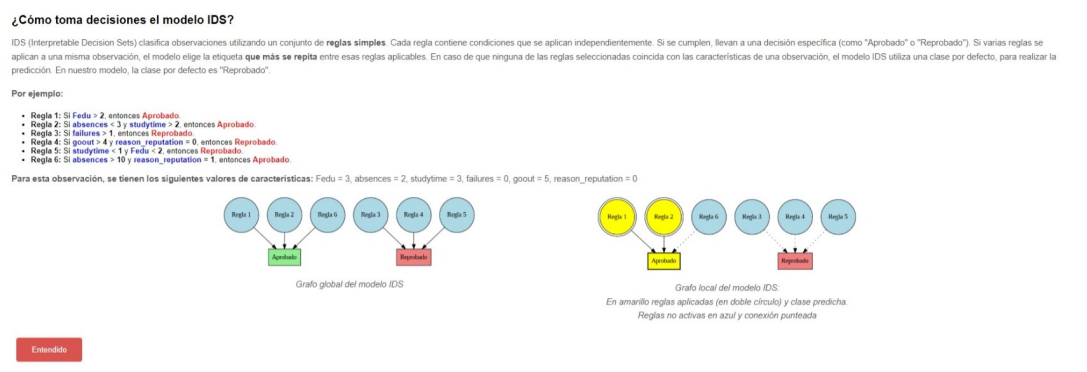


Figura 4.17: Ejemplo de explicación del modelo IDS.

- *Sección de preguntas:* Muestra observaciones específicas acompañadas de las reglas y grafos generados por los modelos, que facilitan la evaluación de aspectos como exactitud, ambigüedad y errores por parte de los participantes. Las

opciones de respuesta se presentan mediante botones rosados, que cambian a azul cuando son seleccionados.

Figura 4.18: Ejemplo de una pregunta.

- *Preguntas de seguimiento:* En este tipo de preguntas los usuarios seleccionan una respuesta y luego califican su nivel de confianza en la decisión mediante una respuesta de seguimiento (*Mucho*, *Poco* o *Nada*). Las respuestas de seguimiento utilizan botones azules que son resaltados en naranja cuando son seleccionados.

Figura 4.19: Ejemplo de una pregunta con respuesta principal y de seguimiento.

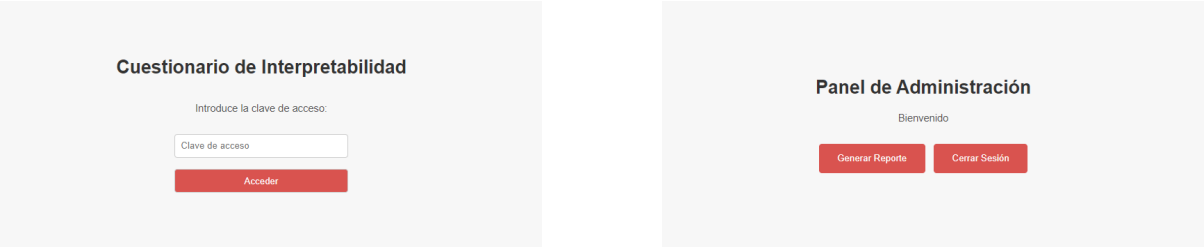
- *Registro de tiempos:* Captura automáticamente el tiempo empleado en responder cada pregunta, proporcionando información valiosa sobre la claridad de las explicaciones.
- *Generación de reportes:* Exporta los resultados en archivos Excel con análisis detallados, facilitando la interpretación y comparación de datos.

La herramienta ha sido probada y validada internamente, y se encuentra preparada para administrar el cuestionario en estudios futuros.

4.6.1. Gestión de Accesos

Survey-XAI-App implementa un sistema de autenticación para gestionar los accesos de dos tipos de usuarios:

- *Usuarios regulares*: Acceden a la funcionalidad de responder el cuestionario.
- *Administrador*: Tiene acceso al panel de administración para generar reportes y gestionar las respuestas.



(a) Interfaz de acceso. (b) Panel de administración.

Figura 4.20: Vistas clave: Panel de administración e interfaz de acceso.

El acceso a cualquiera de los roles requiere el uso de una contraseña, la cual se configura previamente sin necesidad de asignar un usuario individual a cada participante. Estas contraseñas se gestionan a través de variables de entorno.

- Contraseña del usuario: Definida en la variable *USER_ACCESS_KEY*.
- Contraseña del administrador: Definida en la variable *ADMIN_ACCESS_KEY*.

Durante el inicio de sesión, las contraseñas ingresadas por el usuario se comparan con estas claves, y el acceso se concede según el rol del usuario.

4.6.2. Cuestionario

Las preguntas están organizadas y almacenadas en un archivo JSON que incluye:

- Categorías y subcategorías de las preguntas.
- Observaciones específicas asociadas a las características del conjunto de datos.
- Reglas generadas por los modelos para cada observación.
- Visualizaciones globales y locales de los modelos.

Listing 4.1: Ejemplo de configuración de una pregunta en JSON

```
1    {
2        "id": 1,
3        "category": "Exactitud",
4        "sub_category": "Reglas",
5        "instructions": "Selecciona la predicción correcta para esta observación.",
6        "model": "DT-InterpretML",
7        "observation": {
8            "absences": 18,
```

```
            "goout": 2,
            "studytime": 2,
            "reason_reputation": 0,
            "failures": 0,
            "Fedu": 1
        },
        "rules": [
            "si failures <= 0.50 y reason_reputation <= 0.50 y Fedu <= 1.50 entonces Reprobado
                "
        ],
        "prediction_model": {
            "DT-InterpretML": "Reprobado",
            "IDS": "Reprobado"
        },
        "real_class": "Reprobado"
    }
```


El cuestionario completo puede consultarse en el Anexo E.

4.7. Pruebas de la Herramienta de Interpretabilidad

Se crearon tres usuarios de prueba ficticios para simular la contestación del cuestionario y validar el correcto funcionamiento de la herramienta *Survey-XAI-App* y la correcta integración de las preguntas y visualizaciones. El resumen de respuestas se muestra en las siguientes tablas.

Cuadro 4.11: Usuario 1

#	Valor	VPS
1	Aprobado	-
2	Reprobado	-
3	Aprobado	-
4	Reprobado	-
5	Aprobado	-
6	Reprobado	-
7	Aprobado	Mucho
8	Reprobado	Nada
9	Aprobado	Poco
10	Reprobado	Mucho
11	Aprobado	Nada
12	Reprobado	Poco
13	Correcto	Mucho
14	Incorrecto	Nada
15	Correcto	Poco
16	Incorrecto	Mucho
17	Correcto	Nada
18	Incorrecto	Poco
19	IDS	-
20	DT	-

Cuadro 4.12: Usuario 3

#	Valor	VPS
1	Aprobado	-
2	Aprobado	-
3	Reprobado	-
4	Reprobado	-
5	Aprobado	-
6	Aprobado	-
7	Reprobado	Poco
8	Reprobado	Poco
9	Aprobado	Nada
10	Aprobado	Nada
11	Reprobado	Mucho
12	Reprobado	Mucho
13	Correcto	Poco
14	Correcto	Poco
15	Incorrecto	Nada
16	Incorrecto	Nada
17	Correcto	Mucho
18	Correcto	Mucho
19	IDS	-
20	IDS	-

Cuadro 4.13: Usuario 2

#	Valor	VPS
1	Reprobado	-
2	Aprobado	-
3	Reprobado	-
4	Aprobado	-
5	Reprobado	-
6	Aprobado	-
7	Reprobado	Nada
8	Aprobado	Poco
9	Reprobado	Mucho
10	Aprobado	Nada
11	Reprobado	Poco
12	Aprobado	Mucho
13	Incorrecto	Nada
14	Correcto	Poco
15	Incorrecto	Mucho
16	Correcto	Nada
17	Incorrecto	Poco
18	Correcto	Mucho
19	DT	-
20	IDS	-

Figura 4.21: Respuestas de los usuarios de prueba. VPS = Valor en Pregunta de Seguimiento.

Además de las tablas, las gráficas generadas a partir de estos resultados pueden consultarse en el Anexo F. En el archivo *resultados_predicciones.xlsx*, también están disponibles, en las hojas *test_1*, *test_2*, *test_3* y *charts* respectivamente.

Capítulo 5

Análisis de Resultados

En este capítulo se analiza la *interpretabilidad* y el *rendimiento predictivo* de los modelos evaluados (*DT-InterpretML*, *DT-Scikit-learn* e *IDS*). El contenido se organiza en las siguientes secciones:

- *Rendimiento Predictivo [5.1]*: Se evalúa la capacidad de los modelos para clasificar correctamente en las categorías *Aprobado* y *Reprobado*.
- *Propiedades Estructurales [5.2]*: Análisis de parsimonia (simplicidad), cobertura (proporción de datos explicados) y solapamiento (conflicto entre reglas).
- *Métricas de Interpretabilidad [5.3]*: Evaluación del efecto de las propiedades estructurales y el rendimiento predictivo en la percepción de interpretabilidad.
- *Relación Precisión-Parsimonia [5.4]*: Se evalúa el equilibrio entre simplicidad estructural (longitud de reglas) y rendimiento predictivo.
- *Distribución de Probabilidades de Predicción [5.5]*: Análisis de la confianza y estabilidad de las probabilidades asignadas por los modelos en sus predicciones.
- *Generación de Reportes [5.7]*: Descripción del uso de *Survey-XAI-App* para generar reportes consolidados y gráficos categorizados para facilitar el análisis de las respuestas de los usuarios.
- *Visualizaciones [5.8]*: Descripción de grafos locales y globales generados y adaptados para cada modelo con el fin de incluirlos en el cuestionario.

5.1. Evaluación del Rendimiento Predictivo

En esta sección se analiza el desempeño de los modelos *DT-InterpretML*, DT-Scikit-learn e *IDS* utilizando métricas clave como precisión, recall, F1-score y *accuracy* global para ofrecer una visión integral del rendimiento predictivo en las clases *Aprobado* y *Reprobado*.

Cuadro 5.1: Métricas de rendimiento de los modelos evaluados.

Modelo	**Precisión Global**	**Aprobado**			**Reprobado**		
		Precisión	**Recall**	**F1-score**	**Precisión**	**Recall**	**F1-score**
DT-Scikit-learn	0.6639	0.7037	0.7808	0.7403	0.5789	0.4783	0.5238
DT-InterpretML	0.6639	0.6988	0.7945	0.7436	0.5833	0.4565	0.5122
IDS	0.5462	0.6418	0.5890	0.6143	0.4231	0.4783	0.4489

La Tabla 5.1 resume las métricas evaluadas. Tanto *DT-InterpretML* como *DT-Scikit-learn* alcanzaron un *accuracy* global del 66.39 %, superando al modelo IDS (54.62 %). Sin embargo, se observan diferencias significativas entre los modelos:

- *DT-Scikit-learn:* Sobresale en precisión (70.37 %) y recall (78.08 %) para la clase *Aprobado*, mostrando alta capacidad para identificar estudiantes aprobados, pero con un desempeño limitado en la clase *Reprobado* (recall: 47.83 %).
- *DT-InterpretML:* Presenta métricas similares al modelo de *Scikit-learn*, pero con un mayor recall en la clase *Aprobado* (79.45 %).
- *IDS:* Aunque su *accuracy* global es menor, muestra un balance más equitativo entre las clases, con un recall del 58.90 % para *Aprobado* y 47.83 % para *Reprobado*.

La Figura 5.1 ilustra las fortalezas y debilidades de los modelos en las métricas evaluadas. Tanto *DT-InterpretML* como *DT-Scikit-learn* destacan en precisión general y *recall* para la clase *Aprobado*, aunque muestran limitaciones en la detección de la clase *Reprobado*. Por su parte, el modelo *IDS* ofrece un balance más equitativo entre ambas clases, aunque con una menor precisión global.

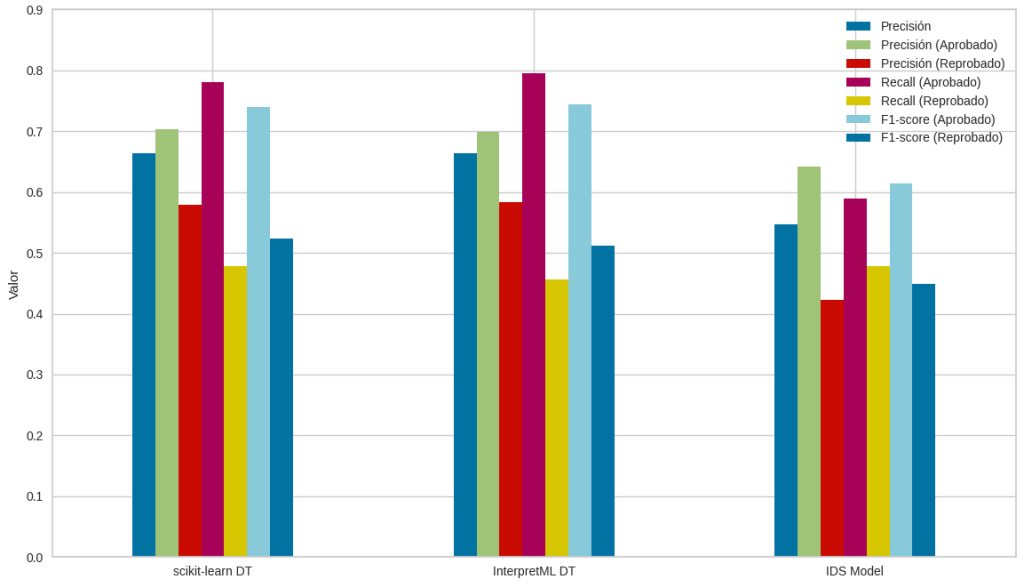


Figura 5.1: Comparación de métricas entre los modelos evaluados.

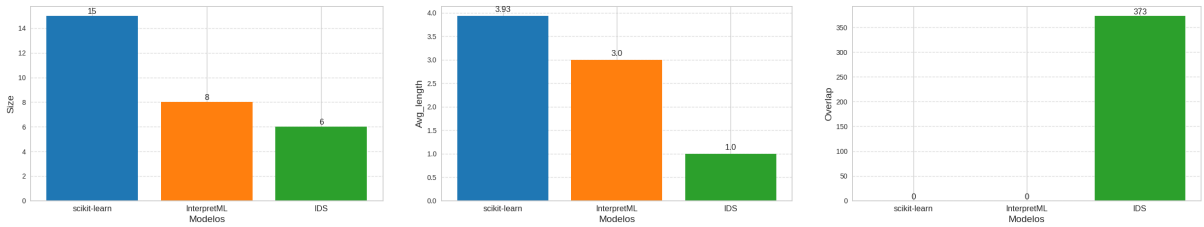
A pesar de estas limitaciones, los modelos mantienen su capacidad para evaluar la interpretabilidad, gracias a la simplicidad y claridad de sus reglas y visualizaciones.

5.2. Propiedades Estructurales

En este trabajo, las propiedades estructurales definidas en los fundamentos teóricos se emplean para analizar las características intrínsecas de los modelos. Estas propiedades incluyen el tamaño (*size*), la longitud promedio de las reglas (*avg_length*), la cobertura (*coverage*) y el solapamiento (*overlap*), las cuales cuantifican la simplicidad, claridad y organización de las reglas.

Las propiedades estructurales se calcularon utilizando el conjunto de datos de entrenamiento balanceado para reflejar las condiciones bajo las cuales los modelos fueron diseñados. Este análisis es crucial para evaluar cómo las características internas del modelo afectan su capacidad de interpretación.

La Figura 5.2 muestra que el modelo *IDS* genera reglas más simples y específicas, permitiendo calcular métricas exclusivas como cobertura y solapamiento. En contraste, los modelos *DT-InterpretML* y *DT-scikit-learn*, al ser jerárquicos, no presentan solapamiento y explican todos los datos, haciendo innecesaria la métrica de cobertura. Estos resultados destacan cómo las propiedades estructurales influyen en la claridad inherente de los modelos.



(a) Número de reglas. (b) Longitud promedio. (c) Solapamiento (IDS).

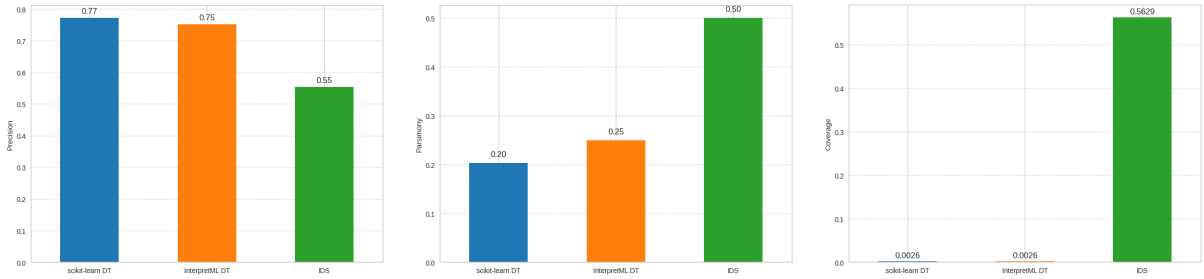
Figura 5.2: Comparación de propiedades estructurales entre modelos.

Métricas complementarias como el Índice de Gini, sparsidad, profundidad total y número de reglas se han analizado para ampliar la comprensión de las propiedades estructurales. Estas métricas, que no forman parte del análisis principal, pueden consultarse en la Sección A del anexo.

5.3. Cálculo de Métricas de Interpretabilidad

Mientras que las propiedades estructurales analizan las características internas de los modelos, las métricas de interpretabilidad evalúan cómo estas propiedades, combinadas con el rendimiento predictivo, impactan la percepción de interpretabilidad y utilidad práctica de los modelos. Estas métricas incluyen precisión, parsimonia, cobertura y sparsidad.

La Figura 5.3 compara estas métricas para los tres modelos evaluados. Adicionalmente, una gráfica complementaria que muestra la cobertura total (*cover*) de los modelos, calculada como la cantidad de instancias explicadas por las reglas generadas, puede consultarse en el Anexo B.



(a) Precisión. (b) Parsimonia. (c) Cobertura (IDS).

Figura 5.3: Comparación de métricas de interpretabilidad entre los modelos.

Se observa que los modelos priorizan distintos aspectos de interpretabilidad y precisión:

- *DT-Scikit-learn:* Prioriza la precisión, pero su complejidad estructural reduce la interpretabilidad.
- *DT-InterpretML:* Equilibra precisión, simplicidad y claridad de reglas.
- *IDS:* Sobresale en parsimonia y cobertura, aunque el solapamiento puede dificultar su interpretación.

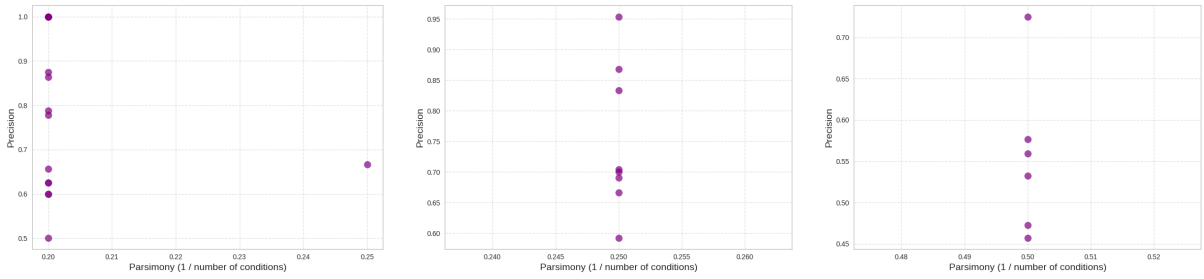
Esto evidencia que la interpretabilidad no depende solo de propiedades estructurales, sino también de cómo los usuarios perciben la simplicidad y claridad de las reglas. Por ello, el cuestionario de este trabajo se centra en evaluar la facilidad de comprensión y aplicación de las reglas generadas.

5.4. Relación Precisión-Parsimonia

Esta sección analiza la relación entre precisión y parsimonia, que refleja el equilibrio entre rendimiento predictivo y simplicidad estructural.

La parsimonia, definida como el inverso de la longitud promedio de las reglas, mide la simplicidad relativa de los modelos. Este análisis complementa los estudios previos al resaltar cómo reglas más simples mejoran la interpretabilidad percibida, especialmente junto con un buen rendimiento predictivo.

La Figura 5.4 compara la precisión y parsimonia de los modelos evaluados:



(a) *DT-Scikit-learn.* (b) *DT-InterpretML.* (c) *IDS.*

Figura 5.4: Relación entre precisión y parsimonia (1/longitud de reglas) para los modelos evaluados.

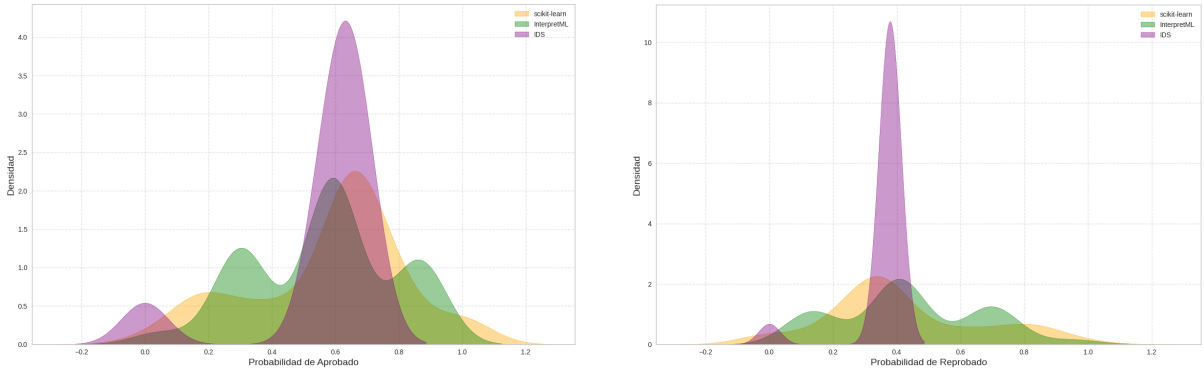
Los resultados reflejan compromisos distintos entre precisión y parsimonia:

- *Modelo Scikit-learn:* Logra alta precisión (0.70), pero su baja parsimonia (0.20) refleja la complejidad de sus reglas, lo que limita su interpretabilidad estructural.
- *Modelo InterpretML:* Presenta un balance intermedio entre parsimonia (0.25) y precisión (0.70), destacándose como un modelo equilibrado en rendimiento y claridad estructural.
- *Modelo IDS:* Priorizando la simplicidad, alcanza una alta parsimonia (0.50) gracias a la brevedad de sus reglas, aunque con una precisión más baja (0.55).

Estos resultados muestran que *IDS* prioriza la simplicidad, *DT-InterpretML* equilibra claridad y precisión, mientras que *DT-Scikit-learn*, aunque el más preciso, presenta desafíos interpretativos por su complejidad. Esto resalta la necesidad de evaluar tanto la simplicidad estructural como el rendimiento predictivo en la percepción del usuario, acorde al cuestionario desarrollado. Adicionalmente, la relación precisión-cobertura, puede consultarse en el Anexo C.

5.5. Distribución de Probabilidades de Predicción

La distribución de probabilidades de predicción proporciona información sobre cómo los modelos asignan confianza a sus decisiones.



(a) Clase *Aprobado* (b) Clase *Reprobado*

Figura 5.5: Distribución de probabilidades de predicción en los modelos evaluados. *IDS* prioriza la consistencia y claridad, mientras que *DT-InterpretML* y *DT-Scikit-learn* ofrecen mayor flexibilidad, aunque con diferentes grados de dispersión.

En la Figura 5.5 destacan las siguientes observaciones:

- *DT-Scikit-learn:* La amplia dispersión y las probabilidades extremas capturan más detalles en las predicciones, pero disminuyen la estabilidad, complicando su interpretación.
- *DT-InterpretML:* La dispersión de probabilidades en un rango amplio (0.3 a 0.9) muestra un equilibrio entre granularidad y variabilidad, aunque puede dificultar su interpretación en ciertos casos.

- *IDS:* Las distribuciones concentradas en valores intermedios (0.6 para *Aprobado* y 0.4 para *Reprobado*) reflejan un enfoque que asigna probabilidades de manera estable, pero reduce la diferenciación entre instancias con alta y baja probabilidad.

5.6. Cálculo de Interpretabilidad

Los resultados de la métrica de Interpretabilidad (4.1) se muestran en la Figura 5.6, donde se observa que:

- *DT-InterpretML:* Obtuvo la mayor interpretabilidad en la mayoría de las configuraciones destacando por su balance entre su baja longitud promedio de reglas y rendimiento predictivo.
- *IDS:* Sobresalió en parsimonia debido a la brevedad de sus reglas.
- *DT-Scikit-learn:* Aunque competitivo en términos de precisión, su mayor solapamiento y menor parsimonia limitaron su puntuación de interpretabilidad.

Estos hallazgos confirman que la configuración seleccionada en la Ecuación 4.1,representa un balance adecuado para este trabajo al priorizar la parsimonía y la precisión de los modelos.

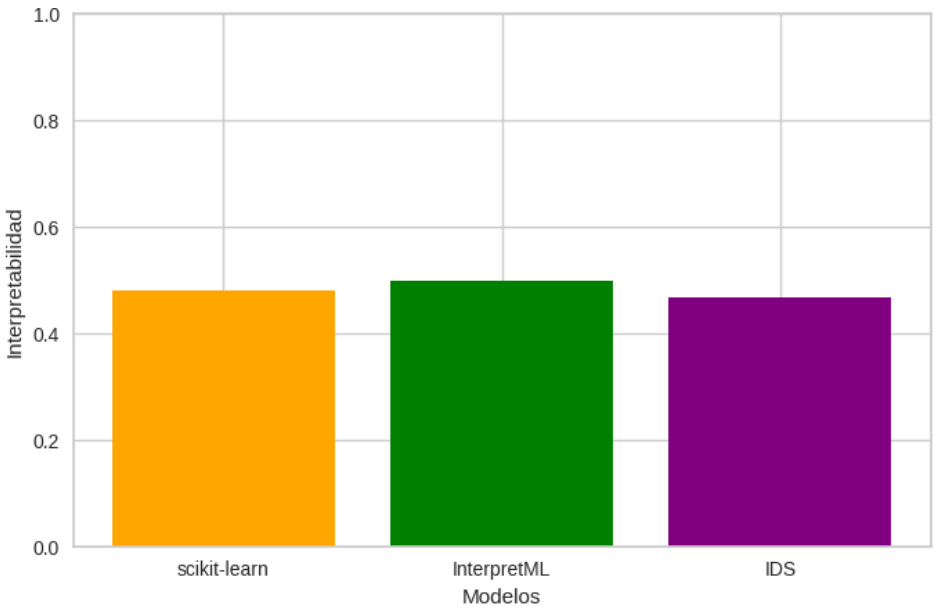


Figura 5.6: Resultados de interpretabilidad para la Configuración 1: $\alpha = 0.3$, $\beta = 0.4$, $\gamma = 0.2$ y $\delta = 0.1$.

Gráficas adicionales de interpretabilidad generadas para el resto de configuraciones de la tabla 4.7 pueden consultarse en el Anexo D.

5.7. Generación de Reportes

Survey-XAI-App genera gráficas y reportes a partir de las respuestas de los usuarios, organizadas según las categorías definidas en el archivo *questions.json*.

Se genera una gráfica para cada subcategoría (*reglas, grafo local y grafo global*), las cuales se organizan en carpetas según la categoría correspondiente (*Ambigüedad, Error, Exactitud y Preferencias de Visualización*). Todas las gráficas se almacenan en la carpeta principal *report*, que incluye una subcarpeta llamada *follow_up_question*, destinada a las gráficas de las respuestas a las preguntas de seguimiento que permiten analizar la confianza de los usuarios en sus respuestas.

La herramienta exporta los resultados en el archivo Excel *all_users_survey_report.xlsx*, como se muestra en la Figura 5.7.

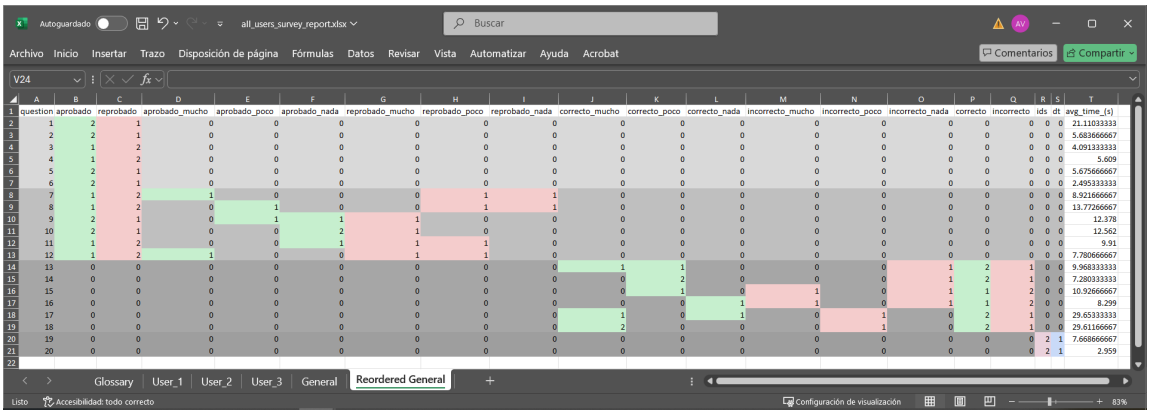


Figura 5.7: Vista del archivo Excel *all_users_survey_report.xlsx*, generado por la herramienta, con el análisis de las respuestas de los usuarios. Las categorías se diferencian mediante un tono de gris, mientras que las respuestas positivas (*Aprobado*) se resaltan en verde y las negativas (*Reprobado*) en rojo.

Para validar los resultados generados, se creó adicionalmente el archivo Excel *resultados_predicciones.xlsx*, que incluye copias de los resultados obtenidos tanto con la herramienta desarrollada como con el notebook de este TFM. La información se presenta en dos tipos de reportes principales:

- *Reportes Individuales:* Datos específicos de cada usuario (*User_1*, *User_2*, *User_3*, etc.). Incluye el tiempo empleado para responder cada pregunta a pesar de modificaciones hechas por el usuario en su respuesta.
- *Reporte Consolidado:* Un resumen general (*report_generated*) que integra la información de todos los usuarios.

El archivo final *all_users_survey_report.xlsx* incluye las hojas *Glossary*, *User_1*, *User_2*, *User_3*, *General* y *Reordered General*. Además de los tiempos individuales, los reportes consolidados incluyen una columna con el tiempo promedio empleado por los usuarios para responder cada pregunta, como se ilustra en la Figura 5.7.

Actualmente, debido a las limitaciones del servicio de alojamiento *Railway*, la generación del reporte sólo es posible de forma local. A continuación, se presentan los tipos de gráficas generadas que se explicaron anteriormente en esta sección y que corresponden a las pruebas realizadas en *Survey-XAI-App*:

- *Tiempo Promedio por Pregunta:* El cálculo se realiza con los tiempos registrados en la base de datos.

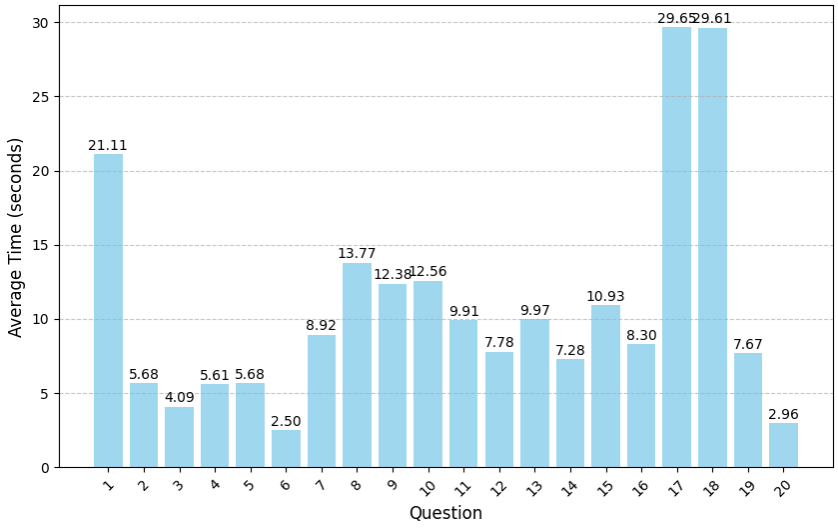
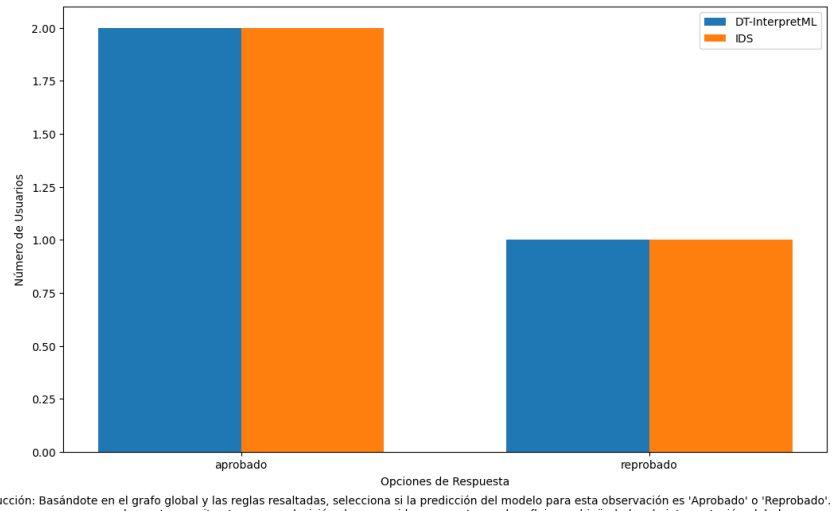


Figura 5.8: Tiempo promedio (en segundos) que los usuarios tomaron para responder cada pregunta del cuestionario.

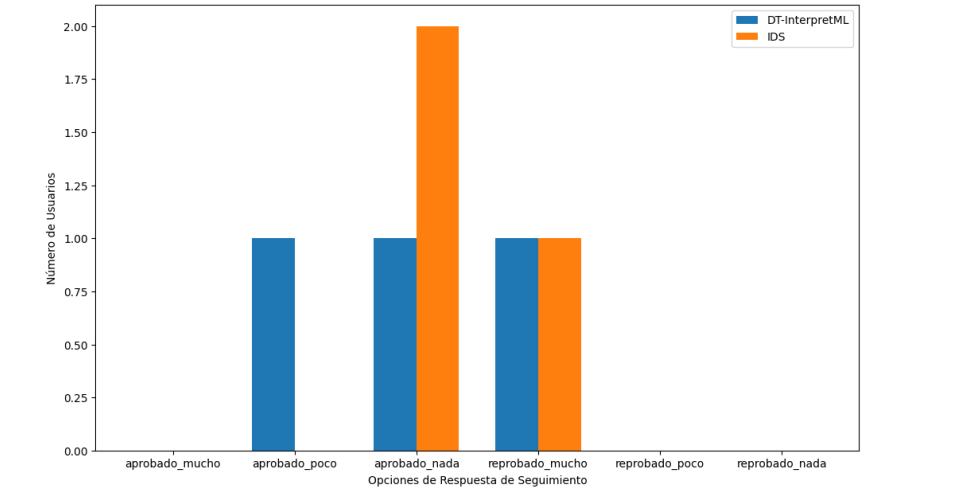
- *Ambigüedad - Grado Global:* La gráfica a continuación muestra el conteo total de respuestas en la categoría de *Ambigüedad*, subcategoría *Grado Global*. Las opciones de respuesta disponibles para cada modelo evaluado son *Aprobado* y *Reprobado*.



Pregunta ID	Modelo	Predicción del Modelo	Clase Real
ID 9	DT-InterpretML	Reprobado	Aprobado
ID 10	IDS	Aprobado	Aprobado

Figura 5.9: Distribución de respuestas para *Ambigüedad - Grado Global*. En la pregunta 9 del modelo *DT-InterpretML*, 2 usuarios calificaron la predicción como *Aprobado*, mientras que 1 usuario seleccionó *Reprobado*. El mismo resultado se observó en la pregunta 10 para la misma observación, pero correspondiente al modelo *IDS*.

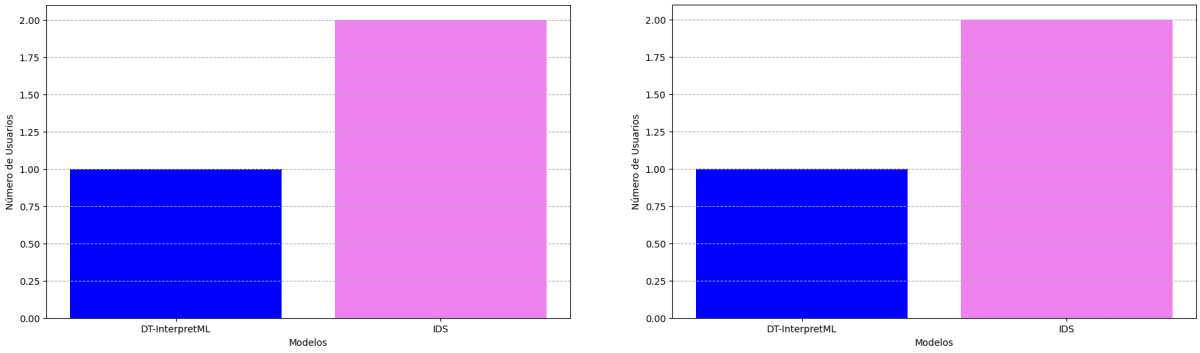
- *Ambigüedad - Seguimiento:* La gráfica muestra el análisis de respuestas de seguimiento para *Ambigüedad*, subcategoría Grado Global.



Pregunta ID	Modelo	Predicción del Modelo	Clase Real
ID 9	DT-InterpretML	Reprobado	Aprobado
ID 10	IDS	Aprobado	Aprobado

Figura 5.10: Conteo total de respuestas de seguimiento para *Ambigüedad - Grado Global*. En la pregunta 9 de *DT-InterpretML*, 2 usuarios calificaron la observación como *Aprobado*: uno estaba seguro y el otro poco seguro. Solo 1 usuario calificó como *Reprobado*, con alta certeza. Para la misma observación en la pregunta 10 de *IDS*, 2 usuarios seleccionaron *Aprobado*, pero sin certeza, y 1 usuario calificó como *Reprobado* con alta certeza.

- *Preferencia de Modelos - Preguntas 19 y 20:* Gráficas que resumen las preferencias de los usuarios sobre la interpretación de los modelos evaluados.



(a) Preferencia de modelos en la Pregunta 19. (b) Preferencia de modelos en la Pregunta 20.

Figura 5.11: Resultados de las preguntas de la categoría *Preferencias de Visualización*.

Las gráficas adicionales generadas por la herramienta *Survey-XAI-App*, se pueden revisar en el Anexo F.

5.8. Implementación de Visualizaciones

Las visualizaciones, o grafos, que fueron implementados en este TFM, y que se presentan a continuación, tienen el objetivo de facilitar la comprensión a nivel global y local de las reglas y predicciones generadas por los modelos evaluados.

5.8.1. Grafo Global del Modelo IDS

El grafo global del modelo *IDS* representa todas las reglas generadas y su conexión con las decisiones finales (*Aprobado* o *Reprobado*, en verde y rojo respectivamente). Esto facilita entender las relaciones entre reglas y su impacto en las predicciones.

Listing 5.1: Características del grafo global del modelo IDS

```
{
    "absences": 0, "goout": 2, "studytime": 2,
    "reason_reputation": 1, "failures": 0, "Fedu": 3
}
```

Cuadro 5.2: Predicciones para la observación del grafo global.

Modelo	**Predicción**
DT-InterpretML	Aprobado
IDS	Aprobado

La Figura 5.12 presenta el grafo global, resaltando en amarillo las reglas activas aplicadas a la observación.

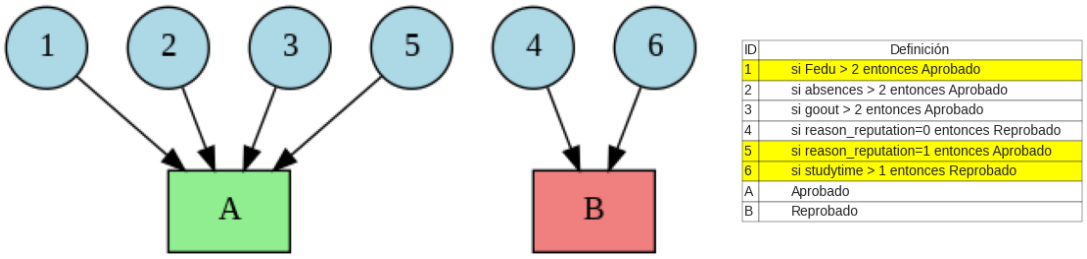


Figura 5.12: Grafo global del modelo IDS mostrando reglas y predicciones.

5.8.2. Grafo Local del Modelo IDS

El grafo local destaca con nodos amarillos en doble círculo solo las reglas activas para una observación específica, mientras que representa con nodos azules y enlaces punteados las reglas inactivas.

Listing 5.2: Características del grafo local del modelo IDS

```
{
    "absences": 8, "goout": 4, "studytime": 2,
    "reason_reputation": 0, "failures": 3, "Fedu": 1
}
```

Cuadro 5.3: Predicciones para la observación del grafo local.

Modelo	**Predicción**
DT-InterpretML	Reprobado
IDS	Reprobado

La Figura 5.13 muestra el grafo local con las reglas activas resaltadas.

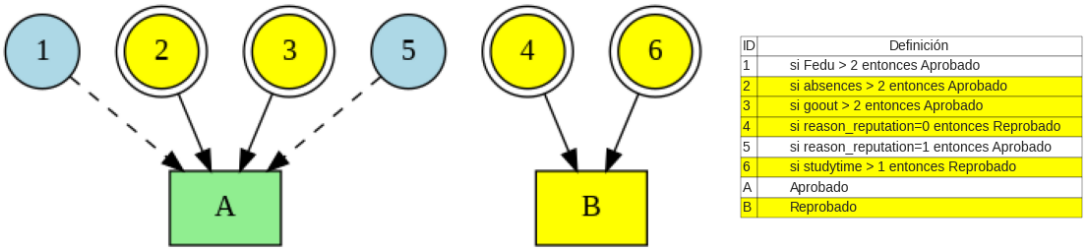


Figura 5.13: Grafo local del modelo IDS resaltando las reglas activas.

5.8.3. Visualizaciones del Modelo DT-InterpretML

El modelo *DT-InterpretML* utiliza un árbol de decisión subyacente (*Scikit-learn*) para generar visualizaciones. En este TFM, se presentan dos tipos de visualizaciones:

- *Explicación Local Nativa*: Resalta el camino seguido por una observación específica en el árbol, como se muestra en la Figura 5.14.

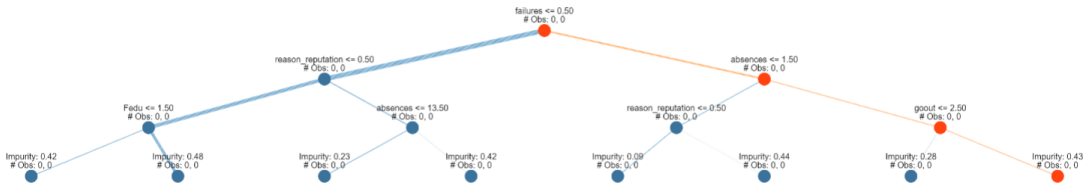


Figura 5.14: Explicación local nativa del modelo *DT-InterpretML*.

- *Árbol Personalizado*: Muestra todos los nodos del modelo, con detalles adicionales como condiciones y métricas clave que facilitan un análisis más completo (Figura 5.15).

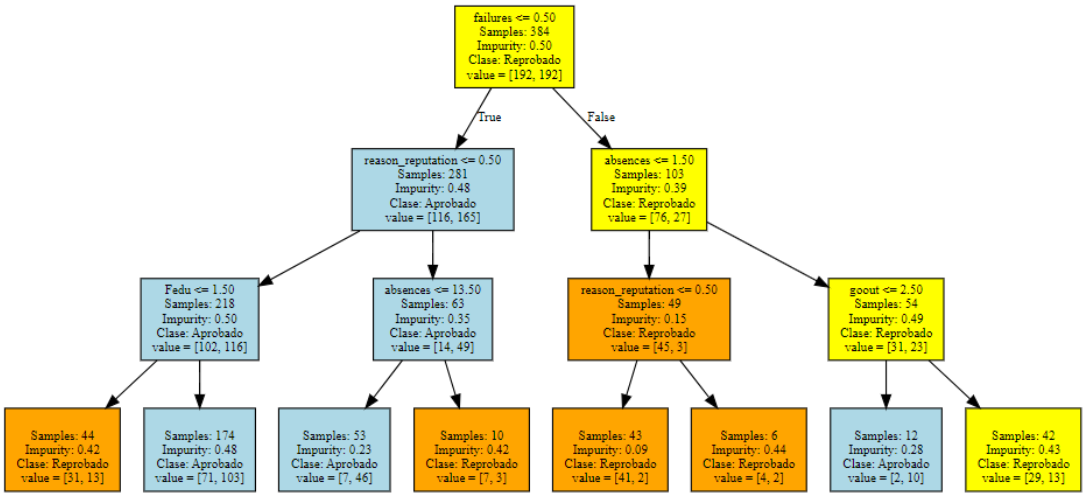


Figura 5.15: Árbol personalizado a partir del modelo *Scikit-learn* subyacente del modelo *DT-InterpretML*.

Capítulo 6

Conclusiones

Este capítulo presenta las conclusiones derivadas del estudio sobre la interpretabilidad de dos modelos transparentes: Árboles de Decisión (*DT-InterpretML*) e Interpretable Decision Sets (*IDS*). Se valida el cumplimiento de los objetivos establecidos (Sección 6.1), se reflexiona sobre las hipótesis planteadas (Sección 6.2) y se analizan las limitaciones del estudio (Sección 6.3). Finalmente, se proponen líneas de trabajo futuro para extender y mejorar este análisis (Sección 6.4) y se incluye una reflexión personal sobre los alcances de este TFM (Sección 6.5).

6.1. Logro de los Objetivos

A continuación, se detalla el grado de cumplimiento de los objetivos definidos en la Sección 1.2:

- *Objetivo General: Evaluar y comparar la interpretabilidad de modelos de decisión DT y IDS desde un enfoque técnico y perceptual.*

    - Se desarrolló una librería personalizada para el modelo *IDS*, capaz de generar grafos locales y globales para facilitar la interpretación de sus reglas.
    - Se desarrolló una herramienta funcional que permitió analizar el rendimiento y la interpretabilidad cuantitativa de los modelos (*notebook*).
    - Se desarrolló una herramienta web (*Survey-XAI-App*) que implementa un cuestionario diseñado para evaluar la percepción de interpretabilidad de los usuarios (interpretabilidad cualitativa).

- *Objetivos Específicos:*

    1. *Desarrollar una herramienta web de evaluación de interpretabilidad:*
       La herramienta fue desarrollada con éxito, integrando visualizaciones globales y locales, almacenamiento en base de datos y generación automatizada de reportes y gráficas.
    2. *Analizar el desempeño técnico de los modelos:*
       Se compararon métricas clave como precisión, *recall* y *F1-score*. Los modelos *DT-InterpretML* y *DT-Scikit-learn* demostraron mayor precisión global,

mientras que el modelo *IDS* mostró un balance entre clases con menor precisión global.

3. *Evaluar las propiedades estructurales de los modelos:*
   Se analizaron propiedades como el tamaño, la longitud promedio de las reglas, solapamiento y cobertura. El modelo *IDS* generó reglas más simples, mientras que los modelos *DT* presentaron estructuras más complejas.
4. *Diseñar un cuestionario para evaluar la percepción de interpretabilidad:*
   Se diseñó un cuestionario con preguntas clasificadas en categorías como *exactitud*, *ambigüedad* y *error*. Se añadieron también preguntas de seguimiento para evaluar la confianza de los usuarios en sus respuestas.
5. *Realizar pruebas internas de la herramienta:*
   Aunque no se aplicó el cuestionario a una muestra externa, las pruebas internas validaron la funcionalidad y el cumplimiento de los objetivos planteados.

6.2. Validación de Hipótesis

El análisis permitió obtener conclusiones preliminares respecto a las hipótesis formuladas en la Sección 1.4:

- *Hipótesis 1:* Las reglas generadas por el modelo *IDS* son más simples y específicas, facilitando su interpretación.
  Conclusión: Validada. El modelo *IDS* generó reglas más cortas que se aplican independientemente para elegir la clase por votación o mediante la clase por defecto en caso de empate.
- *Hipótesis 2:* Las visualizaciones globales y locales de *DT-InterpretML* facilitan la comprensión de las predicciones.
  Conclusión: Parcialmente validada, ya que no se adminitró el cuestionario a una muestra de usuarios reales. Sin embargo, las visualizaciones personalizadas se crearon a partir del modelo subyacente de *DT-InterpretML*, herramienta reconocida en el campo de la inteligenciaartificial explicable por su representación clara del proceso de decisión que facilita la interpretación local y global.
- *Hipótesis 3:* Existe un equilibrio entre precisión y simplicidad estructural en los modelos evaluados.
  Conclusión: Parcialmente validada. *DT-InterpretML* logró un mejor balance entre precisión y parsimonia, mientras que *IDS* priorizó la simplicidad a expensas del rendimiento predictivo.

6.3. Limitaciones del Estudio

El presente estudio presenta las siguientes limitaciones:

- El cuestionario no se aplicó a usuarios reales debido a restricciones logísticas y de tiempo.
- La validación de la percepción de interpretabilidad se basó únicamente en pruebas internas de la herramienta.

- No se consideraron aspectos adicionales como la escalabilidad de los modelos o su rendimiento en grandes volúmenes de datos.

6.4. Trabajo Futuro

Este estudio ofrece una base sólida para futuros trabajos en el campo de la interpretabilidad de modelos transparentes. Las principales líneas de trabajo futuro incluyen:

- *Aplicación del cuestionario en estudios reales:* Validar empíricamente la percepción de interpretabilidad mediante pruebas con usuarios reales, como estudiantes y profesionales del área de inteligencia artificial.
- *Optimización de los modelos evaluados:* Explorar configuraciones avanzadas de *IDS* y *DT-InterpretML*, además de incluir nuevos modelos interpretables como *Explainable Boosting Machine (EBM)*.
- *Ampliación de la herramienta de evaluación:* Incorporar nuevas funcionalidades, como generación dinámica de reportes interactivos, análisis automatizados de respuestas y métricas adicionales de interpretabilidad. Por ejemplo, qué tan alineadas están las reglas locales con las globales, o cómo integrar al cuestionario el contexto y el tipo de audiencia.
- *Publicación de resultados:* Difundir los hallazgos obtenidos en conferencias y revistas científicas especializadas en inteligencia artificial explicable (*XAI*).
- *Alineación con la Ley de Inteligencia Artificial Europea:* La herramienta propuesta podría alinearse con los principios establecidos en la Ley Europea de Inteligencia Artificial [46].

6.5. Reflexión Final

Este trabajo ha demostrado la importancia de combinar análisis técnicos y perceptuales para evaluar la interpretabilidad de modelos transparentes en inteligencia artificial. A lo largo del estudio, se lograron avances significativos, como la implementación de una herramienta funcional y el diseño de un cuestionario estructurado que sienta las bases para futuras investigaciones en este campo.

Los resultados obtenidos reflejan un primer paso sólido en la evaluación de modelos interpretables. Sin embargo, soy consciente de que este es solo el comienzo de un camino prometedor. Reafirmo mi interés y compromiso por dar seguimiento a este proyecto, aplicando los aprendizajes adquiridos durante el máster para continuar desarrollando y optimizando la herramienta. La aplicación del cuestionario en estudios reales, la inclusión de nuevos modelos interpretables y la alineación con los marcos normativos actuales, son algunas de las líneas que pretendo explorar en el futuro.

Este trabajo contribuye al campo de la *XAI* y refleja mi aspiración de seguir aportando al desarrollo de sistemas de IA más transparentes, éticos y confiables para facilitar su adopción y comprensión tanto en el ámbito académico como en la práctica profesional.

Bibliografía

[1] Himabindu Lakkaraju, Stephen H. Bach y Jure Leskovec. «Interpretable Decision Sets: A Joint Framework for Description and Prediction». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, págs. 1675-1684. ISBN: 9781450342322. DOI: 10.1145/2939672.2939874. URL: https://doi.org/10.1145/2939672.2939874.

[2] P. Cortez y A. M. Gonçalves Silva. «Using data mining to predict secondary school student performance». En: 2008. URL: https://api.semanticscholar.org/CorpusID:16621299.

[3] W Samek. *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.

[4] Amnesty International. *El escándalo de los subsidios para el cuidado infantil en Países Bajos, una alerta urgente para prohibir los algoritmos racistas*. Accessed: 2024-08-28. 2021. URL: https://www.amnesty.org/es/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/.

[5] BBC News Mundo. *"Le arruinaron la vida a gente inocente": el escándalo que hizo dimitir en bloque al gobierno de Países Bajos*. Accessed: 2024-08-28. 2021. URL: https://www.bbc.com/mundo/noticias-internacional-55683795.

[6] Euronews. *El escándalo por la discriminación racial en las ayudas familiares cerca al Gobierno de Rutte*. https://es.euronews.com/2021/01/13/el-escandalo-por-la-discriminacion-racial-en-las-ayudas-familiares-cerca-al-gobierno-de-ru. Último acceso: 28 de agosto de 2024. 2021.

[7] BBC Mundo. *¿Cómo en Estados Unidos las matemáticas te pueden meter en prisión?* Último acceso: agosto 28, 2024. 2016. URL: https://www.bbc.com/mundo/noticias-37679463.

[8] Angwin, Julia and Larson, Jeff and Mattu, Surya and Kirchner, Lauren. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Último acceso: 28 de agosto de 2024. 2016.

[9] Finale Doshi-Velez y Been Kim. «Towards a rigorous science of interpretable machine learning». En: *arXiv preprint arXiv:1702.08608* (2017).

[10] David Gunning et al. «DARPA's explainable AI (XAI) program: A retrospective». En: *Applied AI Letters* 2.4 (2021), e61.

[11] W James Murdoch et al. «Interpretable machine learning: definitions, methods, and applications». En: *arXiv preprint arXiv:1901.04592* (2019).

[12] Harmanpreet Kaur et al. «Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning». En: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, págs. 1-14. ISBN: 9781450367080.

[13] Leilani H Gilpin et al. «Explaining explanations: An overview of interpretability of machine learning». En: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, págs. 80-89.

[14] Cynthia Rudin. «Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead». En: *Nature Machine Intelligence* (2019).

[15] Sunil L Kukreja, Johan Löfberg y Martin J Brenner. «A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification». En: *IFAC proceedings volumes* 39.1 (2006), págs. 814-819.

[16] Trevor J Hastie. «Generalized additive models». En: *Statistical models in S*. Routledge, 2017, págs. 249-307.

[17] Ilia Rushkin. «Optimizing the Ptolemaic Model of Planetary and Solar Motion». En: *arXiv preprint arXiv:1502.01967* (2015).

[18] J Nathan Kutz y Steven L Brunton. «Parsimony as the ultimate regularizer for physics-informed machine learning». En: *Nonlinear Dynamics* 107.3 (2022), págs. 1801-1817.

[19] Richard O. Duda, Peter E. Hart y David G. Stork. *Pattern Classification*. 2nd Edition. Wiley, 2000.

[20] Franciso Herrera et al. «An overview on subgroup discovery: foundations and applications». En: *Knowledge and Information Systems* 29.3 (2010).

[21] Stefan Wrobel. «Relational Data Mining». En: Springer, 2001. Cap. Inductive Logic Programming for Knowledge Discovery in Databases.

[22] Harsha Nori et al. «Interpretml: A unified framework for machine learning interpretability». En: *arXiv preprint arXiv:1909.09223* (2019).

[23] Ibomoiye Domor Mienye y Nobert Jere. «A Survey of Decision Trees: Concepts, Algorithms, and Applications». En: *IEEE Access* (2024).

[24] Han Liu, Alexander Gegov y Mihaela Cocea. *Rule Based Systems for Big Data*. Springer, 2015.

[25] Benjamin Letham et al. «Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model». En: *The Annals of Applied Statistics* 9.3 (2015).

[26] Ronald L Rivest. «Learning decision lists». En: *Machine learning* 2 (1987), págs. 229-246.

[27] Alexey Ignatiev et al. «A SAT-Based Approach to Learn Explainable Decision Sets». En: *Automated Reasoning*. Ed. por Didier Galmiche, Stephan Schulz y Roberto Sebastiani. Cham: Springer International Publishing, 2018, págs. 627-645. ISBN: 978-3-319-94205-6.

[28] Leo Breiman et al. *Classification and Regression Trees*. Chapman y Hall, 1984.

[29] Caglar Aytekin. «Neural networks are decision trees». En: *arXiv preprint arXiv:2210.05189* (2022).

[30] Forough Poursabzi-Sangdeh et al. «Manipulating and Measuring Model Interpretability». En: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380966.

[31] Zachary C. Lipton. *The mythos of model interpretability*. Disponible en arXiv. 2016. arXiv: 1606.03490 [stat.ML]. URL: https://arxiv.org/abs/1606.03490.

[32] Marco Tulio Ribeiro, Sameer Singh y Carlos Guestrin. «"Why should i trust you?."Explaining the predictions of any classifier». En: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, págs. 1135-1144.

[33] Bojan Mihaljevic y Esteban García. *Notas del curso XAI del Máster en Inteligencia Artificial de la Universidad Politécnica de Madrid*. 2024.

[34] Scott Lundberg y Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]. URL: https://arxiv.org/abs/1705.07874.

[35] Jimmy Lin et al. «Generalized and Scalable Optimal Sparse Decision Trees». En: *Proceedings of the 37th International Conference on Machine Learning*. Ed. por Hal Daumé III y Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul de 2020, págs. 6150-6160.

[36] Xiyang Hu, Cynthia Rudin y Margo Seltzer. «Optimal Sparse Decision Trees». En: *Advances in Neural Information Processing Systems*. Ed. por H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019.

[37] Sandra G. Hart y Lowell E. Staveland. «Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research». En: *Human Mental Workload*. Ed. por Peter A. Hancock y Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, 1988, págs. 139-183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9.

[38] Microsoft Developer. *The Science Behind InterpretML: Explainable Boosting Machine*. Accedido el: 20 de septiembre de 2024. 2020. URL: https://www.youtube.com/watch?v=MREiHgHgl0k.

[39] Benjamin Bengfort et al. *Yellowbrick*. Ver. 0.9.1. 14 de nov. de 2018. DOI: 10.5281/zenodo.1206264. URL: http://www.scikit-yb.org/en/latest/.

[40] Marco Tulio Ribeiro, Sameer Singh y Carlos Guestrin. «Anchors: High-precision model-agnostic explanations». En: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[41] Ramaravind K Mothilal, Amit Sharma y Chenhao Tan. «Explaining machine learning classifiers through diverse counterfactual explanations». En: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, págs. 607-617.

[42] Paulo Cortez. «Student performance». En: *UCI machine learning repository* (2014).

[43] Jiri Filip. *pyIDS: Python Implementation of Interpretable Decision Sets*. Accessed: 2024-11-21. 2020. URL: https://github.com/jirifilip/pyIDS.

[44] Hima Bindu. *Interpretable Decision Sets*. Accessed: 2024-11-21. 2018. URL: https://github.com/lvhimabindu/interpretable_decision_sets.

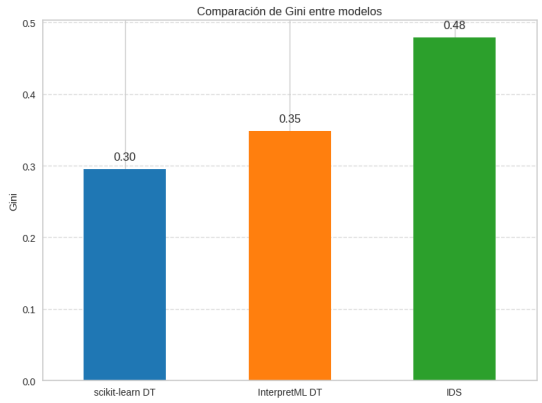
[45] Stuart Mitchell, Michael OSullivan y Iain Dunning. «Pulp: a linear programming toolkit for python». En: *The University of Auckland, Auckland, New Zealand* 65 (2011), pág. 25.

[46] Unión Europea. *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, por el que se establecen normas armonizadas sobre inteligencia artificial*. Accedido el 17 de junio de 2024. 2024. URL: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=OJ:L_202401689.

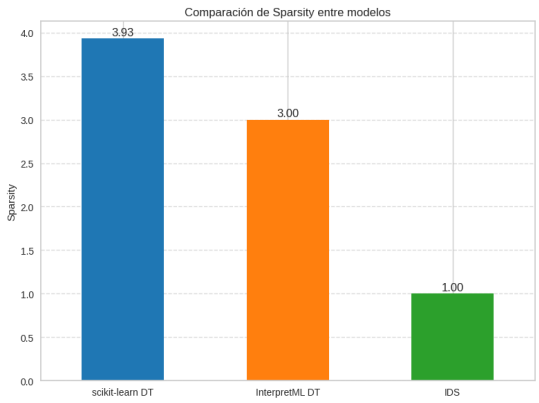
Anexo A

Métricas Complementarias de Interpretabilidad

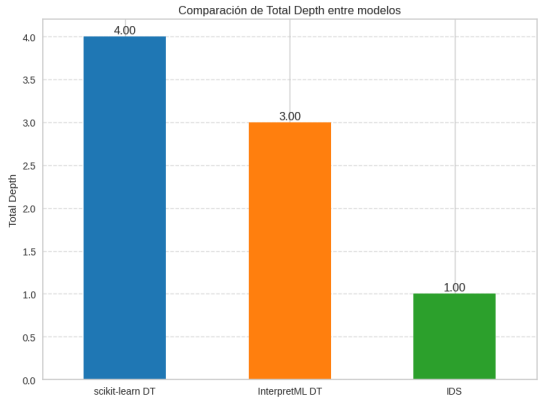
En esta sección se presentan las métricas complementarias utilizadas para evaluar los modelos. Estas gráficas ofrecen detalles adicionales sobre las propiedades estructurales y de complejidad de los modelos evaluados.



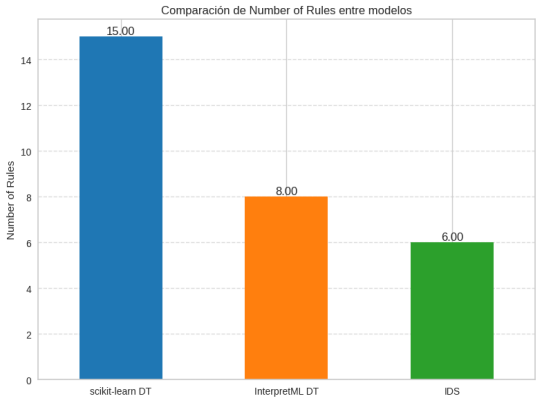
(a) Índice de Gini.



(b) Sparsidad.



(c) Profundidad total.



(d) Número de reglas.

Figura A.1: Métricas complementarias de interpretabilidad.

Anexo B

Propiedad Complementaria: Cobertura

Además de las propiedades estructurales discutidas en la metodología, se calculó la cobertura total (*cover*) para los modelos evaluados. Esta métrica indica la cantidad de instancias explicadas por las reglas generadas. En este caso, los tres modelos cubren exactamente el mismo número de instancias, como se muestra en la Figura B.1.

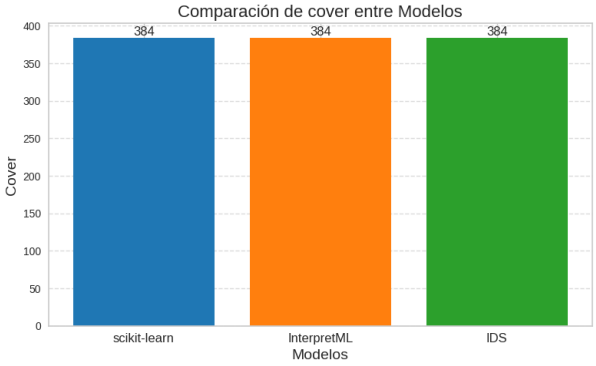


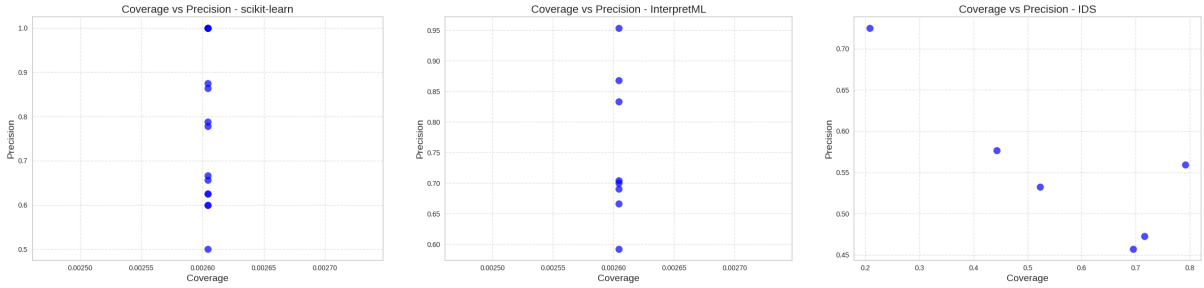
Figura B.1: Cobertura total (*cover*) entre modelos.

El valor uniforme de cobertura (384) entre los modelos confirma que todos explican adecuadamente el conjunto de datos, lo que indica que las diferencias en sus propiedades estructurales no afectan su capacidad para cubrir las instancias del dataset.

Anexo C

Relación Precisión-Cobertura

Como análisis complementario, se evaluó la relación entre la cobertura (*coverage*) y la precisión de los modelos. La cobertura mide la proporción de datos explicados por las reglas. La Figura C.1 presenta los resultados.



(a) *DT-Scikit-learn.* (b) *DT-InterpretML.* (c) *IDS.*

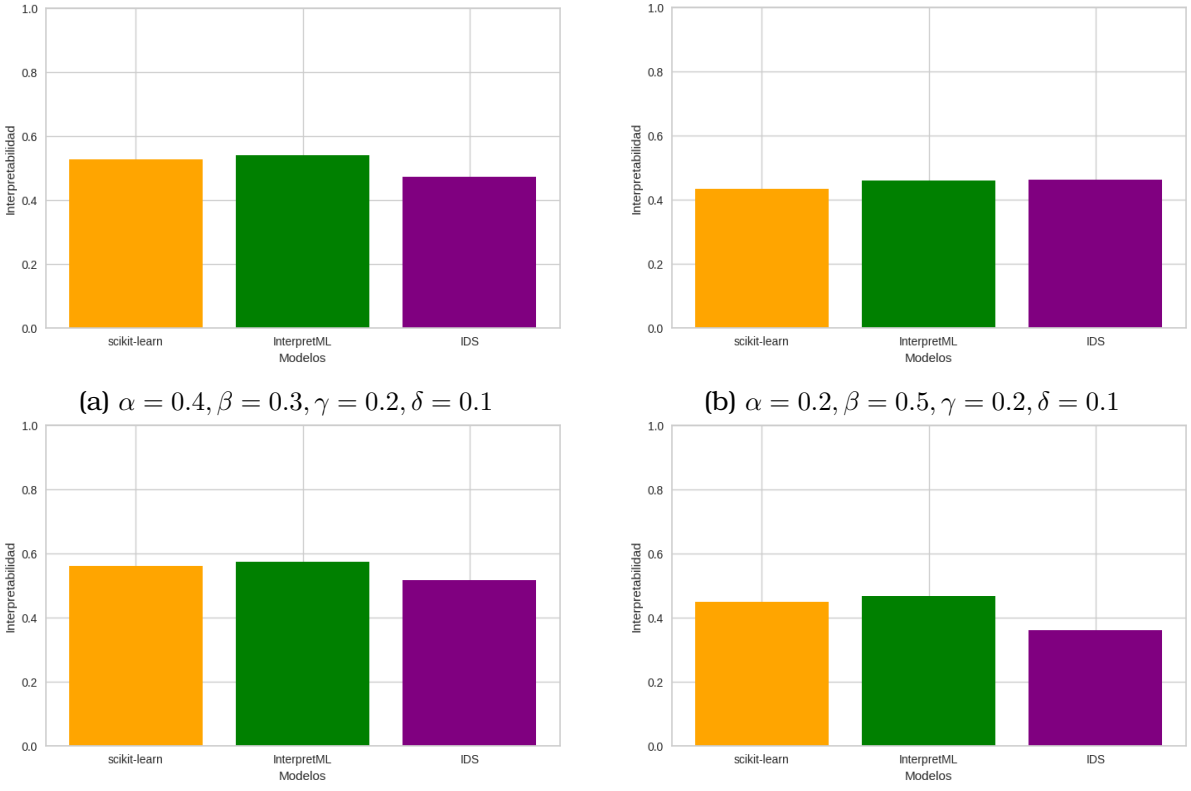
Figura C.1: Relación entre precisión y cobertura (*coverage*) para los modelos evaluados.

Los modelos scikit-learn e InterpretML muestran una cobertura baja (0.0026), lo que indica reglas más específicas. Por otro lado, IDS presenta una amplia cobertura (0.7), aunque a costa de una reducción en la precisión (0.45-0.55).

Anexo D

Configuraciones para la Métrica de Interpretabilidad

En esta sección, se presentan las gráficas correspondientes a las pruebas realizadas para evaluar diferentes configuraciones de los pesos α, β, γ y δ de la métrica de interpretabilidad definida en la Sección 5.6.



(a) $\alpha = 0.4, \beta = 0.3, \gamma = 0.2, \delta = 0.1$

(b) $\alpha = 0.2, \beta = 0.5, \gamma = 0.2, \delta = 0.1$

(c) $\alpha = 0.3, \beta = 0.3, \gamma = 0.3, \delta = 0.1$

(d) $\alpha = 0.1, \beta = 0.4, \gamma = 0.3, \delta = 0.2$

Figura D.1: Resultados de interpretabilidad con diferentes configuraciones de pesos.

Anexo E

Plantilla del Cuestionario de Interpretabilidad

La siguiente plantilla ejemplifica cómo se estructura el cuestionario en el archivo *questions.json*. Para cada pregunta, se incluye la observación, las reglas, las predicciones de los modelos y la respuesta esperada.

E.1. Pregunta 1

ID: 1
Categoría: Exactitud
Subcategoría: Reglas
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en las reglas proporcionadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.1: Observación de pregunta 1.

absences	goout	studytime	reason_reputation	failures	Fedu
18	2	2	0	0	1

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado

Figura E.1: Reglas de decisión para la Pregunta 1 del modelo DT-InterpretML.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Reprobado
- *Usuario:*

- ☐ Aprobado
- ☐ Reprobado

E.2. Pregunta 2

ID: 2
Categoría: Exactitud
Subcategoría: Reglas
Modelo Evaluado: IDS

Instrucciones: Basándote en las reglas proporcionadas y el grafo local, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.2: Observación de pregunta 2.

absences	goout	studytime	reason_reputation	failures	Fedu
18	2	2	0	0	1

si absences > 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado
si studytime > 1 **entonces** Reprobado

Figura E.2: Reglas de decisión para la Pregunta 2 del modelo IDS.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

E.3. Pregunta 3

ID: 3
Categoría: Exactitud
Subcategoría: Grado Global
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo global y las reglas resaltadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.3: Observación de pregunta 3.

absences	goout	studytime	reason_reputation	failures	Fedu
0	2	2	1	0	3

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $>$ 1.50 **entonces** Aprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $>$ 13.50 **entonces** Reprobado

Figura E.3: Reglas de decisión para la Pregunta 3 del modelo DT-InterpretML.

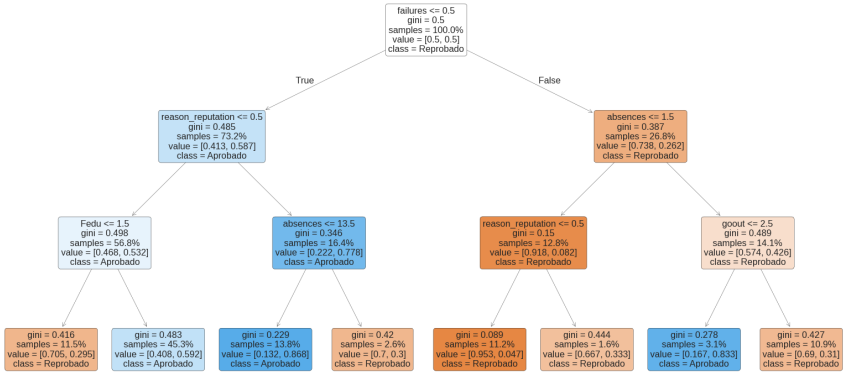


Figura E.4: Grafo global del modelo DT-InterpretML para la Pregunta 3.

Respuestas:

- *DT:* Aprobado
- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

E.4. Pregunta 4

ID: 4
Categoría: Exactitud
Subcategoría: Grado Global
Modelo Evaluado: IDS

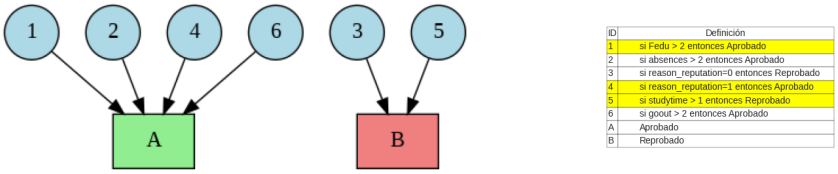
Instrucciones: Basándote en el grafo global y las reglas resaltadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.4: Observación de pregunta 4.

absences	goout	studytime	reason_reputation	failures	Fedu
0	2	2	1	0	3

si Fedu > 2 **entonces** Aprobado
si reason_reputation = 1 **entonces** Aprobado
si studytime > 1 **entonces** Reprobado

Figura E.5: Reglas de decisión para la Pregunta 4 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si reason_reputation=0 entonces Reprobado
4	si reason_reputation=1 entonces Aprobado
5	si studytime > 1 entonces Reprobado
6	si goout > 2 entonces Aprobado
A	Aprobado
B	Reprobado

Figura E.6: Grafo global del modelo IDS para la Pregunta 4.

Respuestas:

- *DT:* Aprobado
- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

E.5. Pregunta 5

ID: 5
Categoría: Exactitud
Subcategoría: Grado Local
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo local y las reglas resaltadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.5: Observación de pregunta 5.

absences	goout	studytime	reason_reputation	failures	Fedu
8	4	2	0	3	1

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $\leq$ 2.50 **entonces** Aprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $>$ 2.50 **entonces** Reprobado

Figura E.7: Reglas de decisión para la Pregunta 5 del modelo DT-InterpretML.

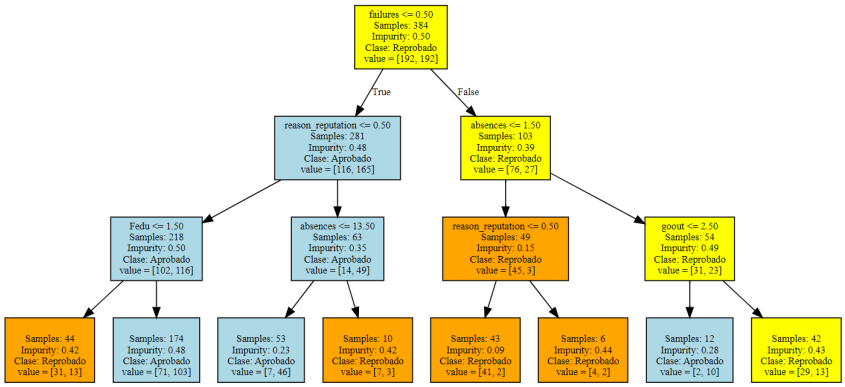


Figura E.8: Grafo local del modelo DT-InterpretML para la Pregunta 5.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

E.6. Pregunta 6

ID: 6
Categoría: Exactitud
Subcategoría: Grado Local
Modelo Evaluado: IDS

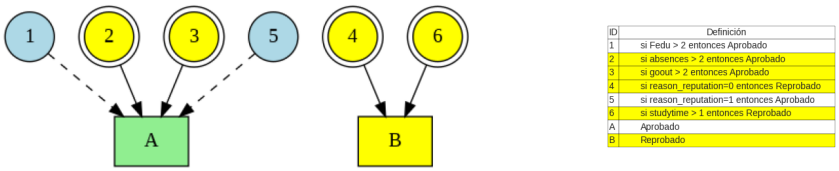
Instrucciones: Basándote en el grafo local y las reglas resaltadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”.

Cuadro E.6: Observación de pregunta 6.

absences	goout	studytime	reason_reputation	failures	Fedu
8	4	2	0	3	1

si absences $>$ 2 **entonces** Aprobado
si goout $>$ 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado
si studytime $>$ 1 **entonces** Reprobado

Figura E.9: Reglas de decisión para la Pregunta 6 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si goout > 2 entonces Aprobado
4	si reason_reputation=0 entonces Reprobado
5	si reason_reputation=1 entonces Aprobado
6	si studytime > 1 entonces Reprobado
A	Aprobado
B	Reprobado

Figura E.10: Grafo local del modelo IDS para la Pregunta 6.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

E.7. Pregunta 7

ID: 7
Categoría: Ambigüedad
Subcategoría: Reglas
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en las reglas proporcionadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”. Si al analizar las reglas no encuentras una respuesta evidente, considera que esto puede reflejar ambigüedad en las reglas.

Cuadro E.7: Observación de pregunta 7.

absences	goout	studytime	reason_reputation	failures	Fedu
10	2	1	0	0	2

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $>$ 1.50 **entonces** Aprobado

Figura E.11: Reglas de decisión para la Pregunta 7 del modelo DT-InterpretML.

Respuestas:

- *DT:* Aprobado
- *IDS:* Reprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.8. Pregunta 8

ID: 8
Categoría: Ambigüedad
Subcategoría: Reglas
Modelo Evaluado: IDS

Instrucciones: Basándote en las reglas proporcionadas, selecciona si la predicción correcta para esta observación es "Aprobado" o "Reprobado". Si al analizar las reglas no encuentras una respuesta evidente, considera que esto puede reflejar ambigüedad en las reglas.

Cuadro E.8: Observación de pregunta 8.

absences	goout	studytime	reason_reputation	failures	Fedu
10	2	1	0	0	2

si absences > 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado

Figura E.12: Reglas de decisión para la Pregunta 8 del modelo IDS.

Respuestas:

- *DT:* Aprobado
- *IDS:* Reprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado

- ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.9. Pregunta 9

ID: 9
Categoría: Ambigüedad
Subcategoría: Grado Global
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo global y las reglas resaltadas, selecciona si la predicción correcta para esta observación es "Aprobado" o "Reprobado". Si el grafo y las reglas no te permiten tomar una decisión clara, considera que esto puede reflejar ambigüedad en la interpretación global.

Cuadro E.9: Observación de pregunta 9.

absences	goout	studytime	reason_reputation	failures	Fedu
6	5	1	0	3	1

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $>$ 2.50 **entonces** Reprobado

Figura E.13: Reglas de decisión para la Pregunta 9 del modelo DT-InterpretML.

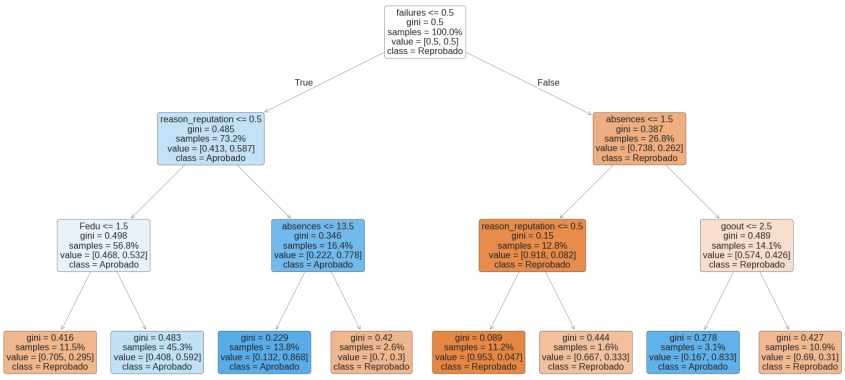


Figura E.14: Grafo global del modelo DT-InterpretML para la Pregunta 9.

Respuestas:

- *DT:* Reprobado

- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.10. Pregunta 10

ID: 10
Categoría: Ambigüedad
Subcategoría: Grado Global
Modelo Evaluado: IDS

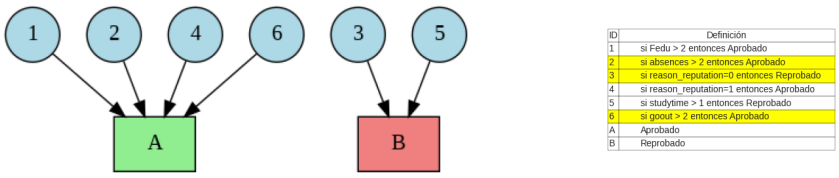
Instrucciones: Basándote en el grafo global y las reglas resaltadas, selecciona si la predicción correcta para esta observación es "Aprobado" o "Reprobado". Si el grafo y las reglas no te permiten tomar una decisión clara, considera que esto puede reflejar ambigüedad en la interpretación global.

Cuadro E.10: Observación de pregunta 10.

absences	goout	studytime	reason_reputation	failures	Fedu
6	5	1	0	3	1

si absences > 2 **entonces** Aprobado
si goout > 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado

Figura E.15: Reglas de decisión para la Pregunta 10 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si reason_reputation=0 entonces Reprobado
4	si reason_reputation=1 entonces Aprobado
5	si studytime > 1 entonces Reprobado
6	si goout > 2 entonces Aprobado
A	Aprobado
B	Reprobado

Figura E.16: Grafo global del modelo IDS para la Pregunta 10.

Respuestas:

- *DT:* Reprobado

- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.11. Pregunta 11

ID: 11
Categoría: Ambigüedad
Subcategoría: Grado Local
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo local y las reglas resaltadas, selecciona si la predicción correcta para esta observación es "Aprobado" o "Reprobado". Si al analizar las reglas y el grafo local no encuentras una justificación clara, considera que esto puede reflejar ambigüedad en la interpretación local.

Cuadro E.11: Observación de pregunta 11.

absences	goout	studytime	reason_reputation	failures	Fedu
19	4	1	0	1	2

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $>$ 2.50 **entonces** Reprobado

Figura E.17: Reglas de decisión para la Pregunta 11 del modelo DT-InterpretML.

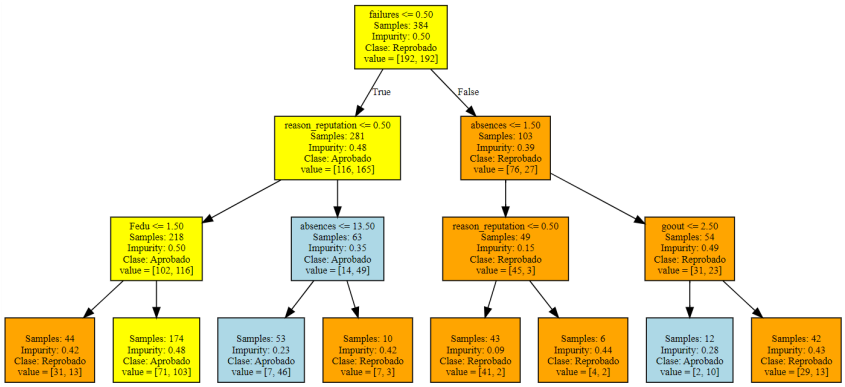


Figura E.18: Grafo local del modelo DT-InterpretML para la Pregunta 11.

Respuestas:

- *DT:* Reprobado
- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.12. Pregunta 12

ID: 12
Categoría: Ambigüedad
Subcategoría: Grado Local
Modelo Evaluado: IDS

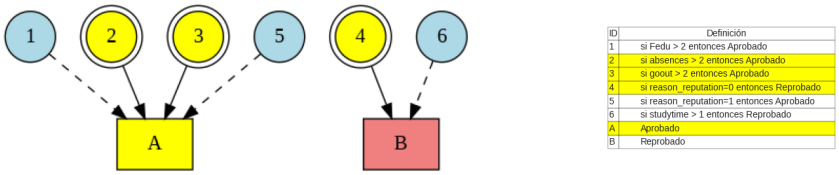
Instrucciones: Basándote en el grafo local y las reglas resaltadas, selecciona si la predicción correcta para esta observación es “Aprobado” o “Reprobado”. Si al analizar las reglas y el grafo local no encuentras una justificación clara, considera que esto puede reflejar ambigüedad en la interpretación local.

Cuadro E.12: Observación de pregunta 12.

absences	goout	studytime	reason_reputation	failures	Fedu
19	4	1	0	1	2

si absences > 2 **entonces** Aprobado
si goout > 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado

Figura E.19: Reglas de decisión para la Pregunta 12 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si goout > 2 entonces Aprobado
4	si reason_reputation=0 entonces Reprobado
5	si reason_reputation=1 entonces Aprobado
6	si studytime > 1 entonces Reprobado
A	Aprobado
B	Reprobado

Figura E.20: Grafo local del modelo IDS para la Pregunta 12.

Respuestas:

- *DT:* Reprobado
- *IDS:* Aprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Aprobado
  - ☐ Reprobado

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.13. Pregunta 13

ID: 13
Categoría: Error
Subcategoría: Reglas
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en la regla resaltada y los valores de la observación, califica si la predicción del modelo es adecuada según las reglas proporcionadas. Si encuentras inconsistencias entre la predicción y las reglas, considera esto al tomar tu decisión.

Cuadro E.13: Observación de pregunta 13.

absences	goout	studytime	reason_reputation	failures	Fedu
6	2	3	1	0	3

si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado

Figura E.21: Regla de decisión para la Pregunta 13 del modelo DT-InterpretML.

Respuestas:

- *DT:* Aprobado
- *IDS:* Aprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.14. Pregunta 14

ID: 14
Categoría: Error
Subcategoría: Reglas
Modelo Evaluado: IDS

Instrucciones: Basándote en la regla resaltada y los valores de la observación, califica si la predicción del modelo es adecuada según las reglas proporcionadas. Si encuentras inconsistencias entre la predicción y las reglas, considera esto al tomar tu decisión.

Cuadro E.14: Observación de pregunta 14.

absences	goout	studytime	reason_reputation	failures	Fedu
6	2	3	1	0	3

si Fedu $>$ 2 **entonces** Aprobado
si absences $>$ 2 **entonces** Aprobado
si reason_reputation = 1 **entonces** Aprobado
si studytime $>$ 1 **entonces** Reprobado

Figura E.22: Reglas de decisión para la Pregunta 14 del modelo IDS.

Respuestas:

- *DT:* Aprobado

- *IDS:* Aprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.15. Pregunta 15

ID: 15
Categoría: Error
Subcategoría: Grado Global
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo global y la regla resaltada, califica si la predicción del modelo para esta observación es adecuada. Si encuentras inconsistencias entre la predicción y la regla, considera esto al tomar tu decisión.

Cuadro E.15: Observación de pregunta 15.

absences	goout	studytime	reason_reputation	failures	Fedu
0	2	2	0	0	1

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $>$ 2.50 **entonces** Reprobado

Figura E.23: Reglas de decisión para la Pregunta 15 del modelo DT-InterpretML.

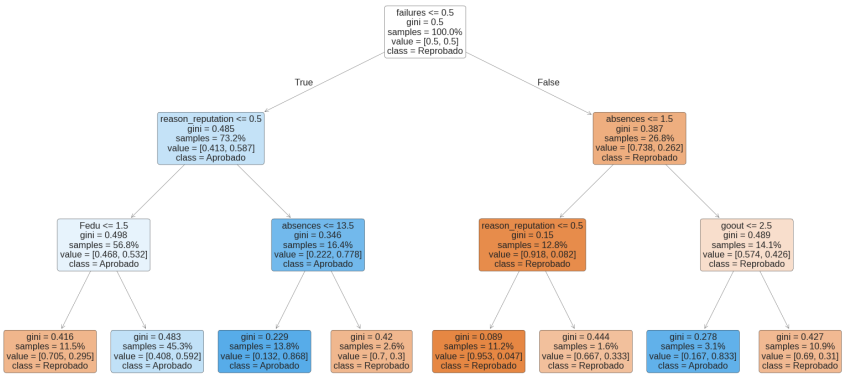


Figura E.24: Grafo global del modelo DT-InterpretML para la Pregunta 15.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.16. Pregunta 16

ID: 16
Categoría: Error
Subcategoría: Grado Global
Modelo Evaluado: IDS

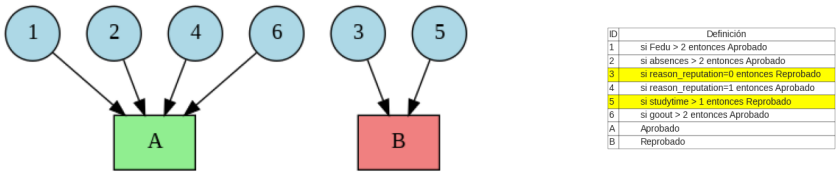
Instrucciones: Basándote en el grafo global y la regla resaltada, califica si la predicción del modelo para esta observación es adecuada. Si encuentras inconsistencias entre la predicción y la regla, considera esto al tomar tu decisión.

Cuadro E.16: Observación de pregunta 16.

absences	goout	studytime	reason_reputation	failures	Fedu
0	2	2	0	0	1

si reason_reputation = 0 **entonces** Reprobado
si studytime > 1 **entonces** Reprobado

Figura E.25: Reglas de decisión para la Pregunta 16 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si reason_reputation=0 entonces Reprobado
4	si reason_reputation=1 entonces Aprobado
5	si studytime > 1 entonces Reprobado
6	si goout > 2 entonces Aprobado
A	Aprobado
B	Reprobado

Figura E.26: Grafo global del modelo IDS para la Pregunta 16.

Respuestas:

- *DT:* Reprobado
- *IDS:* Reprobado
- *Clase Real:* Aprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.17. Pregunta 17

ID: 17
Categoría: Error
Subcategoría: Grado Local
Modelo Evaluado: DT-InterpretML

Instrucciones: Basándote en el grafo local y la regla resaltada, califica si la predicción del modelo para esta observación es adecuada. Si encuentras inconsistencias entre la predicción y la regla, considera esto al tomar tu decisión.

Cuadro E.17: Observación de pregunta 17.

absences	goout	studytime	reason_reputation	failures	Fedu
12	3	2	0	0	4

si failures $\leq$ 0.50 **y** reason_reputation $\leq$ 0.50 **y** Fedu $\leq$ 1.50 **entonces** Reprobado
si failures $\leq$ 0.50 **y** reason_reputation $>$ 0.50 **y** absences $\leq$ 13.50 **entonces** Aprobado
si failures $>$ 0.50 **y** absences $>$ 1.50 **y** goout $>$ 2.50 **entonces** Reprobado

Figura E.27: Reglas de decisión para la Pregunta 17 del modelo DT-InterpretML.

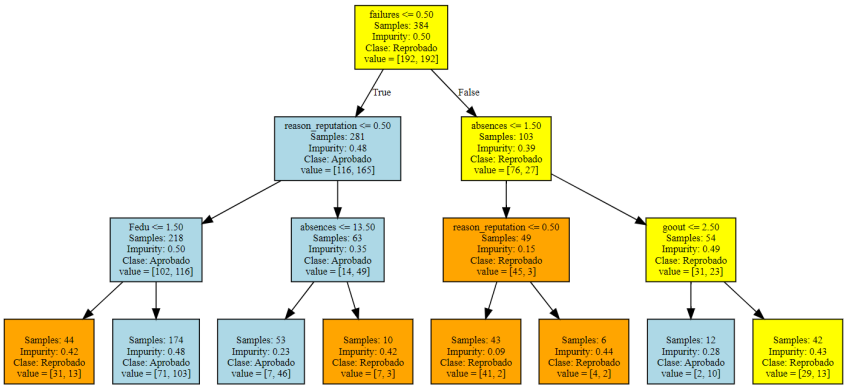


Figura E.28: Grafo local del modelo DT-InterpretML para la Pregunta 17.

Respuestas:

- *DT:* Aprobado
- *IDS:* Aprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.18. Pregunta 18

ID: 18
Categoría: Error
Subcategoría: Grado Local
Modelo Evaluado: IDS

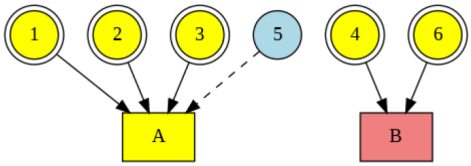
Instrucciones: Basándote en el grafo local y la regla resaltada, califica si la predicción del modelo para esta observación es adecuada. Si encuentras inconsistencias entre la predicción y la regla, considera esto al tomar tu decisión.

Cuadro E.18: Observación de pregunta 18.

absences	goout	studytime	reason_reputation	failures	Fedu
12	3	2	0	0	4

si Fedu > 2 **entonces** Aprobado
si absences > 2 **entonces** Aprobado
si goout > 2 **entonces** Aprobado
si reason_reputation = 0 **entonces** Reprobado
si studytime > 1 **entonces** Reprobado

Figura E.29: Reglas de decisión para la Pregunta 18 del modelo IDS.



ID	Definición
1	si Fedu > 2 entonces Aprobado
2	si absences > 2 entonces Aprobado
3	si goout > 2 entonces Aprobado
4	si reason_reputation=0 entonces Reprobado
5	si reason_reputation=1 entonces Aprobado
6	si studytime > 1 entonces Reprobado
A	Aprobado
B	Reprobado

Figura E.30: Grafo local del modelo IDS para la Pregunta 18.

Respuestas:

- *DT:* Aprobado
- *IDS:* Aprobado
- *Clase Real:* Reprobado
- *Usuario:*
  - ☐ Correcto
  - ☐ Incorrecto

Pregunta de seguimiento: ¿Qué tan seguro(a) estás de tu respuesta?

- ☐ Mucho
- ☐ Poco
- ☐ Nada

E.19. Pregunta 19

ID: 19
Categoría: Preferencias de Visualización

Instrucciones: ¿Cuál de los siguientes grafos encontraste más útil para entender el funcionamiento del modelo?

Opciones:

- ☐ Árbol de decisión (InterpretML)
- ☐ Conjuntos de Decisiones interpretables (IDS)

E.20. Pregunta 20

ID: 20
Categoría: Preferencias de Visualización

Instrucciones: ¿Qué modelo te facilitó comprender la predicción y analizar posibles errores?

Opciones:

- ☐ Árbol de decisión (InterpretML)
- ☐ Conjuntos de Decisiones interpretables (IDS)

E.21. Pregunta 21

ID: 21
Categoría: Pregunta Descriptiva

Instrucciones: ¿Crees que la visualización del grafo y sus reglas debería siempre acompañar las predicciones para mejorar la comprensión?

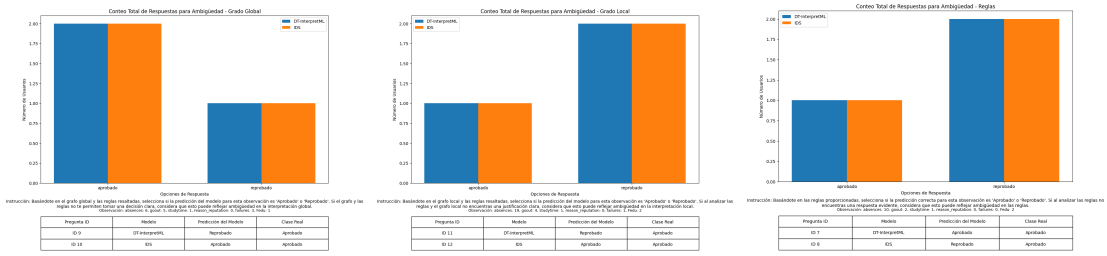
Respuesta del Usuario:

Anexo F

Resultados de los Usuarios de Prueba

En esta sección se presentan las gráficas generadas por los usuarios de prueba al completar el cuestionario implementado en la herramienta web *Survey-XAI-App*.

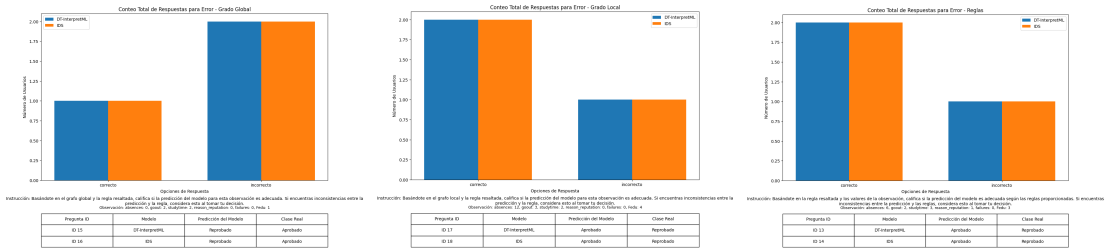
F.1. Ambigüedad



(a) Grado Global (b) Grado Local (c) Reglas

Figura F.1: Distribución total para la categoría Ambigüedad.

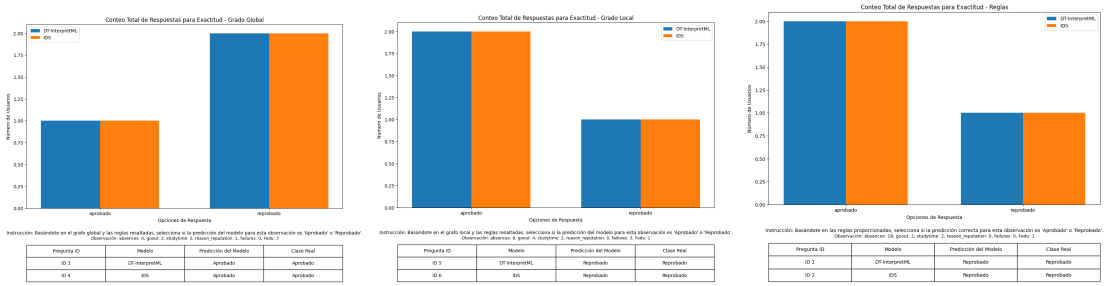
F.2. Error



(a) Grado Global (b) Grado Local (c) Reglas

Figura F.2: Distribución total para la categoría Error.

F.3. Exactitud

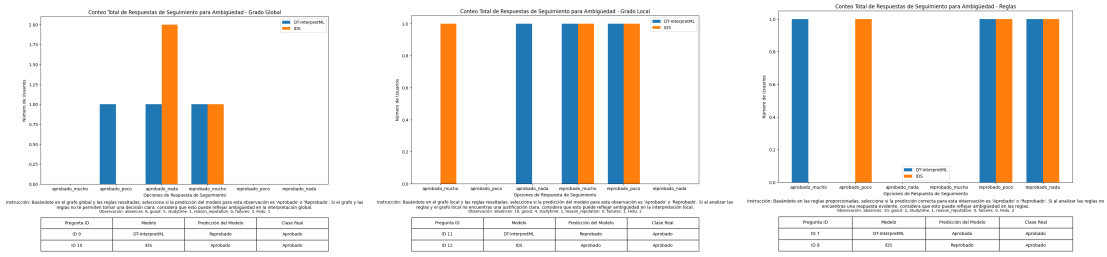


(a) Grado Global (b) Grado Local (c) Reglas

Figura F.3: Distribución total para la categoría Exactitud.

F.4. Preguntas de Seguimiento

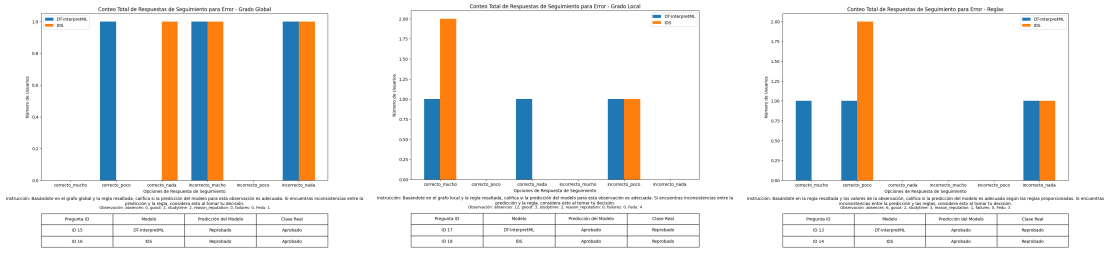
F.4.1. Ambigüedad



(a) Grado Global (b) Grado Local (c) Reglas

Figura F.4: Seguimiento para la categoría Ambigüedad.

F.4.2. Error



(a) Grado Global (b) Grado Local (c) Reglas

Figura F.5: Seguimiento para la categoría Error.