
Facultad de Ciencias Económicas y Empresariales.

# Sesgos no instruccionales en encuestas de evaluación del profesorado

Karen Camacho Vilacoba

5º E6 + Analytics

Director: José Luis Arroyo Barriguete.

Madrid, 2025-26

## Resumen

Las evaluaciones de la enseñanza por parte de los estudiantes (de ahora en adelante SET, por sus siglas en inglés) son ampliamente utilizadas en la educación superior y, con frecuencia, influyen en decisiones administrativas importantes. Sin embargo, un número cada vez mayor de investigaciones indica que los resultados de las SET pueden verse sistemáticamente influenciados por factores no relacionados con la enseñanza (sesgos no instruccionales). Este estudio se centra en la dificultad percibida de los cursos como fuente clave de posible sesgo y examina si su influencia se limita a las calificaciones medias o si también afecta a la estructura interna de las distribuciones de las evaluaciones.

El análisis se basa en el conjunto de datos públicos de RateMyProfessor.com (He, 2020) y adopta un enfoque empírico multinivel (a nivel de estudiantes y profesores) y multimétodo. En primer lugar, se realizan análisis descriptivos y correlacionales para caracterizar las principales variables y establecer la relación de referencia entre la dificultad percibida y las calificaciones de los estudiantes en ambos niveles de agregación. Seguidamente, se aplica el método de *clustering* K-means a las medidas estandarizadas de la dificultad percibida (diff_index) y las calificaciones generales de los profesores (star_rating) para identificar perfiles de evaluación distintos entre los docentes. En una fase posterior del análisis, se examina la relación entre la dificultad percibida y la estructura distributiva de las evaluaciones estudiantiles a nivel docente. Para ello, se estiman modelos de regresión por mínimos cuadrados ordinarios (OLS) que analizan su asociación con la variabilidad de las valoraciones (desviación estándar), su asimetría y su curtosis. En todos los modelos se controla explícitamente por la valoración media del profesor y por el número de reseñas recibidas. Como complemento al enfoque lineal y con el fin de explorar posibles relaciones no lineales entre las variables, se utiliza adicionalmente una red neuronal de tipo perceptrón multicapa (MLP), cuya interpretación se apoya en un análisis de sensibilidad.

En cuanto a los resultados, los análisis exploratorios ponen de manifiesto una asociación negativa clara entre la dificultad percibida y las valoraciones concedidas por los estudiantes, tanto a nivel individual como a nivel agregado por profesor. El análisis de clústeres identifica patrones estructurados consistentes con la existencia de dos perfiles

docentes diferenciados: por un lado, profesores percibidos como menos exigentes, que reciben calificaciones medias más elevadas y presentan una mayor homogeneidad en las evaluaciones; y, por otro, profesores asociados a una mayor dificultad percibida, con valoraciones medias más bajas y una mayor dispersión en las respuestas de los estudiantes.

En conjunto, los resultados aportan evidencia empírica de que la dificultad percibida actúa como un sesgo no instruccional en las SET, con efectos que van más allá de las puntuaciones medias y alcanzan también a las propiedades distributivas de los datos de evaluación. Ello refuerza la importancia de interpretar los resultados de las SET con cautela e invitan a adoptar enfoques analíticos más matizados en su utilización para la toma de decisiones.

## *Palabras Clave*

Evaluación docente; Encuestas de evaluación del profesorado; sesgos no instruccionales; dificultad percibida; análisis multinivel; clustering; regresión; aprendizaje automático.

## Abstract

Student evaluations of teaching (SET) are widely used in higher education and often influence important administrative decisions. However, a growing body of research indicates that SET results may be systematically influenced by factors unrelated to teaching (non-instructional biases). This study focuses on the perceived difficulty of courses as a key source of potential bias and examines whether its influence is limited to average ratings or whether it also affects the internal structure of evaluation distributions.

The analysis is based on the public dataset from RateMyProfessor.com (He, 2020) and adopts a multilevel (student and professor level) and multimodal empirical approach. First, descriptive and correlational analyses are performed to characterize the main variables and establish the reference relationship between perceived difficulty and student grades at both levels of aggregation. Next, the K-means clustering method is applied to standardized measures of perceived difficulty (diff_index) and overall professor ratings (star_rating) to identify distinct evaluation profiles among professors.

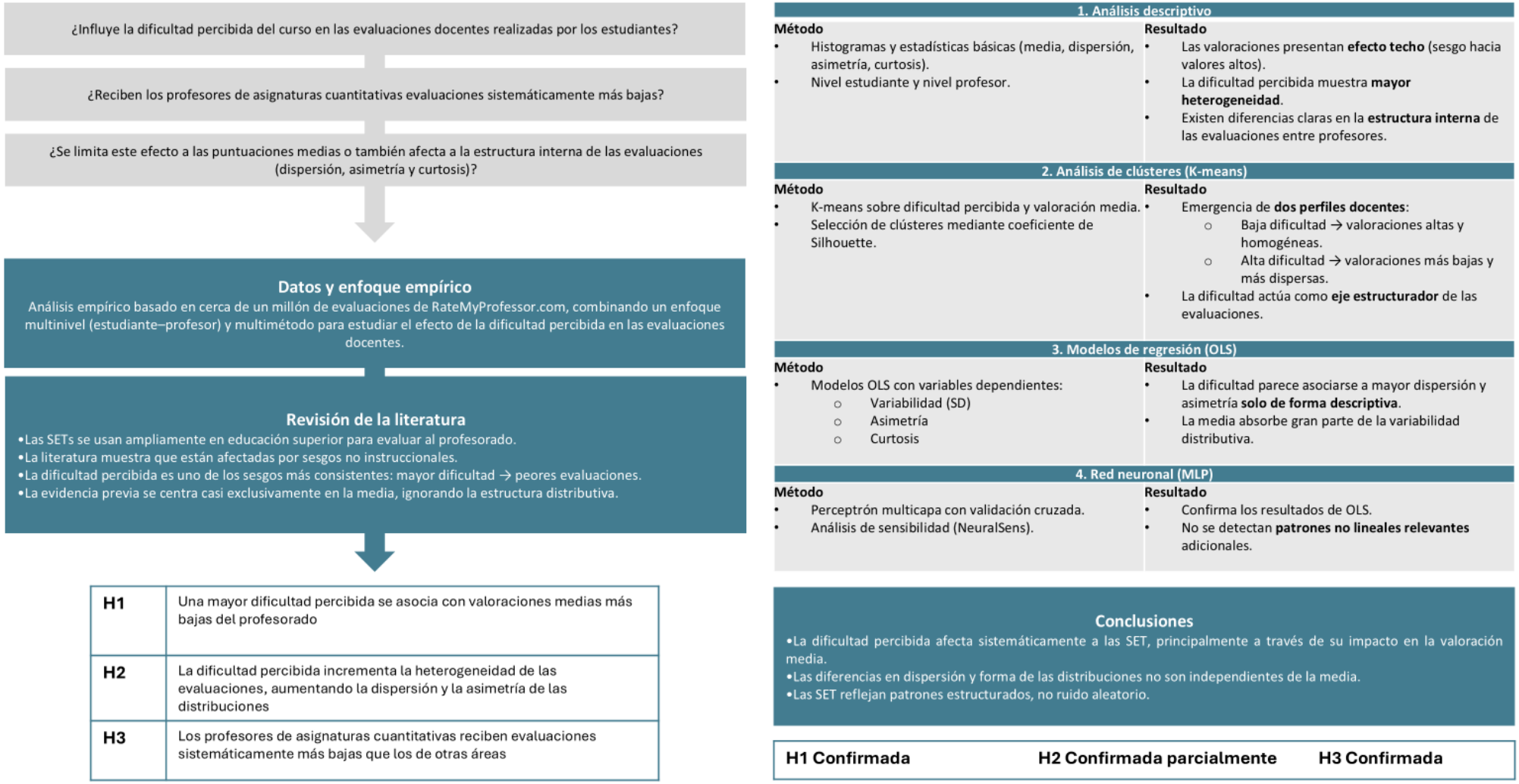
In a later phase of the analysis, the relationship between perceived difficulty and the distributional structure of student evaluations at the teacher level is examined. To do this, ordinary least squares (OLS) regression models are estimated to analyze their association with the variability of the ratings (standard deviation), their asymmetry, and their kurtosis. All models explicitly control the teacher's average rating and the number of reviews received. To complement the linear approach and explore possible nonlinear relationships between variables, a multilayer perceptron (MLP) neural network is also used, whose interpretation is supported by a sensitivity analysis.

In terms of results, exploratory analyses reveal a clear negative association between perceived difficulty and the ratings given by students, both individually and collectively per teacher. Cluster analysis identifies structured patterns consistent with the existence of two distinct teaching profiles: on the one hand, teachers perceived as less demanding, who receive higher average ratings and show greater homogeneity in evaluations; and, on the other hand, teachers associated with greater perceived difficulty, with lower average ratings and greater dispersion in student responses. Taken together, the results provide empirical evidence that perceived difficulty acts as a non-instructional bias in SETs, with effects that go beyond average scores and affect the distributional properties of assessment data. This reinforces the importance of interpreting SET results with caution and calls for more nuanced analytical approaches in their use for decision-making.

## *Key Words*

Student evaluations of teaching; non-instructional biases; perceived difficulty; multilevel analysis; clustering; regression; machine learning.

# Graphical abstract



# Contenido

# 1 Introducción

Las evaluaciones de la enseñanza por parte de los estudiantes (Student Evaluation of Teaching - SET) se utilizan de manera generalizada en las instituciones de educación superior como instrumentos para evaluar la calidad de la enseñanza. La adopción extendida de estas herramientas se debe en gran medida a las ventajas que presentan. Primeramente, son fácilmente implementables a un coste relativamente bajo. Además, cuentan con la legitimidad percibida de los estudiantes como participantes directos en el proceso de enseñanza-aprendizaje. Por ello, en la práctica, los resultados de las SET se emplean con frecuencia no solo con fines formativos, como la mejora de los cursos, sino también para tomar decisiones administrativas de gran importancia, como la renovación de contratos, los ascensos, la titularidad y los incentivos basados en el rendimiento. A pesar de su relevancia institucional y uso generalizado, las SET son una herramienta muy controvertida debido a la creciente evidencia que demuestra que están influenciadas por factores ajenos a la eficacia real de la enseñanza.

En este sentido, uno de los retos fundamentales en la interpretación de las SET radica en su susceptibilidad a estas influencias ajenas a la enseñanza, o sesgos no instruccionales. Las percepciones de los estudiantes sobre los cursos y los profesores no solo están determinadas por las prácticas docentes, sino también por elementos contextuales y subjetivos, como la carga de trabajo y la dificultad percibida del curso. Estas influencias pueden introducir distorsiones sistemáticas en los resultados de la evaluación, lo que puede resultar perjudicial para determinados profesores o disciplinas. Por ello, es relevante comprender cómo afectan estos sesgos a la enseñanza para evaluar la validez y explorar un uso adecuado de las SET.

Partiendo de los sesgos mencionados previamente, la dificultad percibida ocupa un lugar especialmente relevante por su complejidad de interpretación. La dificultad es una característica intrínseca de muchos cursos y disciplinas, pero, al mismo tiempo, una percepción subjetiva de los estudiantes. Esta doble naturaleza la convierte en una fuente potencialmente importante de sesgo en las evaluaciones de los estudiantes. En este sentido, es importante destacar que el efecto de la dificultad puede ir más allá de las calificaciones medias e influir en la estructura interna de las distribuciones de las evaluaciones, como la presencia de evaluaciones extremas. La mayoría de los análisis

empíricos de las SET, que se explorarán más adelante, se han centrado principalmente en las puntuaciones medias de las evaluaciones. Sin embargo, los datos de las SET se caracterizan a menudo por distribuciones sesgadas y respuestas heterogéneas de los estudiantes, lo que sugiere que las características de la distribución pueden contener información relevante que se pierde cuando la atención se limita únicamente a las medias. Por tanto, un análisis más completo de las SET requiere examinar no solo las tendencias centrales, sino también la variabilidad y las propiedades relacionadas con la forma de las distribuciones.

En este contexto, el presente estudio investiga el papel de la dificultad percibida como sesgo no instruccional que influye en las evaluaciones de los estudiantes. El análisis adopta un enfoque multinivel y multi método, que combina el análisis descriptivo, las métricas de distribución, las técnicas de *clustering* y la modelización. Al centrarse tanto en el nivel medio como en la estructura interna de las evaluaciones (dispersión, asimetría y curtosis), este estudio pretende ofrecer una perspectiva empírica más matizada sobre cómo la dificultad percibida influye en los resultados de las SET y contribuir a los debates en curso sobre su interpretación y uso en la educación superior.

### *1.1 Objetivos de la investigación*

El objetivo principal de este estudio es analizar la existencia y la magnitud de los sesgos no instruccionales en las evaluaciones docentes universitarias, utilizando datos a gran escala de RateMyProfessor.com.

Más concretamente, la investigación persigue los siguientes objetivos:

- Examinar la relación entre la dificultad percibida del curso y las evaluaciones de los estudiantes, analizando si los profesores asociados a cursos más exigentes tienden a recibir puntuaciones de evaluación más bajas tanto a escala individual (a nivel de estudiante) como agregada (a nivel de profesor).
- Investigar si la dificultad percibida afecta no solo a las puntuaciones medias de las evaluaciones, sino también a la estructura interna de las distribuciones de las evaluaciones, analizando las medidas de variabilidad, asimetría y curtosis en las calificaciones de los estudiantes a nivel de profesor.
- Identificar si los profesores que imparten asignaturas con un alto contenido cuantitativo, en particular, aquellas relacionadas con las matemáticas, reciben sistemáticamente evaluaciones más bajas en comparación con los profesores de otras áreas académicas, explorando así la presencia de sesgos de evaluación relacionados con la disciplina.
- Explorar la existencia de perfiles de evaluación distintos entre los profesores, basándose en la distribución conjunta de la dificultad percibida y las calificaciones generales, utilizando técnicas de agrupamiento para identificar grupos de profesores con patrones de evaluación similares.

Al abordar estos objetivos, el estudio pretende ofrecer una caracterización empírica más completa de cómo la dificultad percibida opera como factor no docente en las evaluaciones de la enseñanza por parte de los estudiantes y aportar pruebas relevantes para la interpretación y el uso responsable de las SET en contextos de educación superior.

# 2 Revisión de la literatura

## *2.1 Las evaluaciones del profesorado por parte de los estudiantes en la Educación Superior*

Entre los instrumentos más utilizados para evaluar la calidad de la enseñanza en las instituciones de educación superior, se encuentran las evaluaciones de la docencia por parte de los estudiantes (de ahora en adelante SET, por sus siglas en inglés). En términos generales, consisten en cuestionarios estandarizados en los que los estudiantes califican diferentes aspectos de un curso y de su profesor, usualmente al final del período académico. Con el tiempo, las SET han cobrado importancia en los sistemas de garantía de calidad (Marsh, 2007). Además, en la última década se han utilizado frecuentemente para tomar decisiones relacionadas con la promoción del profesorado, la titularidad, la renovación de contratos y los incentivos basados en el rendimiento (Zhao et al., 2022).

A pesar de que las evaluaciones de los estudiantes para calificar la enseñanza han cobrado relevancia en las últimas décadas, sus orígenes se remontan a principios del siglo pasado. Las primeras formas de instrumentos de evaluación de los estudiantes ya se utilizaban en las universidades estadounidenses a principios del siglo XX, y uno de los primeros cuestionarios estandarizados se elaboró en la Universidad de Purdue en 1915 (Zhao et al., 2022). Desde entonces, las SET han sido adoptadas de manera progresiva por instituciones de educación superior de Estados Unidos y Europa y, más recientemente, por las universidades de todo el mundo, convirtiéndose en un elemento habitual de las prácticas de evaluación institucional (Marsh, 2007).

En el contexto de la educación superior actual, las SET se realizan a través de encuestas anónimas, habitualmente *online*. La metodología, en líneas generales, es la siguiente: se pide a los estudiantes que respondan a preguntas que abarcan dimensiones como la calidad general de la enseñanza, la claridad de las explicaciones, la organización del curso, los métodos de evaluación, la carga de trabajo y la dificultad percibida. Finalmente, en muchas instituciones, se presenta una pregunta global que recoge la evaluación general del profesor por parte del estudiante. Esta juega un papel destacado, ya que se suele utilizar como indicador resumido del rendimiento docente y es fácilmente comparable entre cursos y profesores (Zhang et al., 2020). Las preguntas se

suelen formular de manera que el estudiante seleccione una respuesta en una escala Likert., o bien en una escala numérica.

La adopción generalizada de las SET se fundamenta en la perspectiva de que los estudiantes, como participantes directos en el proceso de enseñanza-aprendizaje, resultan idóneos para evaluar la calidad de la enseñanza. Esta visión considera el *feedback* de los estudiantes como una valiosa fuente de información para mejorar la enseñanza y aumentar la responsabilidad docente. Por otro lado, y desde el punto de vista institucional, las SET se consideran también convenientes, ya que su implementación es relativamente económica y permiten la recopilación de datos estandarizados a gran escala en diversos contextos académicos (Spooren et al., 2013).

A pesar de estas ventajas, la validez de las SET como medidas de la calidad y eficacia docentes ha sido objeto de un amplio debate. Numerosos estudios han examinado si las respuestas de los estudiantes reflejan con precisión la calidad de la enseñanza o los resultados del aprendizaje. Por un lado, algunos señalan la existencia de niveles aceptables de fiabilidad y consistencia interna, mientras que otros destacan que las SET son muy sensibles a varios factores no relacionados con la eficacia docente real (Boring et al., 2016). Estas preocupaciones o visiones han dado lugar a un creciente escepticismo respecto al uso acrítico de las evaluaciones de los estudiantes en procesos de toma de decisiones en instituciones de educación superior (Uttl et al., 2017).

### *2.2 Estructura de las SETs*

Las evaluaciones de los docentes por parte de los estudiantes están estructuradas con datos multidimensionales y heterogéneos. En líneas generales, las SETs consisten en calificaciones ordinales proporcionadas por los estudiantes, que se agregan a nivel de profesor o de curso. Una característica clave de los datos de las SETs es la asimetría en la forma en que los estudiantes utilizan las escalas de evaluación. Las pruebas empíricas muestran que las calificaciones tienden a concentrarse en la parte superior de la escala, lo que da lugar a distribuciones sesgadas con un uso limitado de los valores más bajos (Boring et al., 2016). Este “efecto techo” implica que la variación entre los profesores a menudo se limita en un rango poco amplio de puntuaciones altas, de manera que se dificulta la interpretación de las diferencias en las medias. Por ello, en el análisis de los SETs aumenta la relevancia de las propiedades de distribución, como la dispersión, la

asimetría y la curtosis y en consecuencia, los profesores con puntuaciones medias similares pueden diferir significativamente en el grado de consenso o polarización entre los estudiantes.

Adicionalmente, los datos SET suelen estar anidados, es decir, las calificaciones individuales de los estudiantes se agrupan por profesores, que a su vez pertenecen a departamentos e instituciones. Esta estructura jerárquica implica que los resultados de la evaluación no solo están determinados por las percepciones de los estudiantes a nivel individual; sino que también por factores contextuales a niveles de agregación superiores. Sobre esta estructura jerárquica, Uttl et al. (2017) señalan que ignorar esta naturaleza multinivel puede llevar a conclusiones erróneas sobre la eficacia de la enseñanza, ya que la variabilidad entre los profesores suele ser tan informativa como las diferencias entre ellos.

### *2.3 Sesgos no instruccionales en las evaluaciones de los estudiantes*

Como mencionamos en el apartado previo, existe una percepción creciente acerca de las SET como evaluaciones subjetivas moldeadas por las percepciones y experiencias de los estudiantes. Este cambio de perspectiva ha motivado una amplia literatura centrada en la identificación de sesgos sistemáticos en las evaluaciones de los estudiantes, no relacionados con la calidad real de la enseñanza. Estos factores denominados “no docentes” atienden a características del curso, del profesor o del estudiante que influyen en las evaluaciones independientemente de la eficacia de la enseñanza. Algunos estudios indican que estos factores pueden distorsionar significativamente las calificaciones de los estudiantes, lo que suscita dudas sobre la validez de las SET como indicadores del rendimiento pedagógico (Spooren et al., 2007).

Uno de los sesgos no docentes más estudiados se relaciona con los resultados de las evaluaciones y las calificaciones recibidas o esperadas. Las investigaciones empíricas encuentran sistemáticamente una asociación positiva entre las calificaciones que reciben los estudiantes, o esperan recibir, y las evaluaciones que asignan a los profesores (Isely y Singh, 2005). Esta relación se ha interpretado como una prueba de que las SET reflejan en parte la satisfacción de los estudiantes más que la eficacia de la enseñanza, lo que crea incentivos para la “inflación” de las calificaciones y una menor exigencia académica. En este contexto, los profesores se pueden ver “recompensados” (en este

contexto, recibir evaluaciones más altas) implícitamente por prácticas de calificación indulgentes, mientras que aquellos que mantienen estándares más altos corren el riesgo de recibir evaluaciones más bajas a pesar de fomentar un aprendizaje más profundo (Schneider, 2013).

Otro factor que puede interferir en las evaluaciones es la característica del curso impartido. Los cursos percibidos como más "difíciles", en general, aquellos cuantitativamente intensivos u obligatorios, tienden a recibir calificaciones más bajas que los cursos optativos o menos exigentes (Ali & Al Ajmi, 2013). En particular, se ha demostrado que disciplinas como las matemáticas, la ingeniería y las ciencias naturales reciben evaluaciones medias más bajas en comparación con las humanidades y las ciencias sociales, incluso después de controlar otros factores (Uttl et al., 2017). Este patrón sugiere la existencia de un sesgo relacionado con la dificultad, por el que los estudiantes penalizan a los profesores por las características del curso que son intrínsecas a la materia y no reflejan la calidad de la enseñanza.

Finalmente, los factores relacionados con los estudiantes pueden constituir un origen adicional de sesgo. Aspectos como la motivación, el interés previo por la materia, la capacidad académica y las actitudes hacia los procesos de evaluación pueden influir en la forma en que los estudiantes evalúan a sus profesores (Spooren et al., 2007). En este sentido, los estudiantes con un rendimiento inferior o menor motivación, son más propensos a expresar su insatisfacción, lo que puede reflejarse en evaluaciones más bajas. Además, se ha demostrado que los aspectos contextuales del proceso de evaluación, como el momento (por ejemplo, antes o después de los exámenes), el anonimato y el propósito declarado de las evaluaciones, pueden afectar de manera negativa a las evaluaciones (Young et al., 1999).

En resumen, la literatura examinada sobre las SET señala que los resultados pueden venir determinados por una compleja interacción de factores educativos y no educativos. Es importante destacar que, para el presente estudio, las pruebas consistentes sobre los sesgos no docentes proporcionan una sólida base teórica para examinar no solo cómo factores, como la dificultad percibida, afectan a las calificaciones medias, sino también cómo influyen en la distribución global de las evaluaciones (su variabilidad, asimetría y extremidad) entre los profesores.

### *2.4 Estudios empíricos como herramienta de investigación de las SET*

Dentro de este marco estructural, la investigación empírica se ha fundamentado cada vez más en conjuntos de datos observacionales a gran escala y métodos cuantitativos para examinar los factores que juegan un rol clave en los resultados de la evaluación de los estudiantes. La mayoría de los estudios empíricos se basan en conjuntos de datos administrativos o basados en plataformas que contienen miles de evaluaciones de profesores, cursos y diferentes instituciones. Se aplican métodos, principalmente apoyados en la regresión lineal, para evaluar la relación entre las puntuaciones de las SET y los factores tanto relativos a la docencia como los posibles sesgos indicados anteriormente. Para su análisis, estos estudios consideran las características observables, como el nivel del curso o el contexto institucional; y examinan variables que pueden resultar factores de sesgo como la dificultad percibida o la carga de trabajo. Más allá del modelado lineal, algunos análisis empíricos recientes han tratado de captar la naturaleza multidimensional de los datos SET incorporando técnicas de aprendizaje no supervisado. Por ejemplo, Liu (2022) aplica métodos de agrupamiento a los datos de evaluación para identificar patrones ocultos basados en combinaciones de la calidad docente percibida y las exigencias del curso. Estos enfoques son especialmente adecuados para los datos SET, en los que las distribuciones suelen ser sesgadas, heterogéneas y moldeadas por las interacciones entre múltiples dimensiones perceptivas.

Los resultados empíricos de ambos enfoques en los diferentes estudios convergen en una conclusión común: la dificultad percibida desempeña un papel fundamental en la configuración de los resultados de la evaluación. Los profesores que imparten cursos más exigentes académicamente reciben calificaciones medias más bajas y una mayor dispersión interna; mientras que los cursos menos exigentes tienden a agruparse en torno a evaluaciones más altas y homogéneas. Estos resultados sugieren que los resultados de las SET no se distribuyen de forma continua o aleatoria, sino que se organizan en patrones estructurados en torno a la dificultad del curso.

Sin embargo, aunque los estudios anteriormente mencionados evidencian la existencia de tales patrones, se centran en las diferencias medias o en la pertenencia a grupos, dejando abierta la cuestión de cómo la dificultad afecta a la estructura interna de las

distribuciones de las evaluaciones. En este sentido, no hay análisis exhaustivos dedicados a observar si la dificultad percibida influye en la variabilidad, la asimetría y la curtosis en las evaluaciones de los estudiantes, y de qué manera. El presente estudio combina el análisis descriptivo, las técnicas de *clustering* y los enfoques basados en la regresión y el aprendizaje automático, para aportar nuevas evidencias sobre el papel de la dificultad percibida como factor estructural en las SET. De esta manera, se propone una perspectiva nueva sobre el sesgo no docente y sus implicaciones para la interpretación y el uso de las SET en la toma de decisiones académicas.

# 3 Material y Métodos

## *3.1 Datos y procedimiento*

El presente trabajo se ha basado en el conjunto de datos "Big Data Set from RateMyProfessor.com for Professors' Teaching Evaluation" en *Mendeley Data* (He, 2020). Este conjunto, disponible públicamente para fines académicos, recopila información sobre las evaluaciones docentes publicadas en la plataforma *RateMyProfessor.com*. El dataset original, empleado en varios artículos de investigación, incluye cerca de un millón de registros de profesores universitarios de distintas instituciones en Estados Unidos y está disponible en dos formatos: uno compuesto por múltiples archivos .csv individuales por profesor y otro consolidado que agrupa la información de varios docentes en un único archivo. Este último ha sido el utilizado en el presente trabajo.

Este dataset contiene 18 variables relacionadas con la evaluación del profesorado universitario, entre ellas: nombre del profesor, universidad, departamento, estado, año desde la primera reseña, valoración global (*star rating*), índice de dificultad (*difficulty index*), porcentaje de estudiantes que repetirían con el profesor (*take again*), etiquetas descriptivas, asistencia obligatoria, curso con o sin créditos, calificación del estudiante (*grade*), y comentarios textuales.

El primer objetivo del presente estudio es analizar de qué manera la dificultad percibida de la materia, entendida como un sesgo no instruccional externo al desempeño del profesor, influye en la variabilidad y forma de las valoraciones docentes. El segundo objetivo es determinar si existe un sesgo negativo hacia las materias cuantitativas,

siendo los profesores de dichas materias peor evaluados que el resto de profesores. Atendiendo a dichos objetivos, se han seleccionado las siguientes variables principales:

| Tipo de variable | Variable | Descripción |
|---|---|---|
| Dependiente | *student_star* | Valoración individual del profesor por parte del estudiante. |
| Dependiente | *sd(student_star)* | Desviación estándar de las valoraciones del profesor. Esta variable, no disponible en el dataset original, ha sido calculada para el trabajo. |
| Dependiente | *skewness(student_star)* | Asimetría de la distribución de las valoraciones. Esta variable, no disponible en el dataset original, ha sido calculada para el trabajo. |
| Dependiente | *kurtosis(student_star)* | Curtosis de la distribución de las valoraciones. Esta variable, no disponible en el dataset original, ha sido calculada para el trabajo. |
| Explicativa | *diff_index* | Índice de dificultad percibida del profesor. |
| Explicativa | *star_rating* | Valoración global media del profesor. Se calcula como la media de la variable "*student_star"* |
| Control | *n_reviews* | Número total de reseñas recibidas por el profesor. |
| Categóricas | *department_name* | Variables institucionales. |

*Tabla 1: Descripción de las variables. Elaborada por la autora.*

La metodología para el estudio se ha estructurado en cuatro fases:

1. Limpieza y preprocesamiento del conjunto de datos.
2. Análisis descriptivo y correlacional.
3. Agrupamiento de profesores mediante técnicas de *clustering*.
4. Modelización estadística y de aprendizaje automático.

### *3.2 Limpieza y preprocesamiento de datos*

Para la fase de preprocesamiento de datos se llevó a cabo la transformación de variables según su tipología y la agregación de datos a nivel de profesor. En cuanto a la transformación de variables, las de naturaleza cualitativa (department_name, local_name, state_name, tag_professor, grades, for_credits y attence) fueron convertidas a formato factor para su reconocimiento como categorías en los modelos. Por su parte, las variables cuantitativas (take_again, star_rating, diff_index, student_star y student_difficult) se transformaron a tipo numérico para garantizar su adecuada

interpretación en los cálculos. Adicionalmente, se elaboró un conjunto agregado a nivel de profesor (datos_unicos) que contiene un único registro por profesor para facilitar análisis posteriores a nivel individual, promediando el resto de variables.

Una vez transformadas las variables, se procedió a la detección y tratamiento de valores ausentes. En primer lugar, se contaron los valores faltantes (NAs) por variable para evaluar la magnitud del problema. Dado que se identificaron datos ausentes principalmente en student_star y student_difficult, se optó por aplicar el algoritmo de vecinos más cercanos (K-Nearest Neighbors, KNN), técnica especialmente recomendada para datasets multivariantes cuando se busca preservar la estructura original de los datos (Troyanskaya et al., 2001). En este caso, se llevó a cabo la imputación en dos pasos: primero se imputaron los valores faltantes de student_star utilizando como variables de distancia diff_index y student_difficult, y posteriormente se imputaron los valores faltantes de student_difficult empleando diff_index y student_star. Para ambos procesos se utilizó k = 5, ofreciendo un equilibrio entre estabilidad y proximidad a los datos, especialmente en este contexto con muchos registros y pocos valores faltantes. Asimismo, las variables empleadas para calcular las distancias se seleccionaron por su relación directa con la percepción de dificultad y la valoración del profesor, asegurando que la imputación se basara en información relevante para el fenómeno analizado. Este método estima los valores faltantes en función de observaciones con características similares, de manera que se minimiza la pérdida de información. Tras la imputación de los NAs, se verificó la coherencia interna de las variables mediante análisis descriptivos básicos, asegurando que las transformaciones no alteraran la distribución original de los datos. Todo el procesamiento de datos se realizó en lenguaje R (RStudio 2025.09.0 Build 387), utilizando principalmente los paquetes dplyr (Wickham et al., 2023) para la manipulación de datos, VIM (Kowarik & Templ, 2016) para la imputación de valores ausentes, car (Fox & Weisberg, 2019) para diagnósticos estadísticos, y caret (Kuhn, 2022) para el preprocesamiento y validación de modelos.

### *3.3 Agrupación y estructura de los datos*

A partir del conjunto de datos preprocesado y de la versión agregada (datos_unicos), el análisis se ha desarrollado en dos niveles complementarios con el fin de capturar tanto

las percepciones individuales de los estudiantes como los patrones generales a nivel docente.

Nivel estudiante: en este caso cada registro del conjunto de datos representa una valoración individual que un estudiante realiza sobre un profesor, expresada en la variable student_star. Este nivel permite analizar la variabilidad y los posibles sesgos perceptivos a escala micro, es decir, cómo difieren las opiniones entre estudiantes dentro de una misma asignatura o profesor.

Nivel docente: en este nivel, las observaciones se han agregado por profesor, generando un único registro por docente. A partir de las valoraciones individuales (student_star), se calcularon las medidas resumen de media, desviación estándar, asimetría y curtosis. Estas métricas permiten analizar la estructura general de las percepciones hacia cada profesor y detectar patrones, ya sea de consenso o polarización en las evaluaciones.

Esta aproximación combinada de los niveles estudiante y profesor, sigue la recomendación de Uttl et al. (2017), quienes destacan la importancia de analizar simultáneamente la variabilidad individual y los patrones agregados para una interpretación más precisa de los sesgos no instruccionales en las encuestas de evaluación docente.

### *3.4 Análisis de clústeres*

El análisis de clústeres se ha desarrollado con el objetivo de identificar grupos homogéneos de profesores según la dificultad percibida (diff_index) y la valoración global obtenida (star_rating). A partir de este procedimiento se observa si existen perfiles docentes diferenciados, como puede ser el de profesores exigentes, pero bien valorados frente a otros más accesibles con valoraciones inferiores (según la combinación de *diff_index* y *star_rating)*. Para ello, se aplicó el algoritmo K-means (MacQueen, 1967) mediante los paquetes cluster y factoextra en el entorno R (R Core Team, 2025). Cabe mencionar que, antes de ejecutar el algoritmo, las variables se estandarizaron (media 0, desviación estándar 1) con el fin de evitar que diferencias en la escala de medida alterasen el cálculo de las distancias euclídeas, sobre las que se basa K-means.

La determinación del número óptimo de clústeres (k) se realizó mediante el coeficiente de Silhouette (Rousseeuw, 1987), que evalúa simultáneamente la cohesión interna de cada grupo y la separación entre clústeres. Se evaluaron soluciones con valores de k entre 2 y 6, seleccionándose aquella que maximizó la Silhouette media. Una vez determinado el número óptimo, se ejecutó el algoritmo K-means con 25 inicializaciones distintas (nstart = 25) para asegurar la estabilidad de la solución y evitar mínimos locales.

### *3.5 Modelización estadística y aprendizaje automático*

Para analizar la relación entre la dificultad percibida y la distribución de las valoraciones docentes, se desarrollaron dos tipos de modelos complementarios: un modelo de regresión lineal múltiple (OLS) y un modelo de red neuronal (MLP). Ambos se implementaron en R (R Core Team, 2025) utilizando funciones básicas del entorno y los paquetes lmtest (Zeileis & Hothorn, 2002) para diagnósticos de especificación, car (Fox & Weisberg, 2019) para análisis de regresión, caret (Kuhn, 2022) para validación cruzada, y NeuralSens (Portela et al., 2021) para análisis post-hoc del modelo MLP.

#### Construcción de las bases de datos y especificación de variables

Se construyeron tres bases de datos a nivel profesor, cada una con una variable dependiente diferente que caracteriza la distribución de las valoraciones recibidas: (1) la variabilidad de las valoraciones (sd_rating), (2) la asimetría (skew_rating), y (3) la curtosis (kurt_rating). Como variables independientes se incluyeron la dificultad percibida (diff_index), la valoración media (mean_rating), el número de reseñas (n_reviews). En el modelo de regresión se incluyeron de forma explícita sus términos cuadráticos e interacciones, para poder capturar posibles relaciones no lineales. Todas las variables numéricas se estandarizaron para facilitar la comparabilidad de coeficientes entre predictores.

#### Estimación y diagnóstico de los modelos OLS

Los modelos de regresión lineal se estimaron mediante mínimos cuadrados ordinarios, verificando posteriormente los supuestos fundamentales: multicolinealidad mediante los factores de inflación de varianza (VIF), normalidad de residuos mediante el test de Shapiro-Wilk (Shapiro & Wilk, 1965), y homocedasticidad mediante el test de Breusch-Pagan (Breusch & Pagan, 1979). Dado que este último indicó la presencia de

heterocedasticidad, los modelos se reestimaron utilizando errores estándar robustos tipo HC3 (Long & Ervin, 2000), lo que permite mantener la validez inferencial incluso cuando la varianza de los residuos no es constante.

Modelización mediante redes neuronales

En una segunda etapa, se entrenó una red neuronal con arquitectura de una capa oculta para validar y complementar los resultados obtenidos mediante OLS. La red se ajustó mediante validación cruzada de 10 pliegues, explorando un rango de 1 a 3 neuronas en la capa oculta y valores de decay entre $10^{-5}$ y $10^{-1}$, siguiendo un enfoque de búsqueda en rejilla (grid search). Se seleccionó la configuración que minimizó el error cuadrático medio (RMSE) en el conjunto de validación.

La interpretación de este modelo se realizó mediante el algoritmo NeuralSens (Pizarroso et al., 2022), que se basa en un análisis post-hoc basado en el cálculo de sensibilidades.

### *3.6 Análisis comparativo entre departamentos*

Con el fin de explorar la posible existencia de sesgos disciplinares en las evaluaciones docentes, se llevó a cabo un análisis comparativo entre los tres departamentos con mayor número de observaciones, en este caso *Mathematics, English* y *Psychology*. A través de este procedimiento se puede examinar si las diferencias en las valoraciones pueden asociarse al campo de conocimiento y, en particular, si existe un sesgo negativo hacia las áreas percibidas como más difíciles, como las matemáticas.

Exploración descriptiva y contraste de medias

En una primera etapa, se realizó una exploración descriptiva mediante diagramas de caja (*boxplots*) para analizar la distribución de la variable student_star en los diferentes departamentos. Esta visualización permitió identificar diferencias preliminares en las medianas y dispersión de las valoraciones. Posteriormente, se aplicó un ANOVA de un factor, con el objetivo de contrastar formalmente la igualdad de medias entre los grupos.

Verificación de supuestos y ajustes metodológicos

La homogeneidad de varianzas se verificó mediante el test de Levene (Levene, 1960). Dado que este contraste mostró una violación del supuesto de igualdad de varianzas, se optó por emplear el ANOVA de Welch (Welch, 1951), ampliamente recomendado por su

robustez frente a la heterocedasticidad y a los tamaños de muestra desiguales. Como análisis post-hoc, se aplicó la prueba de Tukey HSD, que permitió identificar los pares de departamentos con diferencias estadísticamente significativas en las valoraciones promedio.

Validación mediante análisis agregado

Para comprobar la consistencia de los resultados, el análisis se replicó a nivel agregado por profesor, sustituyendo la valoración individual (student_star) por la valoración media obtenida por cada docente (star_rating). Este enfoque permite evaluar si las diferencias entre áreas se mantienen tanto en las percepciones individuales de los estudiantes como en los patrones a nivel docente.

# 4 Resultados y Discusión

## *4.1 Análisis descriptivo de la muestra*

### *4.1.1 Análisis gráfico*

El análisis descriptivo tuvo como objetivo caracterizar la estructura general de las valoraciones contenidas en el dataset y proporcionar un primer acercamiento a la relación entre la dificultad percibida y las evaluaciones de los estudiantes. Todas las descripciones se han realizado tanto a nivel individual (registro por estudiante) como a nivel agregado (registro por profesor), siguiendo la lógica analítica definida en la metodología.

Este análisis descriptivo permitió identificar patrones generales en la forma en que los estudiantes evalúan a los docentes y en cómo perciben la dificultad de las asignaturas. Los histogramas de student_star mostraron una distribución claramente sesgada hacia valores altos (ver Figura 1), indicando que la mayoría de las valoraciones tienden a ser positivas, aunque existe una cola descendente que refleja un porcentaje no despreciable de evaluaciones negativas.

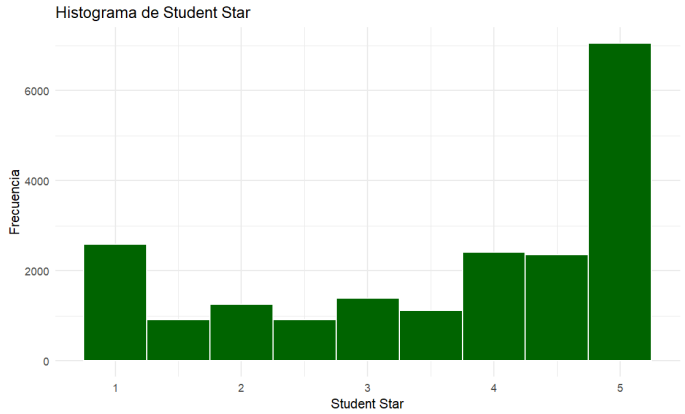



*Figura 1. Histograma de la variable student_star. Elaborado por la autora.*

Por el contrario, la distribución de student_difficult fue más uniforme (ver Figura 2), con presencia significativa de todos los niveles de dificultad, lo que sugiere que los estudiantes utilizan toda la escala para expresar su percepción sobre la exigencia.

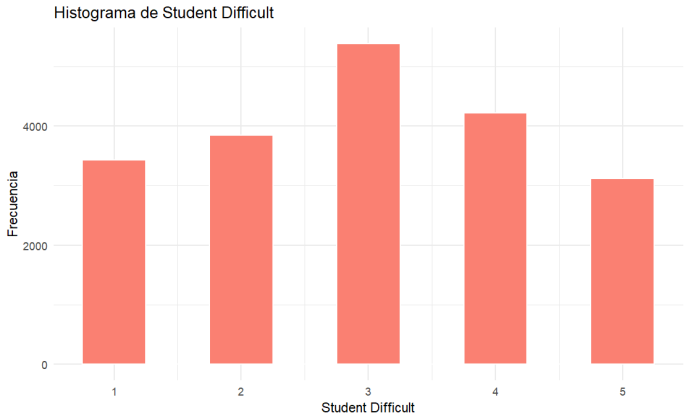



*Figura 2. Histograma de la variable student_difficult. Elaborado por la autora.*

A nivel agregado (media por profesor), star_rating replicó el sesgo hacia valores altos observado en student_star, mientras que diff_index mostró, una vez más, una dispersión notable a lo largo de toda la escala, aunque concentrada en valores medios (ver Figura 3, Figura 4 y Figura 5).

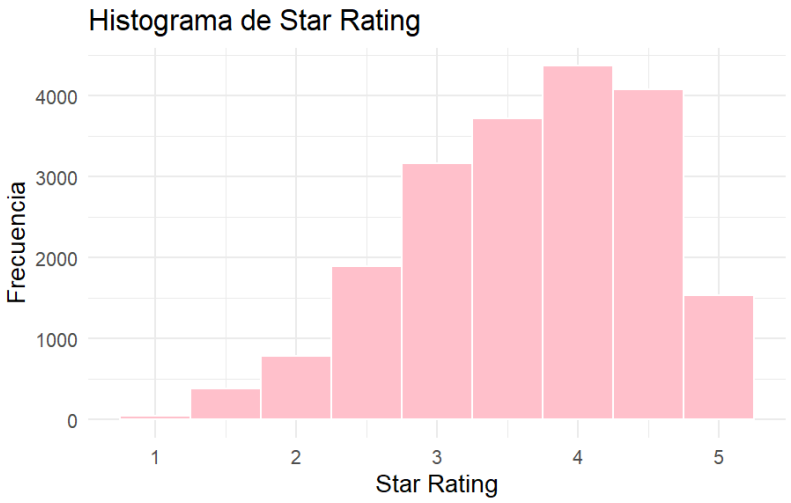



*Figura 3. Histograma de la variable star_rating. Elaborado por la autora.*

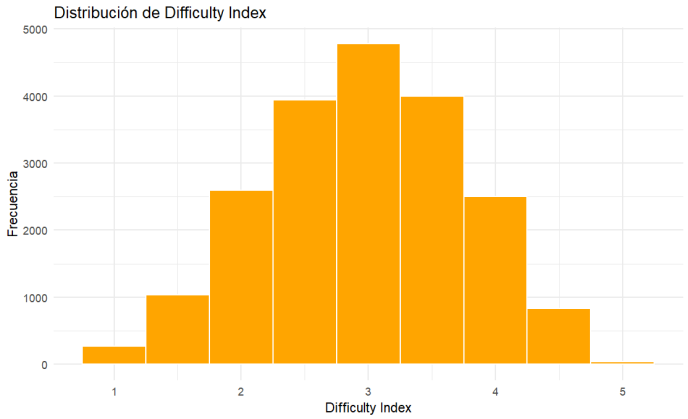



*Figura 4. Histograma de la variable diff_index. Elaborado por la autora.*

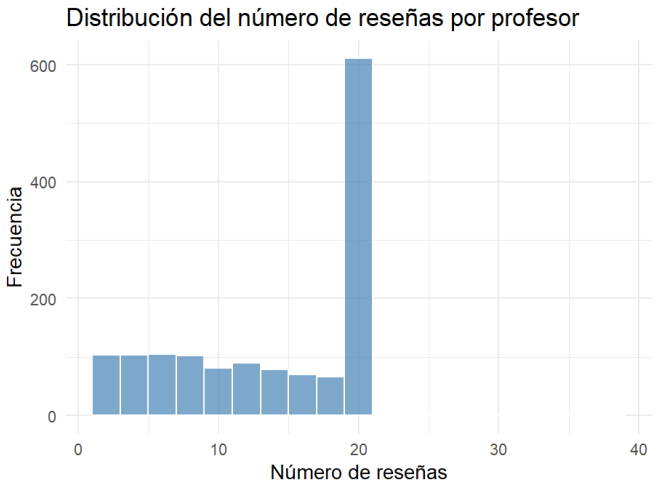



*Figura 5. Histograma del número de reseñas por profesor. Elaborada por la autora.*

Tras examinar los histogramas de las variables principales, se observa que las valoraciones estudiantiles presentan una fuerte concentración en el rango positivo, mientras que la dificultad percibida muestra una distribución mucho más heterogénea a lo largo de toda la escala. Esta diferencia en las distribuciones no permite inferir por sí misma una relación entre ambas variables, pero sí justifica explorar si existe algún tipo de asociación entre ambas dimensiones.

### *4.1.2 Análisis de correlaciones*

Tras la exploración inicial de las variables, el siguiente paso fue analizar la relación entre la dificultad percibida y las valoraciones asignadas por los estudiantes. Para ello, se calcularon correlaciones tanto a nivel individual (alumnos) como a nivel agregado (profesor). A través de este análisis, se pudo comprobar si la idea de que la dificultad podría estar influyendo negativamente en las valoraciones, se sostenía empíricamente.

A nivel individual, la correlación entre student_star y student_difficult fue claramente negativa (coeficiente de correlación = -0.44, p-valor < $2.2\ 10^{-16}$), indicando que cuanto mayor es la dificultad percibida, menor es la valoración asignada. Ello indica que los estudiantes que perciben una asignatura o profesor como más exigente tienden, en promedio, a otorgar puntuaciones más bajas.

Cuando el análisis se replica a nivel docente, agregando las observaciones por profesor, el patrón se reproduce: star_rating (valoración media del profesor) y diff_index (índice de dificultad promedio) presentan igualmente una correlación negativa (coeficiente de correlación = -0.52, p-valor < $2.2\ 10^{-16}$).

Estos resultados son especialmente relevantes, ya que muestran que la relación inversa entre dificultad y valoración no es un fenómeno aleatorio, sino que se observa en ambos niveles de análisis (estudiante y profesor) del dataset. En este sentido, las correlaciones confirman que existe una relación negativa entre dificultad percibida y valoración en las dos escalas analizadas. Esta coherencia observada en ambos niveles de análisis refuerza la hipótesis de que la dificultad actúa como un factor sistemático capaz de distorsionar las evaluaciones docentes. Además, establece una base sólida para examinar cómo esta relación influye en la estructura completa de las distribuciones (variabilidad, asimetría y curtosis) y en los perfiles docentes identificados posteriormente mediante análisis de clústeres.

### *4.1.3 Asimetría, curtosis y dispersión*

Tras establecer la relación negativa entre dificultad percibida y valoración media, el análisis se amplió a aspectos más finos de la distribución de las evaluaciones a nivel docente. El objetivo fue determinar si la dificultad influye únicamente en la media o si también afecta a la estructura completa de las valoraciones, es decir, su dispersión, su asimetría y la relación entre la concentración central de la distribución y el peso relativo en las colas(curtosis). Para ello se calcularon, para cada profesor, la desviación estándar, la asimetría y la curtosis de las valoraciones (student_star) agregadas.

La variabilidad mostró una relación positiva entre ambas variables: los profesores con mayor dificultad percibida presentaron, en general, desviaciones estándar más altas (ver Figura 6). En otras palabras, a medida que aumenta diff_index, las valoraciones se vuelven menos homogéneas y más dispersas. Esto implica que los docentes percibidos como más exigentes no generan únicamente valoraciones más bajas en promedio, sino también juicios más divididos entre los estudiantes. Una posible interpretación preliminar es que algunos estudiantes perciben la exigencia como un indicador de calidad, mientras que otros la experimentan como un obstáculo, lo que deriva en una mayor polarización de opiniones. No obstante, estos resultados deben entenderse como descriptivos: en esta fase exploratoria no es posible determinar si la dificultad por sí sola explica esta mayor dispersión o si la relación está condicionada por otras características del profesor, como su valoración media. Este aspecto se aborda de forma detallada en

los modelos posteriores, donde se evalúa qué variables mantienen efectos significativos una vez controlada la estructura completa de las evaluaciones.

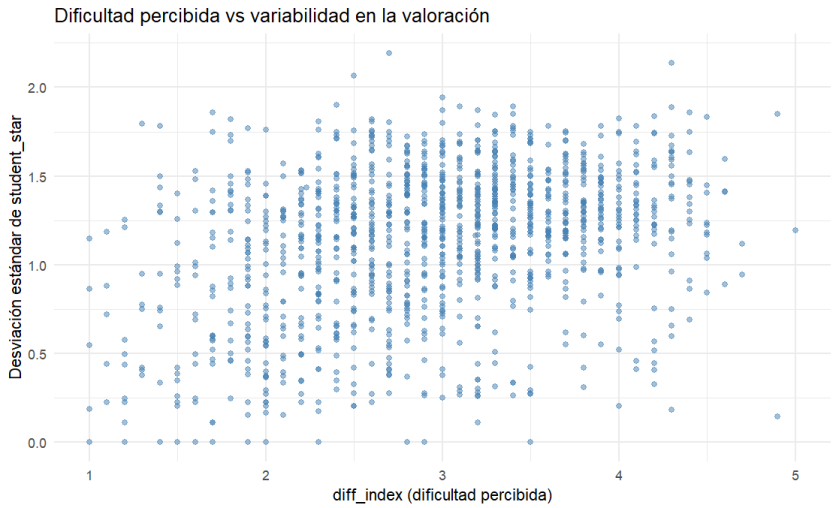



*Figura 6. Relación entre la dificultad percibida (diff_index) y la variabilidad de las valoraciones (sd_rating). Elaborada por la autora.*

En cuanto a la asimetría (ver Figura 7), las distribuciones tendieron a desplazarse hacia valores positivos conforme aumentaba la dificultad. Esto indica que, entre profesores percibidos como más exigentes, las valoraciones bajas son más frecuentes que las altas, lo que coincide con la tendencia ya observada en las correlaciones. En este sentido, la asimetría positiva sugiere que la dificultad genera una mayor concentración de opiniones desfavorables, incluso en contextos donde la media global pueda no ser extremadamente baja. De nuevo, como se comprobará posteriormente en los modelos, esta aparente relación es realmente un efecto derivado de la valoración media.

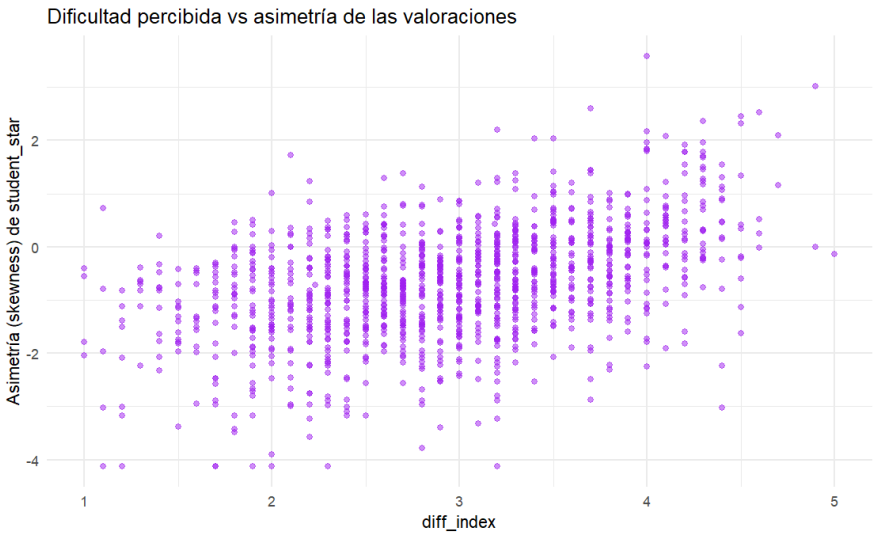



*Figura 7. Relación entre la dificultad percibida (diff_index) y la asimetría de las valoraciones estudiantiles. Elaborada por la autora.*

Respecto a la curtosis, no se aprecia una estructura clara (ver Figura 8).

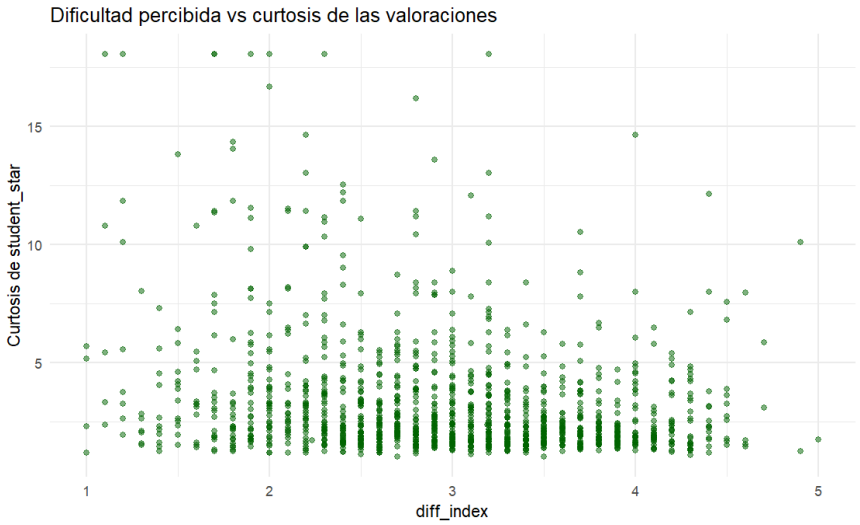



*Figura 8. Asociación entre la dificultad percibida (diff_index) y la curtosis de las valoraciones estudiantiles. Elaborada por la autora.*

Estos resultados parecen apuntar a que la dificultad no solo desplaza las valoraciones hacia niveles más bajos, sino que altera su distribución.

Con el fin de determinar si estas asociaciones descriptivas se mantienen una vez controlada la estructura completa de las evaluaciones, se estimaron modelos lineales con términos cuadráticos e interacciones, incorporando errores estándar robustos ante la presencia de heterocedasticidad (Tabla 2). Asimismo, se complementó el análisis con una red neuronal multicapa cuyo análisis de sensibilidad permite evaluar la relevancia relativa de cada variable explicativa (Tabla 3).[1]

| Variable | Estimate | Std. Error | p-value |
|---|---|---|---|
| Intercept | 0.581 | 0.023 | <0.001 |
| diff_index | 0.010 | 0.018 | 0.578 |
| I(diff_index²) | -0.006 | 0.014 | 0.633 |
| mean_rating | -0.852 | 0.019 | <0.001 |
| I(mean_rating²) | -0.571 | 0.014 | <0.001 |
| n_reviews | -0.049 | 0.018 | 0.006 |
| mean_rating × n_reviews | 0.036 | 0.021 | 0.093 |
| diff_index × n_reviews | -0.033 | 0.020 | 0.089 |

*Tabla 2: Resultados del modelo OLS con errores estándar robustos (HC3) para la variabilidad de las valoraciones (sd_rating). Elaborada por la autora.*

[1] Nota: Se reportan errores estándar robustos (HC3). No se detectan problemas de multicolinealidad (VIF < 5).

En este modelo, todos los factores de inflación de la varianza (VIF) son inferiores a 5, lo que sugiere ausencia de multicolinealidad relevante. El test de Breusch–Pagan detecta heterocedasticidad, por lo que se reportan errores estándar robustos HC3. Respecto a la variable de interés, diff_index no resulta estadísticamente significativo (p = 0.578), ni tampoco su término cuadrático, indicando que una vez controladas la valoración media (mean_rating) y el número de reseñas, la dificultad percibida no explica de forma independiente la variabilidad de las evaluaciones.

| Variable | Mean | Std |
|---|---|---|
| diff_index | 0.011 | 0.043 |
| mean_rating | -0.935 | 1.330 |
| n_reviews | -0.059 | 0.093 |

*Tabla 3: Resultados del análisis de sensibilidad (NeuralSens) para la variabilidad de las valoraciones (sd_rating). Elaborada por la autora.*

El análisis NeuralSens indica que diff_index presenta valores reducidos de media y desviación típica, lo que sugiere baja relevancia predictiva en la MLP. En cambio, mean_rating muestra una desviación típica elevada, consistente con un efecto relevante y potencialmente no lineal en la red neuronal. En conjunto, los resultados de NeuralSens son coherentes con el modelo OLS al señalar que el comportamiento de la variabilidad se asocia principalmente a mean_rating y no a diff_index.

Los resultados del modelo OLS usando como variable dependiente la asimetría se muestran en la Tabla 4, y el modelo de red neuronal con análisis post-hoc basado en el cálculo de sensibilidades en la Tabla 5.

| Variable | Estimate | Std. Error | p-value |
|---|---|---|---|
| Intercept | 0.008 | 0.017 | 0.641 |
| diff_index | 0.003 | 0.014 | 0.847 |
| I(diff_index²) | 0.010 | 0.013 | 0.411 |
| mean_rating | -0.890 | 0.019 | <0.001 |
| I(mean_rating²) | -0.026 | 0.015 | 0.080 |
| n_reviews | -0.119 | 0.014 | <0.001 |
| mean_rating × n_reviews | -0.141 | 0.017 | <0.001 |
| diff_index × n_reviews | -0.005 | 0.015 | 0.744 |

*Tabla 4: Resultados del modelo OLS con errores estándar robustos (HC3) para la asimetría de las valoraciones (skew_rating). Elaborada por la autora.*

Los VIF del modelo son inferiores a 5, descartando multicolinealidad sustancial. Dado que el test de Breusch–Pagan indica heterocedasticidad, se reportan errores estándar

robustos HC3. diff_index no es estadísticamente significativo (p = 0.847), ni su término cuadrático, por lo que no se observa un efecto independiente de la dificultad sobre la asimetría al controlar por mean_rating y n_reviews. La asimetría se asocia principalmente con mean_rating y con su interacción con el número de reseñas.

| Variable | Mean | Std |
|---|---|---|
| diff_index | -0.002 | 0.030 |
| mean_rating | -0.891 | 0.656 |
| n_reviews | -0.061 | 0.223 |

*Tabla 5: Resultados del análisis de sensibilidad (NeuralSens) para la asimetría de las valoraciones (skew_rating). Elaborada por la autora*

En la MLP, diff_index mantiene valores bajos de media y desviación típica, lo que apunta a irrelevancia predictiva. Por el contrario, mean_rating presenta mayor sensibilidad, sugiriendo un efecto relevante (y con posible componente no lineal). Esto refuerza la consistencia entre el enfoque lineal (OLS) y el enfoque flexible (MLP).

Los resultados del modelo OLS usando como variable dependiente la curtosis se muestran en la Tabla 6, y el modelo MLP con análisis post-hoc en la Tabla 7.

| Variable | Estimate | Std. Error | p-value |
|---|---|---|---|
| Intercept | -0.472 | 0.023 | <0.001 |
| diff_index | 0.004 | 0.023 | 0.866 |
| I(diff_index²) | 0.022 | 0.021 | 0.304 |
| mean_rating | 0.680 | 0.035 | <0.001 |
| I(mean_rating²) | 0.464 | 0.028 | <0.001 |
| n_reviews | 0.212 | 0.019 | <0.001 |
| mean_rating × n_reviews | 0.181 | 0.029 | <0.001 |
| diff_index × n_reviews | -0.040 | 0.022 | 0.078 |

*Tabla 6: Resultados del modelo OLS con errores estándar robustos (HC3) para la curtosis de las valoraciones (kurt_rating). Elaborada por la autora.*

Todos los VIF son inferiores a 5, por lo que no se aprecian problemas relevantes de multicolinealidad. El test de Breusch–Pagan detecta heterocedasticidad, motivo por el que se reportan errores estándar robustos HC3. En este modelo, diff_index tampoco resulta significativo (p = 0.866) ni su término cuadrático, lo que sugiere que la dificultad percibida no explica de forma independiente la curtosis una vez controladas mean_rating y n_reviews.

| Variable | Mean | Std |
|---|---|---|

| | | |
|---|---|---|
| diff_index | 0.005 | 0.061 |
| mean_rating | 1.267 | 4.482 |
| n_reviews | 0.176 | 0.328 |

*Tabla 7: Resultados del análisis de sensibilidad (NeuralSens) para la curtosis de las valoraciones (kurt_rating). Elaborada por la autora.*

NeuralSens confirma que diff_index tiene baja sensibilidad y variabilidad en la misma (media y desviación típica reducidas). En cambio, mean_rating presenta una desviación típica muy elevada, compatible con un efecto relevante y marcadamente no lineal dentro de la MLP. De nuevo, el patrón es consistente con OLS: la curtosis se asocia principalmente con mean_rating más que con diff_index.

### *4.2 Análisis clúster*

Tras el análisis descriptivo y correlacional inicial, se exploró si la combinación entre la dificultad percibida y la valoración global del profesorado generaba patrones estructurados dentro del conjunto de datos. Para ello, tras estandarizar las variables, se aplicó un análisis clúster no jerárquico mediante el algoritmo K-means, previamente justificado en la metodología. El coeficiente de Silhouette indicó que la solución más coherente era la formada por dos grupos (Silhouette = 0.44, el mayor de todos los k evaluados), lo que sugiere que los docentes tienden a organizarse en dos perfiles diferenciados atendiendo a estas dos dimensiones. No obstante, se debe indicar que este valor de Silhouette muestra una agrupación moderadamente buena, válida para análisis exploratorios, pero no indica clústeres extremadamente nítidos. La representación gráfica de los clústeres (ver Figura 9) muestra las dos agrupaciones.

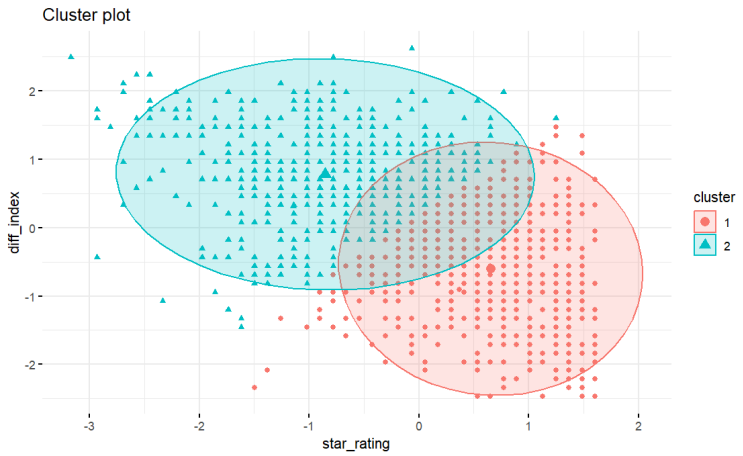



*Figura 9. Agrupación de profesores mediante K-means según valoración global y dificultad percibida. Elaborada por la autora.*

El análisis de clústeres permitió identificar dos grupos bien diferenciados de profesores en función de la valoración global (star_rating) y la dificultad percibida (diff_index). La tabla siguiente resume el número de profesores en cada clúster y los valores medios asociados:

| Clúster | Nº Profesores | Valoración Media (star_rating) | Desviación Típica Valoración | Dificultad Media (diff_index) |
|---|---|---|---|---|
| 1 | 711 | 4,20 | 0,47 | 2,47 |
| 2 | 545 | 2,94 | 0,65 | 3,56 |

*Tabla 8: Valores de clústeres. Elaborada por la autora.*

Los valores recogidos en la Tabla 8 permiten describir con claridad la composición de los grupos identificados mediante el análisis de clústeres. El clúster 1, que agrupa a 711 profesores, se caracteriza por presentar valoraciones globales elevadas (media = 4.20) junto con niveles bajos o moderados de dificultad percibida (media = 2.47). Además, este grupo muestra una menor dispersión en las valoraciones (desviación estándar = 0.473), lo que indica que las evaluaciones tienden a ser más homogéneas. Este patrón refuerza lo observado previamente en los histogramas y análisis correlacionales: cuando los estudiantes perciben al profesor como relativamente accesible, las puntuaciones se concentran en la parte superior de la escala y existe un mayor grado de consenso en la valoración de la experiencia docente.

En contraste, el clúster 2, compuesto por 545 profesores, presenta un perfil claramente diferenciado, con índices de dificultad perceptiblemente más altos (media = 3.56) y valoraciones medias más bajas (media = 2.94). En este grupo, la variabilidad de las evaluaciones es mayor (desviación estándar = 0.650), lo que refleja una estructura de valoraciones menos uniforme. Esta mayor dispersión sugiere la coexistencia de opiniones más diversas entre los estudiantes, coherente con los resultados obtenidos en los análisis de variabilidad, asimetría y curtosis, donde los profesores percibidos como más exigentes tendían a generar distribuciones más irregulares y con mayor presencia de respuestas extremas.

En conjunto, el análisis de clústeres confirma y refuerza los patrones detectados en las fases descriptiva y correlacional. La identificación de dos grupos bien diferenciados sugiere que la dificultad percibida actúa como un eje estructurador de las evaluaciones

docentes, dando lugar a perfiles coherentes: por un lado, profesores con menor dificultad percibida y valoraciones altas y estables; y, por otro, profesores más exigentes, con valoraciones medias más bajas y mayor heterogeneidad en las evaluaciones recibidas. Esta consistencia entre los distintos niveles de análisis subraya la pertinencia de profundizar en estos resultados mediante técnicas de regresión y redes neuronales en las secciones posteriores.

### *4.3 Análisis comparativo entre departamentos*

Con el objetivo de explorar la posible existencia de sesgos disciplinares en las evaluaciones docentes, se llevó a cabo un análisis comparativo entre los departamentos con mayor número de observaciones en el conjunto de datos. En concreto, se seleccionaron *Mathematics, English* y *Psychology* por el volumen de registros y, al representar áreas académicas con distintos niveles tradicionalmente asociados de dificultad percibida. En una primera etapa, se realizó un análisis descriptivo mediante *boxplots* de valoración por departamento, tanto a nivel estudiante (Figura 10) como a nivel profesor, es decir, agregado (Figura 11).

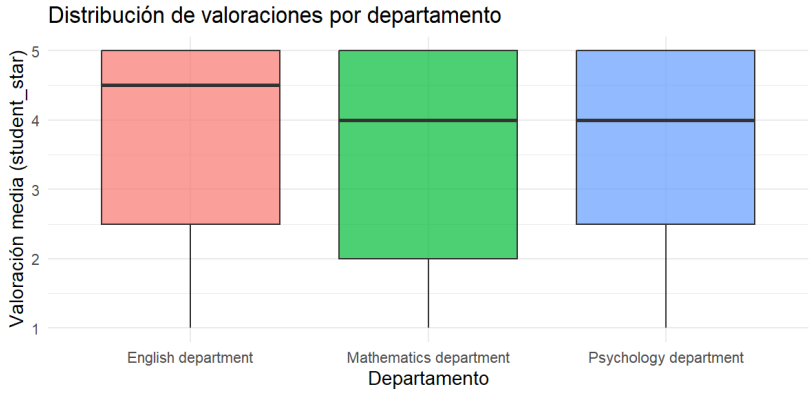



*Figura 10: Distribución de student_star por departamento (nivel estudiante). Elaborada por la autora.*

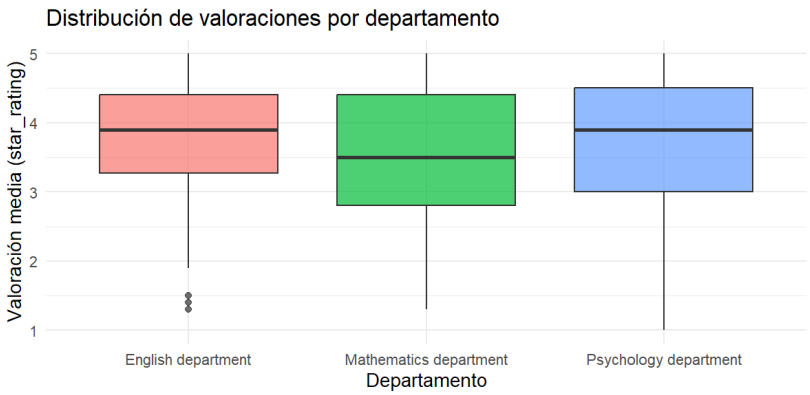



*Figura 11: Distribución de star_rating por departamento (nivel profesor). Elaborada por la autora.*

Los resultados muestran diferencias visibles en la distribución de las valoraciones entre áreas académicas. En particular, el departamento de *Mathematics* presenta una mediana inferior y una mayor dispersión relativa en comparación con *English* y *Psychology*. Ello apunta a que los profesores de disciplinas cuantitativas podrían recibir evaluaciones sistemáticamente más bajas por parte de los estudiantes.

Para contrastar formalmente estas diferencias, se estimó un ANOVA de un factor, previamente, verificando el supuesto de homogeneidad de varianzas mediante el test de Levene, el que indicó que no se cumplía. Por ello, se empleó el ANOVA de Welch, que es robusto frente a heterocedasticidad y tamaños muestrales desiguales (Tabla 9).

| Fuente | gl | Suma de cuadrados | Media cuadrática | F | p-valor |
|---|---|---|---|---|---|
| Departamento | 2 | 86.000 | 43.150 | 20.020 | $p < 0.001$ |
| Residual | 5368 | 11569.000 | 2.160 | | |

*Tabla 9: ANOVA – Valoraciones a nivel estudiante (sin agregar por profesor). Elaborada por la autora.*

El efecto del departamento sobre las valoraciones estudiantiles es estadísticamente significativo ($F = 20.020$, $p < 0.001$). Esto indica que existen diferencias sistemáticas entre departamentos en la valoración media otorgada por los estudiantes. Estos resultados confirman la existencia de diferencias estadísticamente significativas en las valoraciones medias entre departamentos. El análisis posterior, mediante la prueba de Tukey, permitió identificar qué pares de departamentos presentan diferencias significativas. Los resultados mostraron que *Mathematics* difiere significativamente de *English* y *Psychology*, mientras que la diferencia es menor entre estas dos últimas. Este resultado es coherente con la hipótesis de que las disciplinas percibidas como más exigentes o cuantitativamente intensivas tienden a recibir evaluaciones más bajas. Además, para

reforzar la robustez de estos resultados, y continuando con la metodología multinivel, el análisis se replicó a nivel agregado por profesor (Tabla 10), utilizando la variable star_rating en lugar de student_star.

| Fuente | gl | Suma de cuadrados | Media cuadrática | F | p-valor |
|---|---|---|---|---|---|
| Departamento | 2 | 5.060 | 2.528 | 3.171 | 0.043 |
| Residual | 351 | 279.870 | 0.797 | | |

*Tabla 10: ANOVA – Valoraciones a nivel profesor (agregado por profesor). Elaborada por la autora.*

Los resultados obtenidos mantienen el mismo esquema general observado a nivel individual, lo que sugiere que las diferencias disciplinares no se deben únicamente a variaciones en respuestas individuales aisladas, sino que reflejan tendencias estructurales en las evaluaciones. Estos hallazgos sugieren que factores relacionados la disciplina pueden influir sistemáticamente en las evaluaciones docentes. Es importante señalar que, a pesar de que este análisis no permite establecer relaciones causales, sí sugiere la existencia de un posible sesgo disciplinar que debe considerarse en la interpretación de las SET, especialmente cuando se utilizan con fines comparativos entre áreas académicas diversas.

# 5 Conclusiones

Este estudio examina la existencia y la magnitud de los sesgos no instruccionales en las evaluaciones de la docencia realizadas por los estudiantes (SET), prestando especial atención al papel de la dificultad percibida del curso y la disciplina de este. Utilizando un conjunto de datos a gran escala extraído de RateMyProfessor.com, el análisis se ha realizado combinando estadísticas descriptivas, análisis de correlación, técnicas de *clustering* y modelos basados en regresión y machine learning para proporcionar una valoración empírica de cómo las percepciones relacionadas con la dificultad influyen en los resultados de la evaluación. Los resultados proporcionan información valiosa que contribuye al debate en curso sobre la validez y la interpretación de las SET en la educación superior.

A nivel descriptivo, el análisis confirma patrones en las propiedades distributivas de las evaluaciones de los estudiantes. Las calificaciones asignadas por los estudiantes se inclinan fuertemente hacia el extremo superior de la escala, lo que indica que la mayoría de los profesores reciben evaluaciones relativamente favorables. Este efecto techo,

ampliamente documentado en investigaciones anteriores, limita el contenido informativo de las calificaciones medias y complica las comparaciones entre profesores. Por el contrario, la dificultad percibida muestra una distribución mucho más heterogénea, ya que los estudiantes utilizan ampliamente toda la escala para expresar lo exigentes que consideran que son los cursos. Es importante destacar que los datos descriptivos también ponen de relieve una heterogeneidad sustancial a nivel de profesor. A pesar de que las calificaciones medias tienden a agruparse en un rango de valores altos, el número de reseñas por profesor varía considerablemente y la estructura interna de las evaluaciones difiere notablemente entre los distintos docentes. En este sentido, algunos profesores reciben evaluaciones muy homogéneas, mientras que otros generan respuestas muy polarizadas entre los estudiantes

Más allá de los efectos medios, una de las contribuciones centrales de este estudio radica en su examen de cómo la dificultad percibida se relaciona con la estructura interna de las distribuciones de las evaluaciones. El análisis descriptivo de la variabilidad parece apuntar a que los profesores percibidos como más exigentes tienden a recibir evaluaciones más dispersas. En la práctica, esto significa que los cursos difíciles no solo hacen que las evaluaciones bajen de manera uniforme, sino que generan un menor consenso entre los estudiantes. Algunos estudiantes pueden interpretar las altas exigencias como una señal de rigor y calidad, mientras que otros las perciben como excesivas, lo que da lugar a respuestas más polarizadas. Del mismo modo, el análisis de la asimetría sugiere que una mayor dificultad percibida se asocia con distribuciones que se inclinan más hacia calificaciones más bajas. Este hallazgo sugiere que, entre los cursos exigentes, las evaluaciones desfavorables se vuelven más frecuentes en relación con las favorables, incluso cuando las puntuaciones medias se mantienen dentro de un rango relativamente estrecho. Por el contrario, el análisis de la curtosis no revela un patrón claro o sistemático con respecto a la dificultad, lo que indica que la presencia de evaluaciones extremas no se debe únicamente a las exigencias del curso.

El análisis clúster refuerza aún más el papel de la dificultad percibida como dimensión estructural en las evaluaciones de los estudiantes. Considerando de manera conjunta la dificultad y la valoración global de los docentes, el análisis identifica dos clústeres. El primero se caracteriza por una dificultad percibida relativamente baja o moderada y

calificaciones medias altas, acompañadas de una mayor homogeneidad en las evaluaciones de los estudiantes. El segundo grupo está formado por profesores que imparten materias de una mayor dificultad percibida y calificaciones medias más bajas, así como a una mayor dispersión interna. La existencia de estos grupos sugiere que los resultados de la SET no se distribuyen de forma aleatoria entre los profesores, sino que se organizan de acuerdo con perfiles distintos. En este sentido, la dificultad percibida actúa como un factor estructural de las evaluaciones docentes.

En líneas generales, estos resultados sugieren que la dificultad percibida afecta no solo al nivel de las evaluaciones, sino también a su forma. La dificultad parece actuar como un factor que aumenta la heterogeneidad y la asimetría en los juicios de los estudiantes, alterando así la estructura informativa de los datos SET. Sin embargo, es importante destacar que estos hallazgos son de naturaleza descriptiva. Como se muestra en la etapa de modelización posterior, algunas de estas relaciones aparentes están mediadas mayoritariamente por la propia calificación media, que absorbe gran parte de la variación en las características de distribución una vez incluidas explícitamente en los modelos. Esto pone de manifiesto la importancia de una especificación cuidadosa del modelo a la hora de interpretar los patrones de distribución en los datos SET.

En cuanto a los análisis de correlación, proporcionan pruebas de una asociación negativa entre la dificultad percibida y las evaluaciones de los estudiantes. Esta relación se manifiesta tanto a nivel de los estudiantes como a nivel agregado de los profesores, observándose correlaciones negativas de moderadas a fuertes en ambos casos. El hecho de que este patrón se repita en ambos niveles de agregación sugiere que la relación entre la dificultad y las puntuaciones de las evaluaciones no está determinada por percepciones individuales aisladas, sino que refleja una tendencia sistemática dentro de los datos. Los profesores asociados a cursos más exigentes tienden, en promedio, a recibir evaluaciones más bajas por parte de los estudiantes. Estos hallazgos corroboran los descubrimientos de trabajos empíricos previos que documentan sesgos relacionados con la dificultad en las SET y refuerzan la interpretación de la dificultad percibida como un factor no instructivo que afecta sistemáticamente a los resultados de la evaluación.

Los análisis de regresión y machine learning proporcionan información adicional sobre los mecanismos que subyacen a estos patrones. Una vez que las calificaciones medias se

incluyen explícitamente en los modelos, los efectos directos de la dificultad percibida en las medidas de distribución, como la variabilidad y la asimetría, se debilitan sustancialmente y dejan de ser estadísticamente significativos. Esto indica que gran parte de la influencia aparente de la dificultad en la forma de las distribuciones de evaluación opera indirectamente a través de su efecto en las calificaciones medias. En otras palabras, la dificultad percibida afecta principalmente a la puntuación media alta o baja que se otorga a los profesores, y estos cambios en la media, a su vez, determinan las propiedades de distribución observadas.

Adicionalmente, el análisis comparativo entre departamentos aporta evidencia coherente con la existencia de un sesgo disciplinar. El departamento de Mathematics presenta valoraciones sistemáticamente inferiores tanto a nivel individual como agregado por profesor, incluso tras emplear contrastes robustos ante heterocedasticidad. Este resultado sugiere que determinadas áreas académicas, especialmente aquellas percibidas como más exigentes, pueden estar estructuralmente penalizadas en las evaluaciones docentes, lo que refuerza la necesidad de interpretar las SET con cautela en procesos de toma de decisiones académicas.

Por último, este estudio contribuye a la bibliografía existente al trasladar la atención de los análisis basados en la media hacia un examen más completo de la estructura de las distribuciones de las evaluaciones. Al integrar enfoques descriptivos, de agrupación, de regresión y de aprendizaje automático en un único marco empírico, ofrece una perspectiva matizada sobre cómo funciona la dificultad percibida en los datos de las SET. Si bien es cierto que el análisis tiene sus limitaciones, sobre todo por la naturaleza observacional de los datos y la dependencia de las percepciones autoinformadas, ofrece sin embargo pruebas sólidas de que la dificultad funciona como un factor organizativo clave en las evaluaciones de los estudiantes.

Las investigaciones futuras podrían ampliar este trabajo examinando los mecanismos causales con mayor profundidad, examinando la dinámica longitudinal en las evaluaciones o investigando cómo las políticas institucionales interactúan con los sesgos no relacionados con la enseñanza. No obstante, los presentes hallazgos ya subrayan un mensaje central: las evaluaciones de la enseñanza por parte de los estudiantes están

determinadas por fuerzas sistemáticas que van más allá de la calidad de la enseñanza, y comprender estas fuerzas es esencial para su interpretación y uso responsables.

## 6 Declaración respecto al uso de Chat GPT u otras herramientas de IAG

Por la presente, yo, Karen Camacho Vilacoba, estudiante del Doble Grado de Relaciones Internacionales y Business Analytics, de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Sesgos no instruccionales en encuestas de evaluación del profesorado", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
2. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
4. Interpretador de código: Para realizar análisis de datos preliminares.
5. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
6. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
7. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
8. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
9. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
10. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
11. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
12. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

# 7 Referencias


Ali, H. I. H., & Al Ajmi, A. A. S. (2013). Exploring non-instructional factors in student evaluations. *Higher Education Studies, 3*(5), 81–93. https://doi.org/10.5539/hes.v3n5p81

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica, 47*(5), 1287–1294. https://doi.org/10.2307/1911963

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

He, J. (2020). *Big data set from RateMyProfessor.com for professors' teaching evaluation* (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/fvtfjyvw7d.2

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education, 36*(1), 29–42. https://doi.org/10.3200/JECE.36.1.29-42

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software, 74*(7), 1–16. https://doi.org/10.18637/jss.v074.i07

Liu, R. (2022). Data analysis of educational evaluation using K-means clustering method. *Computational Intelligence and Neuroscience, 2022*, Article 3762431. https://doi.org/10.1155/2022/3762431

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *The American Statistician, 54*(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). University of California Press.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of*

*teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer. https://doi.org/10.1007/1-4020-5742-3_9

Pizarroso, J., Portela, J., & Muñoz, A. (2022). NeuralSens: Sensitivity analysis of neural networks. *Journal of Statistical Software, 102*(7), 1–36. https://doi.org/10.18637/jss.v102.i07

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.4) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Schneider, G. (2013). Student evaluations, grade inflation and pluralistic teaching: Moving from customer satisfaction to student learning and critical thinking. *Forum for Social Economics, 42*(1), 122–135. https://doi.org/10.1080/07360932.2013.771128

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.2307/2333709

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education: Development of an instrument based on 10 Likert scales. *Assessment & Evaluation in Higher Education, 32*(6), 667–679. https://doi.org/10.1080/02602930601117191

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness and student evaluation of teaching ratings. *Studies in Educational Evaluation, 54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38*(3–4), 330–336. https://doi.org/10.2307/2332579

Young, P., Rush, L., Shaw, D., & Inglis, S. (1999). Student evaluation of faculty: Effects of purpose on pattern. *Journal of Personnel Evaluation in Education, 13*(2), 179–190. https://doi.org/10.1023/A:1008124606262

Zhang, H., & Luo, F. (2020). The development of psychological and educational measurement in China. *Chinese/English Journal of Educational Measurement and Evaluation, 1*(1), Article 7. https://doi.org/10.59863/BUAI8988

Zhao, L., Xu, P., Chen, Y., & Yan, S. (2022). A literature review of the research on students' evaluation of teaching in higher education. *Frontiers in Psychology, 13*, Article 1004487. https://doi.org/10.3389/fpsyg.2022.1004487

# 8 Anexo

```
################################################################################
# TRABAJO FIN DE GRADO
# Karen Camacho Vilacoba
# Sesgos no instruccionales en encuestas de evaluación del profesorado
################################################################################

################################################################################
# 1. Librerías
################################################################################

library(dplyr)
library(ggplot2)
library(VIM)
library(corrplot)
library(cluster)
library(factoextra)
library(moments)
library(caret)
library(car)
library(lmtest)
library(sandwich)
library(NeuralSens)
library(nnet)

################################################################################
# 2. Carga y preparación inicial de datos
################################################################################

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

datos <- read.csv("RateMyProfessor_Sample_data.csv")
summary(datos)
str(datos)

# Variables categóricas
datos$department_name <- as.factor(datos$department_name)
datos$school_name     <- as.factor(datos$school_name)
datos$professor_name  <- as.factor(datos$professor_name)
datos$local_name      <- as.factor(datos$local_name)
datos$state_name      <- as.factor(datos$state_name)
datos$tag_professor   <- as.factor(datos$tag_professor)
datos$grades          <- as.factor(datos$grades)
datos$for_credits     <- as.factor(datos$for_credits)
datos$attence         <- as.factor(datos$attence)

# Variables numéricas
datos$take_again        <- as.numeric(datos$take_again)
datos$star_rating       <- as.numeric(datos$star_rating)
datos$diff_index        <- as.numeric(datos$diff_index)
datos$student_star      <- as.numeric(datos$student_star)
datos$student_difficult <- as.numeric(datos$student_difficult)

################################################################################
# 3. Imputación de valores perdidos (KNN)
################################################################################

colSums(is.na(datos))

# Imputación de student_star usando diff_index y student_difficult
datos_star_knn <- datos %>% select(student_star, diff_index, student_difficult)
datos_star_imp <- kNN(
  datos_star_knn,
  variable = "student_star",
  k = 5,
```

```
  dist_var = c("diff_index", "student_difficult"),
  imp_var = FALSE
)
datos$student_star <- datos_star_imp$student_star

# Imputación de student_difficult usando diff_index y student_star
datos_diff_knn <- datos %>% select(student_difficult, diff_index, student_star)
datos_diff_imp <- kNN(
  datos_diff_knn,
  variable = "student_difficult",
  k = 5,
  dist_var = c("diff_index", "student_star"),
  imp_var = FALSE
)
datos$student_difficult <- datos_diff_imp$student_difficult

colSums(is.na(datos[c("student_star", "student_difficult")]))

################################################################################
# 4. Análisis descriptivo y correlaciones
################################################################################

# Conteos básicos
n_prof  <- n_distinct(datos$professor_name)
n_sch   <- n_distinct(datos$school_name)
n_dept  <- n_distinct(datos$department_name)
n_state <- n_distinct(datos$state_name)

# Histogramas
ggplot(datos, aes(x = diff_index)) +
  geom_histogram(binwidth = 0.5, fill = "orange", color = "white") +
  theme_minimal() +
  labs(title = "Distribución de Difficulty Index", x = "Difficulty Index", y =
"Frecuencia")

ggplot(datos, aes(x = student_star)) +
  geom_histogram(binwidth = 0.5, fill = "darkgreen", color = "white") +
  theme_minimal() +
  labs(title = "Histograma de Student Star", x = "Student Star", y = "Frecuencia")

ggplot(datos, aes(x = student_difficult)) +
  geom_histogram(binwidth = 0.5, fill = "salmon", color = "white") +
  theme_minimal() +
  labs(title = "Histograma de Student Difficult", x = "Student Difficult", y =
"Frecuencia")

ggplot(datos, aes(x = star_rating)) +
  geom_histogram(binwidth = 0.5, fill = "pink", color = "white") +
  theme_minimal() +
  labs(title = "Histograma de Star Rating", x = "Star Rating", y = "Frecuencia")

# Número de reseñas por profesor
prof_reviews <- datos %>%
  group_by(professor_name) %>%
  summarise(n_reviews = n(), .groups = "drop")

ggplot(prof_reviews, aes(x = factor(n_reviews))) +
  geom_bar(fill = "steelblue", color = "white", alpha = 0.8) +
  labs(
    title = "Distribución del número de reseñas por profesor",
    x = "Número de reseñas (n_reviews)",
    y = "Número de profesores"
  ) +
  theme_minimal(base_size = 14) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
# Base única (una fila por profesor)
datos_unicos <- datos %>% distinct(professor_name, .keep_all = TRUE)

# Correlaciones
cor.test(datos$student_star, datos$student_difficult, use = "pairwise.complete.obs")
cor.test(datos_unicos$star_rating, datos_unicos$diff_index)

# Matrices de correlación
num_est <- datos[, c("student_star", "student_difficult")]
num_est[] <- lapply(num_est, function(x) as.numeric(as.character(x)))
M_est <- cor(num_est, use = "pairwise.complete.obs")
corrplot(M_est, method = "color", addCoef.col = "black", tl.col = "black", tl.srt =
45)

num_prof <- datos_unicos[, c("star_rating", "diff_index")]
num_prof[] <- lapply(num_prof, function(x) as.numeric(as.character(x)))
M_prof <- cor(num_prof, use = "pairwise.complete.obs")
corrplot(M_prof, method = "color", addCoef.col = "black", tl.col = "black", tl.srt =
45)

################################################################################
# 5. Análisis clúster (K-means) a nivel profesor
################################################################################

vars_evaluacion_prof <- c("star_rating", "diff_index")

datos_profesores <- datos %>%
  group_by(professor_name) %>%
  summarise(
    n_reviews = n(),
    across(all_of(vars_evaluacion_prof), ~ mean(.x, na.rm = TRUE)),
    .groups = "drop"
  )

# Filtramos profesores con pocas reseñas
min_n <- 5
datos_profesores <- datos_profesores %>% filter(n_reviews >= min_n)

X_prof <- datos_profesores %>%
  select(all_of(vars_evaluacion_prof)) %>%
  mutate(across(everything(), as.numeric)) %>%
  as.data.frame()

keep <- complete.cases(X_prof)
datos_profesores <- datos_profesores[keep, ]
X_prof <- X_prof[keep, ]

X_prof_s <- scale(X_prof)

probe_k <- function(Xs, kmin = 2, kmax = 6) {
  out <- data.frame(k = kmin:kmax, silhouette = NA_real_)
  for (i in seq_len(nrow(out))) {
    k <- out$k[i]
    set.seed(123 + k)
    km <- kmeans(Xs, centers = k, nstart = 25)
    sil <- silhouette(km$cluster, dist(Xs, method = "euclidean"))
    out$silhouette[i] <- mean(sil[, "sil_width"])
  }
  out
}

res_k_prof <- probe_k(X_prof_s, 2, 6)
print(res_k_prof)

best_k_prof <- res_k_prof$k[which.max(res_k_prof$silhouette)]
```

```r
set.seed(123)
km_prof <- kmeans(X_prof_s, centers = best_k_prof, nstart = 25)
datos_profesores$cluster_prof <- factor(km_prof$cluster)

perfil_prof <- datos_profesores %>%
  group_by(cluster_prof) %>%
  summarise(
    n = n(),
    across(all_of(vars_evaluacion_prof), ~ mean(.x, na.rm = TRUE), .names =
"mean_{.col}"),
    .groups = "drop"
  )
print(perfil_prof)

fviz_cluster(
  km_prof, data = X_prof_s,
  geom = "point", show.clust.cent = TRUE,
  ellipse.type = "norm",
  repel = TRUE, ggtheme = theme_minimal()
)

tabla_clusters <- datos_profesores %>%
  group_by(cluster_prof) %>%
  summarise(
    n_profesores = n(),
    mean_star_rating = mean(star_rating, na.rm = TRUE),
    sd_star_rating   = sd(star_rating, na.rm = TRUE),
    mean_diff_index  = mean(diff_index, na.rm = TRUE),
    sd_diff_index    = sd(diff_index, na.rm = TRUE),
    .groups = "drop"
  )
tabla_clusters

################################################################################
# 6. Modelos: dificultad y estructura distributiva (sd, skew, kurt)
################################################################################

set.seed(123)

datos_shape <- datos %>%
  group_by(professor_name) %>%
  summarise(
    n_reviews   = n(),
    diff_index  = mean(diff_index, na.rm = TRUE),
    mean_rating = mean(student_star, na.rm = TRUE),
    sd_rating   = sd(student_star, na.rm = TRUE),
    skew_rating = skewness(student_star, na.rm = TRUE),
    kurt_rating = kurtosis(student_star, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  filter(n_reviews >= 5)

# Filtramos momentos no definidos y casos degenerados (evitamos fallos en train())
datos_shape <- datos_shape %>%
  filter(
    !is.na(sd_rating),
    !is.na(skew_rating),
    !is.na(kurt_rating),
    sd_rating > 0
  )

# Gráficos exploratorios
ggplot(datos_shape, aes(x = diff_index, y = sd_rating)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  theme_minimal() +
  labs(
```

```
    title = "Dificultad percibida vs variabilidad en la valoración",
    x = "diff_index (dificultad percibida)",
    y = "Desviación estándar de student_star"
  )

ggplot(datos_shape, aes(x = diff_index, y = skew_rating)) +
  geom_point(alpha = 0.5, color = "purple") +
  theme_minimal() +
  labs(
    title = "Dificultad percibida vs asimetría de las valoraciones",
    x = "diff_index",
    y = "Asimetría (skewness) de student_star"
  )

ggplot(datos_shape, aes(x = diff_index, y = kurt_rating)) +
  geom_point(alpha = 0.5, color = "darkgreen") +
  theme_minimal() +
  labs(
    title = "Dificultad percibida vs curtosis de las valoraciones",
    x = "diff_index",
    y = "Curtosis de student_star"
  )

# Control de tuning para MLP (común)
ctrl_tune <- trainControl(
  method = "cv",
  number = 10,
  summaryFunction = defaultSummary,
  returnResamp = "final",
  savePredictions = TRUE
)

tune_grid <- expand.grid(
  size  = 1:3,
  decay = 10^seq(-5, -1, length.out = 5)
)

num_vars <- c("diff_index", "mean_rating", "n_reviews")

# Función auxiliar: prepara datos (dummies + escalado) y estandariza Y
prep_model_data <- function(df, y_name) {
  df[num_vars] <- scale(df[num_vars])

  dv <- dummyVars(~ department_name + school_name, data = df)
  dummies <- as.data.frame(predict(dv, newdata = df))

  out <- cbind(df[, c(y_name, num_vars)], dummies)
  out[[y_name]] <- scale(out[[y_name]])
  out <- as.data.frame(out)

  # Seguridad: eliminamos NAs residuales si existieran
  out <- na.omit(out)
  out
}

# --- 6.1 Variabilidad (sd_rating) ---
datos_sd <- datos_shape %>%
  left_join(datos %>% distinct(professor_name, department_name, school_name), by =
"professor_name") %>%
  select(sd_rating, diff_index, mean_rating, n_reviews, department_name, school_name)

datos_final_sd <- prep_model_data(datos_sd, "sd_rating")

modelo_sd <- lm(
  sd_rating ~ diff_index + I(diff_index^2) +
    mean_rating + I(mean_rating^2) +
```

```
    n_reviews + n_reviews:mean_rating +
    diff_index:n_reviews,
  data = datos_final_sd
)

vif(modelo_sd)
bptest(modelo_sd)

modelo_sd_rob <- vcovHC(modelo_sd, type = "HC3")
coeftest(modelo_sd, vcov. = modelo_sd_rob)

set.seed(2025)
mlp_sd <- train(
  sd_rating ~ diff_index + mean_rating + n_reviews,
  data = datos_final_sd,
  method = "nnet",
  linout = TRUE,
  maxit = 250,
  tuneGrid = tune_grid,
  trControl = ctrl_tune,
  metric = "RMSE"
)

sens_sd <- SensAnalysisMLP(mlp_sd)
summary(sens_sd)

# --- 6.2 Asimetría (skew_rating) ---
datos_skew <- datos_shape %>%
  left_join(datos %>% distinct(professor_name, department_name, school_name), by =
"professor_name") %>%
  select(skew_rating, diff_index, mean_rating, n_reviews, department_name,
school_name)

datos_final_skew <- prep_model_data(datos_skew, "skew_rating")

modelo_skew <- lm(
  skew_rating ~ diff_index + I(diff_index^2) +
    mean_rating + I(mean_rating^2) +
    n_reviews + n_reviews:mean_rating +
    diff_index:n_reviews,
  data = datos_final_skew
)

vif(modelo_skew)
bptest(modelo_skew)

modelo_skew_rob <- vcovHC(modelo_skew, type = "HC3")
coeftest(modelo_skew, vcov. = modelo_skew_rob)

set.seed(2025)
mlp_skew <- train(
  skew_rating ~ diff_index + mean_rating + n_reviews,
  data = datos_final_skew,
  method = "nnet",
  linout = TRUE,
  maxit = 250,
  tuneGrid = tune_grid,
  trControl = ctrl_tune,
  metric = "RMSE"
)

sens_skew <- SensAnalysisMLP(mlp_skew)
summary(sens_skew)

# --- 6.3 Curtosis (kurt_rating) ---
datos_kurt <- datos_shape %>%
```

```r
  left_join(datos %>% distinct(professor_name, department_name, school_name), by =
"professor_name") %>%
  select(kurt_rating, diff_index, mean_rating, n_reviews, department_name,
school_name)

datos_final_kurt <- prep_model_data(datos_kurt, "kurt_rating")

modelo_kurt <- lm(
  kurt_rating ~ diff_index + I(diff_index^2) +
    mean_rating + I(mean_rating^2) +
    n_reviews + n_reviews:mean_rating +
    diff_index:n_reviews,
  data = datos_final_kurt
)

vif(modelo_kurt)
bptest(modelo_kurt)

modelo_kurt_rob <- vcovHC(modelo_kurt, type = "HC3")
coeftest(modelo_kurt, vcov. = modelo_kurt_rob)

set.seed(2025)
mlp_kurt <- train(
  kurt_rating ~ diff_index + mean_rating + n_reviews,
  data = datos_final_kurt,
  method = "nnet",
  linout = TRUE,
  maxit = 250,
  tuneGrid = tune_grid,
  trControl = ctrl_tune,
  metric = "RMSE"
)

sens_kurt <- SensAnalysisMLP(mlp_kurt)
summary(sens_kurt)

################################################################################
# 7. Comparativa entre departamentos (sesgo negativo hacia matemáticas)
################################################################################

datos_filtrados <- datos %>%
  group_by(department_name) %>%
  filter(n() > 1200) %>%
  ungroup() %>%
  mutate(department_name = droplevels(department_name))

summary(datos_filtrados$department_name)

ggplot(datos_filtrados, aes(x = department_name, y = student_star, fill =
department_name)) +
  geom_boxplot(alpha = 0.7) +
  theme_minimal() +
  labs(
    title = "Distribución de valoraciones por departamento",
    x = "Departamento",
    y = "Valoración media (student_star)"
  ) +
  theme(legend.position = "none")

leveneTest(student_star ~ department_name, data = datos_filtrados)

anova_model_sinfiltrar <- aov(student_star ~ department_name, data = datos_filtrados)
summary(anova_model_sinfiltrar)

TukeyHSD(anova_model_sinfiltrar)
plot(TukeyHSD(anova_model_sinfiltrar), las = 1, col = "blue")
```

```r
oneway.test(student_star ~ department_name, data = datos_filtrados, var.equal =
FALSE)

# Repetimos a nivel profesor
datos_filtrados_unicos <- datos_filtrados %>%
  distinct(professor_name, .keep_all = TRUE)

ggplot(datos_filtrados_unicos, aes(x = department_name, y = star_rating, fill =
department_name)) +
  geom_boxplot(alpha = 0.7) +
  theme_minimal() +
  labs(
    title = "Distribución de valoraciones por departamento",
    x = "Departamento",
    y = "Valoración media (star_rating)"
  ) +
  theme(legend.position = "none")

leveneTest(star_rating ~ department_name, data = datos_filtrados_unicos)

anova_model <- aov(star_rating ~ department_name, data = datos_filtrados_unicos)
summary(anova_model)

TukeyHSD(anova_model)

plot(TukeyHSD(anova_model), las = 1, col = "blue")
```