Universidad Nacional
de Educación a Distancia

Escuela Técnica Superior de Ingeniería Informática

Predicción de abandono en MOOC mediante aprendizaje automático: revisión sistemática y meta-análisis

Jorge Tenorio Berrío

Director: Emilio Letón Molina
Co-director: Jorge Pérez Martín

Trabajo de Fin de Máster

Máster Universitario
en Ingeniería y Ciencia de Datos

Febrero-2024

Agradecimientos

“Estoy en deuda con usted. Hay en mí muchas cosas que no podrían explicarse sin su generosa y aleccionadora amistad. No intento saldar mi deuda con estas páginas que hoy le ofrezco. Entre mis defectos no está, creo yo, el de no saber ver las cosas como son, sobre todo cuando, como en este caso, son claras como la luz de una bombilla. Yo le envío este libro con otra intención. Cuando las deudas no se pagan porque no se puede, lo mejor es no hablar de ellas y barajar. (...)

La Alcarria es un hermoso país al que a la gente no le da la gana ir. Yo anduve por él unos días y me gustó. (...) Por la Alcarria fui siempre apuntando en un cuaderno todo lo que veía, y esas notas fueron las que me sirvieron de cañamazo para el libro. No vi en todo el viaje nada extraño, ni ninguna barbaridad gorda (...), y ahora me alegro, porque, como pensaba contar lo que hubiera visto (porque este libro no es una novela, sino más bien una geografía), si ahora, al escribirlo, me caigo pintando atrocidades, iban a decir que exageraba y nadie me había de creer. En la novela vale todo, con tal de que vaya contado con sentido común; pero en la geografía, como es natural, ya no vale todo, y hay que decir siempre la verdad, porque es como una ciencia.

Pues bien, (...) esto es todo lo que hay. Poco es; pero, en fin, menos da una piedra. Le mando también una flor que arranqué de una cuneta; la tuve todo este tiempo metida en un libro y ya está disecada. Yo creo que es bonita. Le ruego que acepte usted este regalo que le ofrece, con la mejor intención del mundo, su devoto.”

Viaje a la Alcarria, Camilo José Cela

Resumen

El principal problema de los MOOC (*Massive Open Online Courses*) es su alta tasa de abandono. En los últimos años se han desarrollado gran cantidad de técnicas para predecir abandonos de forma temprana y automatizada. La gran mayoría de ellas basadas en los últimos avances en Inteligencia Artificial, especialmente en técnicas de aprendizaje automático.

En este trabajo se realiza una síntesis cuantitativa del rendimiento de técnicas de aprendizaje automático para la predicción temprana de abandono en MOOC. Para lograrlo se han empleado las técnicas de revisión sistemática y meta-análisis. Se han extraído los datos de 11 artículos recopilados de distintas bases de datos científicas siguiendo los estándares de la guía PRISMA. El análisis se ha realizado utilizando distintos modelos, implementados tanto con un enfoque frecuentista como bayesiano. Además, mediante análisis de subgrupos se ha estudiado la relación de algunas de las características de los estudios con el rendimiento obtenido.

Los resultados indican que este tipo de sistemas son capaces de detectar un alto porcentaje de abandonos. Sin embargo, esto no se puede afirmar al completo debido a la heterogeneidad presente en todos los experimentos.

Palabras clave: MOOC, meta-análisis, aprendizaje automático, modelos bayesianos

Abstract

The main problem with MOOCs (Massive Open Online Courses) is their high drop-out rate. In recent years, a large number of techniques have been developed to predict dropouts early and automatically. Most of them are based on the latest advances in Artificial Intelligence, especially in machine learning techniques.

This work provides a quantitative synthesis of machine learning techniques' performance for early dropout prediction in MOOCs. The systematic review and meta-analysis techniques were used to achieve this. Data were extracted from 11 articles collected from various scientific databases following the PRISMA guide standards. The analysis was carried out using different models, implemented with both frequentist and Bayesian approaches. Furthermore, subgroup analysis was conducted to examine the relationship between certain study characteristics and the resulting performance.

The results indicate that such systems can detect a high percentage of dropouts. However, this cannot be fully confirmed due to the heterogeneity present in all experiments.

Keywords: MOOC, meta-analysis, machine learning, bayesian models

Glosario

***Forest Plot* o diagrama de efectos** Gráfico utilizado para mostrar los resultados del meta-análisis.

***Funnel Plot* o diagrama de embudo** Gráfico utilizado para comprobar el sesgo de publicación en un meta-análisis.

Predictor En estadística, variable que se utiliza como entrada del algoritmo.

ANN *Artificial Neural Net* o Red Neuronal Artificial.

AUC Area Under the ROC Curve.

Clickstream Secuencia ordenada de acciones de un estudiante en una plataforma de aprendizaje.

DL *Deep learning* o Aprendizaje Profundo.

EDM *Educational Data Mining* o Minería de Datos Educativos.

Especificidad Métrica de evaluación que mide el ratio de aciertos en los casos negativos.

GLMM *Generalized Linear Mixture Model*.

GRU *Gated Recurrent Unit*.

LOO *Leave One Out*.

LSTM *Long-Short Term Memory*.

ML *Machine Learning* o Aprendizaje automático.

MOOC *Massive Open Online Course*.

Prior En estadística bayesiana, distribución de los parámetros antes de ver los datos.

PRISMA *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*.

Sensibilidad Métrica de evaluación que mide el ratio de aciertos en los casos positivos (también llamada *recall* o exhaustividad).

SVM *Support Vector Machine* o Máquina Vectores Soporte.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El auge de empresas educativas en línea ha marcado una revolución en la forma en que nos formamos. En la última década, estas plataformas han surgido como pioneras en la democratización de la educación, derribando barreras geográficas y ofreciendo oportunidades de aprendizaje flexibles a una escala global. En 2022 el valor global de este mercado se estimó en más de 28 miles de millones de dólares, y se predice que en 2030 alcance los 360 miles de millones[Research and Markets, 2024]. Fue en 2022 cuando Coursera[1], la empresa más grande en este sector, alcanzó los 118 millones de usuarios y 1401 empleados, obteniendo unos ingresos totales de 523 millones de dólares [United States Securities and Exchange Commision, 2022]. 2U, la empresa propietaria de otra de las grandes plataformas en el sector, edX[2]; alcanzó los 70 millones de usuarios en el mismo año [2U, 2023]. Este fenómeno no se ha dado únicamente en países occidentales, la plataforma china XuentangX[3] superó los 100 millones de usuarios [International Centre For Engineering Education (ICEE), 2022], y en India, la plataforma pública Swayam[4] alcanzó los 24 millones [Times of India, 2023]. El principal acontecimiento que motivó este crecimiento fue la pandemia del Covid-19. Sólo en 2020 aumentó un 50 % el número de alumnos registrados en plataformas de educación online[Shah, 2020].

Estas empresas se han abanderado con un tipo de producto que ha estado ligado a su crecimiento. Son los denominados *Massive Online Open Courses* (MOOC), un tipo de cursos no reglados que se caracterizan por su flexibilidad y sus escasos requisitos de acceso. Estos cursos resultan idóneos para personas con poco tiempo disponible, ya que se adaptan a cualquier horario; y permiten poner a prueba los intereses del estudiante sin apenas coste, puesto que no suelen tener apenas requisitos de matrícula. Esto lo hace un complemento educativo perfecto para una estudiantes y trabajadores.

Este tipo de cursos no son únicamente un asunto de grandes empresas. Otras institu-

[1] `https://www.coursera.org/`
[2] `https://www.edx.org/`
[3] `https://www.xuetangx.com/`
[4] `https://swayam.gov.in/`

ciones, especialmente universidades, se han implicado en la creación de MOOC o incluso desarrollando sus propias plataformas. Un ejemplo es la Universidad Nacional de Educación a Distancia (UNED) y su plataforma UNED Abierta[5]. En la figura 1.1 se puede ver como esta plataforma no ha parado de crecer en los últimos años.

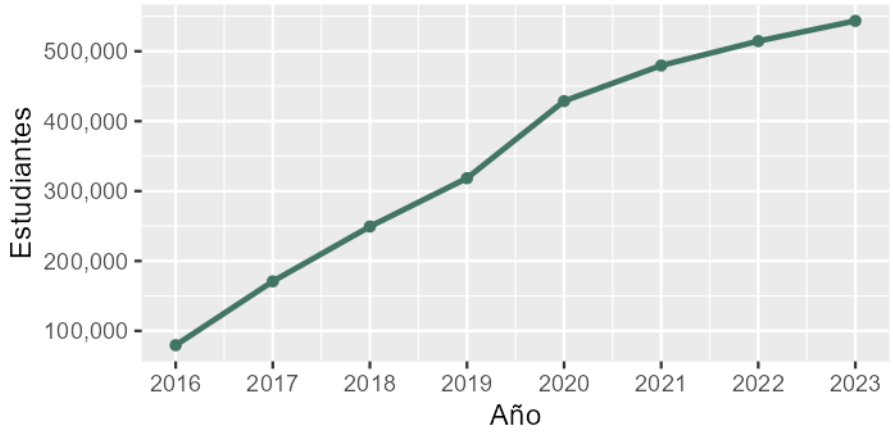

Figura 1.1: Estudiantes registrados en UNED Abierta.

Sin embargo, los MOOC tienen un problema, y es que el porcentaje de alumnos que completan este tipo de cursos es muy bajo. Dependiendo de la fuente y el curso, se estima que la tasa de abandono está en torno al 90 % (ver Gütl et al. [2014] o Floratos et al. [2015]). Las causas de este hecho son múltiples, y van desde la falta del tiempo, dificultades con el tema en cuestión, o actividades poco motivadoras [Gütl et al., 2014]. En Onah et al. [2014] se ha comprobado que hay cierto abandono que es prácticamente ineludible, principalmente provocado por unos requisitos de matrícula prácticamente inexistentes. Un ejemplo son los llamados *recolectores*, cuyo único fin es descargar el contenido del curso. Sin embargo, hay cierta tasa de abandonados que se debe poder evitar, por lo que existe un interés en reducirla mediante la detección temprana e intervención. Esta tarea, debido a su carácter a distancia y masivo, es mucho más difícil que en la educación presencial.

La mayor parte de los enfoques que han tratado de atajar este problema lo han hecho mediante el uso de la Inteligencia Artificial. Este campo ha evolucionado mucho en los últimos años gracias a las técnicas de aprendizaje automático o *machine learning*, que han sido aplicadas en prácticamente todas las áreas del conocimiento. La educación a distancia no ha quedado de lado, puesto que han sido utilizadas para diversas tareas como recomendación de cursos [Liu et al., 2022], detectar alumnos desmotivados [Bhardwaj et al., 2021], predecir el rendimiento académico [Kőrösi and Farkas, 2020], analizar sentimientos en mensajes de foros [Moreno-Marcos et al., 2018], o, en la tarea en la que se centra este trabajo: detectar posibles abandonos [Moreno-Marcos et al., 2020].

Todos estos factores han despertado un interés en la comunidad científica provocando un

[5] `https://iedra.uned.es/`

aumento año tras año en el número de publicaciones en este tema. En la figura 1.2 se muestra el crecimiento del número de publicaciones en este ámbito. Se puede ver que el crecimiento de las publicaciones en predicción de abandono ha estado ligado al aprendizaje automático.

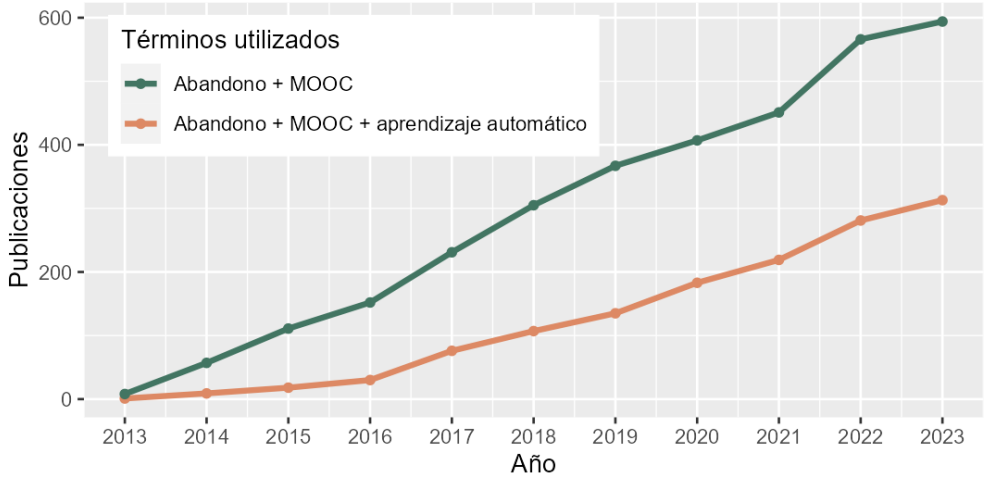

Figura 1.2: Publicaciones por año disponibles en Scopus relacionadas con abandono en MOOC.

En ciencia, cuando el número de estudios de una línea de investigación crece considerablemente se requieren métodos que sinteticen las conclusiones y de esta forma alcanzar el consenso científico.

1.1. Motivación

En este contexto, este trabajo pretende comprobar el estado en que se encuentra la línea de investigación «predicción de abandono en MOOC mediante técnicas de aprendizaje automático» mediante la técnica del meta-análisis. Este trabajo resulta innovador por dos aspectos:

- A pesar de que existan trabajos que sintetizan el estado del arte en esta línea (véase Dalipi et al. [2018] y Moreno-Marcos et al. [2019]). Éstos lo hacen desde un punto de vista cualitativo. La técnica del meta-análisis permite utilizar un enfoque cuantitativo para evaluar el rendimiento global de las técnicas utilizadas en la predicción.

- El uso de la técnica del meta-análisis está muy extendido principalmente en medicina y psicología, pero su uso en ciencia de datos es una línea abierta actualmente (algunos ejemplos en Azeem et al. [2019] o Ayitey Junior et al. [2023]).

1.2. Propuesta y objetivos

Este Trabajo de Fin de Máster pretende analizar en qué medida las técnicas de aprendizaje automático son útiles para la predicción de abandono en MOOC.

Se plantean las siguientes preguntas de investigación:

- **Pregunta de investigación 1**: ¿En qué medida son útiles las técnicas de aprendizaje automático para la detección de abandono en MOOC?
- **Pregunta de investigación 2**: ¿Existen diferencias de rendimiento entre sistemas con distintas características?
- **Pregunta de investigación 3**: ¿Es viable utilizar la técnica del meta-análisis para evaluar el rendimiento global de técnicas de aprendizaje automático?

1.3. Estructura del documento

El resto del documento se estructura del siguiente modo:

- En el capítulo 2 «Estado del arte y marco teórico» se revisa el estado del arte desde dos perspectivas: el problema del abandono en MOOC y la técnica del meta-análisis. En ambos casos se presentarán conceptos teóricos para entender el resto del documento.
- En el capítulo 3 «Materiales y métodos» se revisan los artículos utilizados en el análisis y la forma de obtenerlos. Adicionalmente se definen las técnicas utilizadas en el análisis así como los experimentos realizados
- En el capítulo 4 «Resultados» se presentan los resultados de los experimentos realizados.
- En el capítulo 5 «Discusión» se discuten los resultados obtenidos y las limitaciones del estudio tratando de responder a las preguntas de investigación.
- En el capítulo 6 «Conclusiones y trabajos futuros» se resume el trabajo realizado y proponen futuras líneas de investigación.

Antes de continuar, es importante mencionar que, a pesar de que este trabajo está escrito en español, se ha utilizado el punto como separador decimal por concordancia con el formato de las figuras producidas por los paquetes software utilizados.

Capítulo 2

Estado del arte y marco teórico

En ciencia, para realizar una afirmación fundamentada debe existir consenso científico; y para ello, deben existir múltiples evidencias que confirmen una determinada hipótesis. Esto sólo es posible en líneas de investigación lo suficientemente avanzadas como para que se hayan publicado un número considerable de artículos sobre la misma hipótesis. Además, se requieren métodos para sintetizar los resultados del conjunto de artículos. Existen al menos tres métodos de hacerlo [Borenstein, 2013]:

- **Revisiones narrativas**: escritas por expertos o autoridades. No definen el método por el cual los estudios han sido seleccionados ni cómo elaborar conclusiones, lo que conduce a que estén sesgadas en favor del autor.
- **Revisiones sistemáticas**: se definen reglas de antemano que hace que la síntesis sea reproducible. Tratan de cubrir todos los estudios posibles mediante una síntesis cualitativa.
- **Meta-análisis**: técnica de revisión que trata de resumir la evidencia de forma cuantitativa. Incluye la revisión sistemática de forma implícita para que la obtención de estudios sea reproducible.

Como se ha mencionado en el capítulo anterior, no se ha encontrado ningún trabajo que trate de realizar un meta-análisis en lo que respecta a la predicción de abandono en MOOC con aprendizaje automático. Este capítulo se ha planteado de forma que combina una introducción teórica al resto de la memoria a la vez que presenta los trabajos relacionados. Por lo tanto, el resto del capítulo está estructurado del siguiente modo:

- En la sección 2.1 se describe qué es el meta-análisis y cuáles son los principales métodos para llevarlo a cabo.
- La sección 2.2 se introduce qué es el aprendizaje automático, cómo funciona y para qué sirve.

- En la sección 2.3 se describe cómo se han aplicado las técnicas de aprendizaje automático en la predicción de abandono.
- Por último, en la sección 2.4 se comentan algunos trabajos previos en dos líneas: trabajos que han tratado de sintetizar el estado del arte en predicción de abandono y trabajos que han utilizado el meta-análisis para evaluar otras tareas que utilizan aprendizaje automático.

2.1. Meta-análisis

El objetivo fundamental del meta-análisis es validar una determinada evidencia o hipótesis de forma cuantitativa mediante métodos estadísticos. Esta validación se realiza mediante la síntesis de los resultados obtenidos en los trabajos que analizan esa evidencia.

En esta sección se van a ver cada uno de los elementos que conforman el meta-análisis:

- Un valor que mide la evidencia a validar. Debido a su origen en medicina, esta medida se suele denominar **tamaño o medida de efecto**.
- Una técnica o **modelo** estadístico para llevar a cabo el análisis.
- Una medida de la precisión del análisis, la cual se mide mediante la **heterogeneidad** del conjunto de estudios.
- Un problema importante en los meta-análisis: el **sesgo de publicación**.

Además de estos elementos, también se introduce qué es un **modelo bayesiano** y su aplicación al meta-análisis como alternativa al enfoque tradicional.

2.1.1. Tamaño de efecto

El tamaño o medida de efecto cuantifica la evidencia que se quiere medir. Esta medida puede ser, si hablamos de medicina, la proporción de pacientes curados mediante un medicamento o la incidencia de una determinada enfermedad en la población.

En Borenstein [2013] se mencionan las características que debe cumplir esta medida:

- **Comparable**: debe medir lo mismo en todos los estudios y no debe depender del diseño del estudio.
- **Computable**: debe poderse calcular directamente de la información publicada y no debe requerir re-análisis de los datos.
- Sustancialmente **interpretable**: los investigadores del área deben considerar útil la medida de efecto.

Cada estudio incluido en el meta-análisis debe incorporar su propio valor, que será utilizado para obtener una **medida de efecto global** (también denominada agrupada o resumen) cuyo valor determina la validez de la hipótesis.

Cualquier valor cuantitativo que cumpla estas condiciones es susceptible de considerarse medida de efecto, por lo que pueden existir muchos tipos. En Harrer et al. [2021] las clasifican según el diseño del estudio del que proceden:

- En los diseños de un único grupo se pueden encontrar principalmente medias, proporciones y correlaciones.

- En los diseños con grupo de control se encuentran diferencias de medias y distintos ratios (*risk ratio*, *odds ratio* o ratio de tasas de incidencia).

2.1.2. Modelos

Desde un punto de vista formal, en un meta-análisis se pretende estimar un tamaño de efecto global θ en base a las medidas observada de cada estudio $\hat{\theta}_k$.

El enfoque más simple de calcular el tamaño de efecto global asume que las medidas de efecto radican en una población común homogénea. Por lo que todos los estudios tienen las mismas características y las diferencias entre sus tamaños de efecto están únicamente producidas por el error de muestreo ϵ_k. Este error es más pequeño cuanto más grande es un estudio. Este es el **modelo de efecto fijo** y puede ser definido mediante la siguiente expresión:

$$\hat{\theta}_k = \theta + \epsilon_k$$

La suposición del modelo de efecto fijo es muy optimista para el mundo real, ya que las características de cada estudio siempre variarán de algún modo. El **modelo de efectos aleatorios** tiene en cuenta estas diferencias, asumiendo que no existe un tamaño de efecto real único, si no una distribución de ellos. De este modo, en el modelo efectos aleatorios se asume que existen dos fuentes de variabilidad. Además del error de muestreo ϵ_k, se incluye ζ para tener en cuenta la variabilidad entre estudios. De este modo, la expresión para este modelo queda:

$$\hat{\theta}_k = \theta + \epsilon_k + \zeta$$

En esta expresión subyace una variable aleatoria más: los denominados efectos reales de cada estudio θ_k. Este efecto depende de las características del estudio que no podemos controlar en el análisis. En el modelo de efecto fijo, los efectos reales de cada estudio eran el mismo. Descomponiendo la expresión anterior podemos ver todas las variables del modelo

de efectos aleatorios:

$$\hat{\theta_k} = \theta_k + \epsilon_k$$
$$\theta_k = \theta + \zeta$$

Es común asumir que ϵ_k, y ζ siguen una distribución Gaussiana con media 0 y desviación σ_k y τ respectivamente. De esta forma, surge un modelo jerárquico compuesto por dos distribuciones:

$$\begin{aligned} \hat{\theta}_k &\sim \mathcal{N}(\theta_k, \sigma_k) \\ \theta_k &\sim \mathcal{N}(\theta, \tau) \end{aligned} \tag{2.1}$$

Esta expresión es importante ya que se utilizará en el capítulo siguiente.

2.1.3. Heterogeneidad

Resulta ingenuo pensar que todos los estudios van a compartir efectos similares, en la realidad siempre existirá cierta variabilidad entre los resultados de cada estudio que es importante cuantificar. Como se ha visto, esta variabilidad puede tener dos fuentes: la varianza intra-estudios, fruto del error de cada estudio, y la varianza inter-estudios o **heterogeneidad**, fruto de las diferencias entre las características de cada estudio (σ_k y τ en la expresión anterior, respectivamente).

Estimar la heterogeneidad es uno de los objetivos del meta-análisis ya que está directamente relacionada con el error asociado a la estimación del efecto agrupado, y por lo tanto a la validación de la evidencia. Se pueden considerar cuatro métodos para medir la heterogeneidad. Tres de ellos son índices cuantitativos, el test Q, el estadístico I^2 y la varianza inter-estudios τ^2; y uno gráfico, el denominado *forest plot*.

Antes de continuar con estos métodos es necesario mencionar algo importante. Que un meta-análisis tenga una heterogeneidad muy alta no tiene por qué ser algo negativo, ya que se pueden buscar patrones en esa heterogeneidad y probar nuevas hipótesis. Entre otros, existen dos métodos para analizar la heterogeneidad. El **análisis de subgrupos** trata de agrupar los efectos en subgrupos de estudios en función de sus características, para luego comprobar si existen diferencias entre ellos. El **análisis de influyentes** busca aquellos estudios que más impacto tienen en el meta-análisis. De este modo, es posible buscar características en esos estudios que no se apreciaron inicialmente, como errores o motivos para excluirlo.

El test Q de Cochran

Los dos primeros índices siguen la misma filosofía y están estrechamente relacionados. Tratan de buscar qué parte de la variabilidad entre efectos no está justificada mediante la

variabilidad intra-estudios. El primer estadístico es la **Q de Cochran**. Mide la desviación de los efectos observados $\hat{\theta}_k$ del efecto agrupado estimado $\hat{\theta}$ mediante la suma ponderada de cuadrados. Se utiliza el inverso de la varianza de cada estudio como peso w_k, de esta forma aquellos estudios con una alta precisión tendrán más aporte en este índice.

$$Q = \sum_{k=1}^{K} w_k(\hat{\theta}_k - \hat{\theta})^2$$

El índice Q sigue aproximadamente una distribución χ^2 con $K-1$ grados de libertad (K = número de estudios) si las diferencias en el tamaño de efecto se deben únicamente al error de muestreo. Por ello, se suele utilizar un test χ^2 para comprobar estas diferencias. Si el valor del test es significativo (≤ 0.05) indicará que las diferencias entre estudios son sustanciales.

La Q de *Cochran* es muy importante, ya que es muy utilizada en la bibliografía y otros índices se basan en ella. Sin embargo, según Harrer et al. [2021] tiene el inconveniente de que su poder estadístico está influido en gran medida del tamaño del meta-análisis debido a que el valor de Q crece con el número de estudios.

Estadístico I^2 de Higgins y Thompson

El **estadístico** I^2 mide la proporción de variabilidad que no está causada por el error de muestreo. Está directamente relacionado con la Q de Cochran, pero no es sensible a cambios en el número de estudios. Mide el exceso del valor observado de Q respecto a su valor esperado si no hubiese heterogeneidad (que es $K-1$). Su valor se da en términos de tanto por ciento.

$$I^2 = \frac{Q-(K-1)}{Q} \times 100 \tag{2.2}$$

Es muy típico publicar la heterogeneidad de un meta-análisis mediante este índice. Sin embargo, su valor depende en un alto grado en la precisión de los estudios. Por lo tanto, si el tamaño de los estudios aumenta, I^2 tiende a 100 % [Rücker et al., 2008].

Varianza inter-estudios τ^2

La **varianza inter-estudios** τ^2 ya fue introducida en la sección anterior, cuando se mencionaba que θ_k seguía una distribución gaussiana. La varianza de esa distribución se suele denominar τ^2, e indica la variabilidad en el conjunto de tamaños de efectos reales θ_k. Existen varios métodos para calcular este parámetro.

La varianza inter-estudios τ^2 tiene el inconveniente de que puede ser difícil de interpretar si no se tienen en cuenta las varianzas intra-estudios y la medida de efecto global.

Forest plot

Un gráfico que se suele publicar muy frecuentemente en meta-análisis es el ***forest plot***, también conocido como **diagrama de efectos**. Este gráfico ofrece un vistazo general a los resultados del meta-análisis, y es un buen método para comprobar la heterogeneidad.

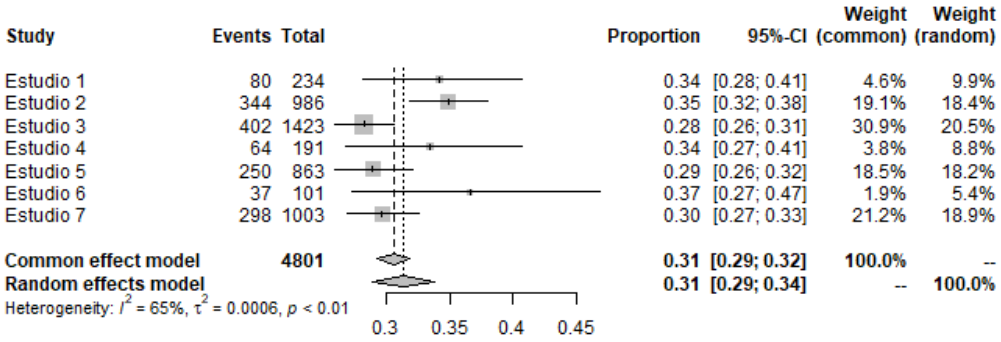

Figura 2.1: Ejemplo de *forest plot*

En el eje X de un *forest plot* se representan los tamaños de efecto y sus errores asociados. Se posicionan de tal forma que los distintos estudios quedan en el eje Y. Dependiendo del tipo de modelo utilizado se puede incluir alguna forma geométrica alrededor de la estimación representando el peso del estudio en el meta-análisis. También se suele incluir la estimación y error asociado a la medida de efecto global. En la figura 2.1 se puede ver un ejemplo de este tipo de gráficos. Se incluyen 6 estudios de ejemplo con sus efectos estimados. Además, en la parte inferior se incluyen las estimaciones de la medida de efecto global tanto con el modelo de efecto fijo así como el de efectos aleatorios. Normalmente, en un *forest plot* se suelen incluir otros datos de cada estudio, como un identificador, la medida de efecto exacta, etc.

Una forma de estimar la heterogeneidad mediante este tipo de gráficos es visualizando las líneas discontinuas verticales, la cuales representan la medida de efecto global estimada a lo largo de los estudios que conforman el meta-análisis. Si esta línea se encuentra dentro de las barras de error, la variabilidad se puede achacar a la varianza intra-estudios. En el ejemplo, tanto la línea del modelo de efecto fijo como la del modelo de efectos aleatorios no cortan las barras de error de los estudios 2 y 3. Esto indica cierta heterogeneidad, tal y como se puede ver en el valor de I^2.

2.1.4. Sesgo de publicación

Existe cierta evidencia de que los resultados de un artículo afectan a su publicación. El **sesgo de publicación** se produce por el hecho de que los estudios con resultados no

significativos o negativos suelen tener una menor tasa de publicación. Esto puede provocar que los estudios incluidos en el meta-análisis no sean representativos. Según Page et al. [2021a], este problema se da también a nivel de resultado. Pues dentro de un estudio es más probable que sólo se publiquen los resultados positivos.

Existen métodos estadísticos para analizar el riesgo de sesgo. Aquí se expondrá la forma tradicional de evaluarlo: el **gráfico de embudo** o ***funnel plot***.

Un gráfico de embudo se basa en lo siguiente: debería ser más probable que dos resultados sean similares cuanto más precisos sean sus resultados. Para comprobarlo se representan los tamaños de efecto de cada estudio contra su error (ϵ_k). De esta forma, los estudios con mayor precisión deberían mostrarse cercanos. Cuanto menor es la precisión del estudio más deberían alejarse. Por lo tanto, si no existiese ningún sesgo, la disposición de los estudios debería tener forma triangular. En caso de que no sea simétrico, se puede entender que existe algún sesgo de publicación. En la figura 2.2 se muestra un ejemplo de este gráfico. Se puede ver que existe un vacío en la esquina inferior izquierda, esto indica que tales resultados no se han obtenido. Como se puede ver, se suele invertir el eje y, por lo que el resultado es en forma de triángulo o embudo, de aquí su nombre.

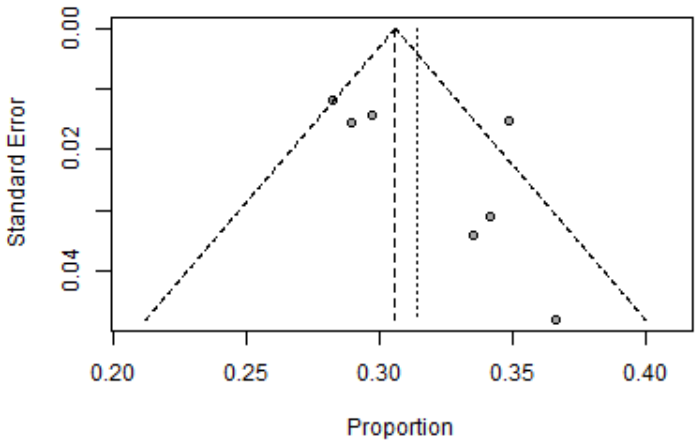

Figura 2.2: Ejemplo de gráfico de embudo.

2.1.5. Enfoque frecuentista vs bayesiano

En estadística existen dos enfoques cuya principal diferencia radica en el concepto filosófico de probabilidad. Por un lado están los frecuentistas, el enfoque más clásico. Para estos, la probabilidad de un suceso se fundamenta en su aparición en un experimento realizado un número muy alto de veces. Por otro lado, el enfoque bayesiano tiene en cuenta la probabilidad como una medida de incertidumbre o credibilidad Gelman et al. [1995]. Esto permite realizar

afirmaciones en experimentos en los que no es posible asumir la repetición del experimento (por ejemplo, la probabilidad de una hipótesis sobre el origen del universo).

Para reflejar esta incertidumbre, un modelo bayesiano siempre tiene asociada una distribución de probabilidad a cada parámetro, la cual define el conocimiento subjetivo que se tiene sobre el valor del mismo. Esta distribución tiene sus propios parámetros como media, moda o varianza. En el modelo frecuentista se asume que existe un valor exacto o puntual para los parámetros del modelo y que puede ser calculado.

Dicho de otro modo, en el enfoque frecuentista los parámetros del modelo son fijos, y los datos son tratados como variables aleatorias; mientras que en el modelo Bayesiano se hace al revés. Esto afecta directamente al concepto de intervalo. Un **intervalo de confianza** en el enfoque frecuentista permite afirmar que si repetimos el experimento 100 veces, un tanto por ciento de ellas el valor obtenido estará en ese rango. Los **intervalos de credibilidad** bayesianos, permiten afirmar que se conoce que el valor está en ese rango con una certidumbre del tanto por ciento.

Introducción al modelo Bayesiano

El enfoque bayesiano se fundamenta en el Teorema de Bayes, el cual establece una igualdad entre tres elementos.

$$p(\text{parámetros}|\text{datos}) = \frac{p(\text{datos}|\text{parámetros}) \times p(\text{parámetros})}{p(\text{datos})}$$

El primer elemento es la **distribución a posteriori** $p(\text{parámetros}|\text{datos})$. Establece la función de densidad de probabilidad de los parámetros dados los datos. Es la distribución que se pretende calcular y que define los posibles valores del parámetro que se pretende obtener.

En el otro lado se tiene la denominada **verosimilitud** $p(\text{datos}|\text{parámetros})$. Se representa como la probabilidad de los datos dados los parámetros; sin embargo, no puede ser considerada una función de densidad ya que es una función de los parámetros, no de los datos. Por ello se pueden encontrar otras representaciones como $L(\text{parámetros}|\text{datos})$ (la L proviene del término utilizado en inglés *likelihood*).

La **distribución a priori** o comúnmente llamados *prior* es el elemento más controvertido del modelo bayesiano y que marca la gran diferencia con los modelos frecuentistas. Dicha distribución a priori está representada por la función de densidad de probabilidad de los parámetros $p(\text{parámetros})$ y define el conocimiento previo sobre los parámetros antes de implementar el modelo. En la literatura se definen tres posibles tipos de *priors* dependiendo de la información que dispongamos. Usar uno u otro debe estar debidamente justificado antes de la implementación del modelo.

Se usan *priors* **no informativos** cuando no se dispone de ninguna información, es difícil de construir o se pretende que afecte lo mínimo al cálculo. Se suelen utilizar funciones de

densidad muy difusas, se dice que estos *priors* dejan hablar a los datos. Un prior **débilmente informativo** contiene la suficiente información para delimitar el resultado dentro de un rango de valores, pero siguen sin aportar información. Por último, los prior **informativos** incluyen la información de anteriores estudios o conocimiento experto mediante una distribución claramente definida.

Por último, se tiene la probabilidad de los datos p(datos) la cual se considera una constante por lo que su papel es menos relevante.

Métodos de cálculo de parámetros

El principal inconveniente del enfoque bayesiano es su complejidad. Para algunos modelos sencillos es posible obtener una solución analítica. Sin embargo, a medida que el modelo se hace más complejo también lo hace el álgebra para resolverlo. Por ello se han desarrollado métodos para hacerlo mediante simulación. Esto facilita el uso de este tipo de modelos aunque conlleva un mayor coste computacional.

El método más común de simulación es *Markov Chain Monte Carlo* (MCMC). Este método utiliza una muestra aleatoria finita de parámetros obtenidos de una distribución aproximada (método de Monte Carlo). Esto se realiza secuencialmente, de forma que cada muestra depende del anterior valor (Cadena de Markov).

Este método tiene dos dificultadesGelman et al. [1995]:

- **Falta de convergencia**: se da cuando no se realizan las suficientes iteraciones para que las simulaciones sean representativas de la función objetivo.
- **Autocorrelación**: cuando no existen suficientes muestras independientes provocando que la inferencia sea menos precisa.

En ambos casos, existen métodos para evaluarlos que se verán en el siguiente capítulo.

Enfoque bayesiano en meta-análisis

Durante mucho tiempo, se ha utilizado principalmente el enfoque frecuentista en meta-análisis. Sin embargo, se ha podido ver que el enfoque bayesiano tiene algunas ventajas. Un ejemplo lo tenemos en Thompson and Semma [2020], en el que se comparan ambos enfoques en un meta-análisis en psicología. Concluye que los resultados son equivalentes, y que las grandes diferencias radican en las afirmaciones de probabilidad que el modelo bayesiano permite.

Las principales ventajas del enfoque bayesiano dentro del meta-análisis es que puede ser más adecuado cuando se disponen de pocos artículo gracias al uso de la información a priori Borenstein [2013]. También tiene algunos inconvenientes, como por ejemplo que no existe una estimación bayesiana de los estadístico Q e I^2, por lo que la estimación de heterogeneidad queda limitada a τ^2.

2.2. Aprendizaje automático

Tal y como se mencionó más arriba, este trabajo se va a centrar en métodos que utilizan aprendizaje automático. Según Géron [2022], **aprendizaje automático** es la ciencia de programar ordenadores de forma que pueden aprender de los datos. Es una técnica especialmente útil en problemas en los que la programación tradicional no alcanza una solución óptima, o es necesario un algoritmo muy complejo con una gran cantidad de reglas para alcanzarla.

El proceso de aprender de los datos se suele denominar entrenamiento, y el conjunto de datos para llevarlo a cabo **conjunto de entrenamiento**. Una vez se ha realizado el entrenamiento, se suele analizar su rendimiento mediante una **métrica de evaluación**. Además, el algoritmo debe de poder generalizar a otros datos fuera del conjunto de entrenamiento, por lo que la evaluación se suele realizar en un conjunto de datos distinto al de entrenamiento, denominado **conjunto de evaluación**.

Los sistemas, algoritmos o modelos de aprendizaje automático se pueden clasificar dependiendo de la supervisión que reciban durante el aprendizaje:

- Aprendizaje **supervisado**: cuando la solución deseada de cada dato está incluida en el conjunto de entrenamiento, la cual se suele denominar etiqueta. El resto de características del conjunto de datos denominan variables predictoras o *predictors*. Si la etiqueta es numérica se dice que es un problema de regresión, en cambio, si es categórica se denomina clasificación.

- Aprendizaje **no supervisado**: cuando no se incluye una etiqueta entre los datos de entrenamiento. Las posibles tareas son más variadas dentro de este enfoque: son el agrupamiento, detección de anomalías, reducción de dimensionalidad, etc.

- Aprendizaje **semi-supervisado**: son algoritmos que pueden lidiar con conjuntos de datos parcialmente etiquetados

- Aprendizaje **por refuerzo**: este es un caso muy distinto al resto en el que el sistema de aprendizaje (agente) observa el entorno, toma acciones y recibe recompensas.

Dentro de cada uno de los tipos de aprendizaje existen múltiples tipos de algoritmos. Más adelante se revisarán aquellos aplicados al problema en cuestión. Sin embargo, antes de continuar es necesario hablar a un tipo de algoritmo que da lugar a un campo entero dentro del aprendizaje automático: las redes neuronales artificiales y el *Deep Learning*.

Una **Red Neuronal Artificial** (*Artificial Neural Net* o ANN) es un tipo de algoritmo de aprendizaje automático inspirado en las redes de neuronas biológicas del cerebro [Géron, 2022]. Pese a que se desarrollaron múltiples modelos a lo largo del siglo XX, el algoritmo que realmente marcó la diferencia fue el Perceptrón Multicapa. Este algoritmo es un tipo de ANN compuesto por al menos tres capas de neuronas: una capa de entrada, una capa oculta

y una capa de salida. Las neuronas de una capa se conectan con las neuronas de la capa siguiente. Los datos se sitúan en la denominada capa de entrada, la cual debe tener tantas neuronas como *predictors* tengan los datos de entrada. De igual manera, la capa de salida tantas como resultados requiera el problema. En la figura 2.3 se puede ver un esquema de esta arquitectura.

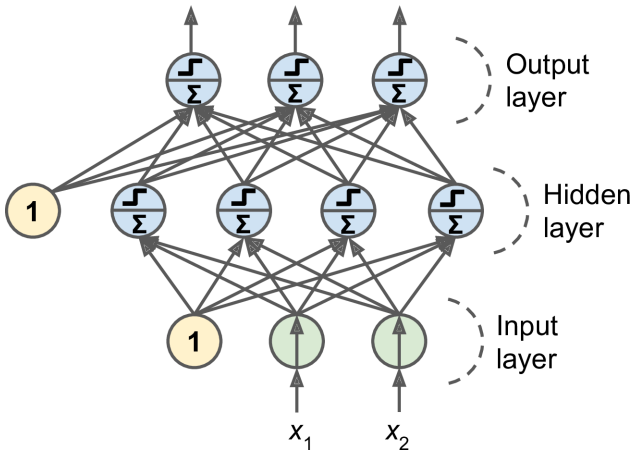

Figura 2.3: Esquema del perceptrón multicapa. Fuente: Géron [2022].

La forma en que funcionan este tipo de algoritmos es realizando una operación no lineal promediada por una serie de parámetros en cada neurona de las capas ocultas y salida. El entrenamiento de este algoritmo trata de estimar estos parámetros.

La red neuronal presentada utiliza el tipo de capa más sencillo, las denominadas *Fully Conected Layer*. Pero existen otros tipos de capas que dan lugar a otras familias de redes neuronales, como las redes neuronales convolucionales [LeCun et al., 2015] o las redes neuronales recurrentes [Goodfellow et al., 2016].

Cuando una ANN contiene una cantidad considerable de capas ocultas se denomina Red Neuronal Profunda (*Deep Neural Net* o DNN). Estos algoritmos pueden lograr un gran rendimiento, pero requieren de una gran cantidad de datos y de capacidad de cómputo. Está demostrado que con suficiente profundidad se pueden abstraer cualquier característica de un conjunto de datos. El área del conocimiento que se encarga del estudio de estos algoritmos se denomina *Deep Learning*.

2.3. Predicción de abandono en MOOC con aprendizaje automático

La tarea de predicción de abandono se engloba dentro de lo que se conoce como **Minería de Datos Educacionales** (*Educational Data Mining* o EDM), cuya tarea es la de analizar y encontrar patrones de datos para ayudar a la toma de decisiones [Alhothali et al., 2022]. En este caso, se trata de tomar cierta información registrada en las plataformas de aprendizaje

para predecir si el estudiante abandona o no en cierto momento del curso. Por lo tanto, entre los tipos de problemas de aprendizaje automático vistos en la sección anterior, se puede definir dentro del aprendizaje supervisado como un problema de clasificación binaria [Prenkaj et al., 2021]. Las predicciones realizadas por estos sistemas se pueden utilizar para tomar acciones proactivas para mejorar la experiencia de aprendizaje del estudiante y establecer estrategias de intervención, por lo tanto existe un interés general para que esta predicción se realice cuanto antes [Alhothali et al., 2022].

Esta tarea parece sencilla: se toma la información de los estudiantes en forma de *predictors* y una etiqueta asociada a si abandona o no en cierto momento, se entrena un algoritmo de clasificación y se evalúa utilizando ciertas métricas. Sin embargo, este problema tiene dos características que hacen que la cantidad de posibles soluciones sea muy grande. Estas son:

1. No existe un consenso en la comunidad en cuanto a la **definición de abandono**.
2. Las características del estudiante requieren un **modelado** adicional del que tampoco existe un estándar.

En los siguientes apartados, se verán estos dos problemas además de completar el resto de elementos para definir el problema de clasificación.

2.3.1. Definición de abandono

Aunque parezca difícil de creer, no existe un consenso en la comunidad científica en cuanto a qué significa abandono [Alhothali et al., 2022]. Por lo que cada autor lo define en función de los intereses del docente, la institución o la plataforma. Algunos estudios consideran que un estudiante ha abandonado si ha estado inactivo durante un periodo de tiempo, por ejemplo Chen et al. [2019], Jin [2023]. Otros, si la nota final es 0 [Chi et al., 2023] o si completa un tanto por ciento del curso [Dass et al., 2021].

Puesto que no existe un criterio común, algunos autores han tratado de sintetizar en categorías los distintos tipos de definiciones [Dass et al., 2021, Prenkaj et al., 2021]. En este trabajo se ha contemplado la categorización realizada por Nagrecha et al. [2017], en la que utilizan únicamente tres categorías:

1. Falta de **participación**: si el estudiante no está activo durante un periodo determinado. Por ejemplo: no interactuar con la plataforma durante un período de tiempo o no presentar tareas.
2. Objetivos de **aprendizaje** no cumplidos: si el estudiante no logra algunas de las metas del curso. Como no obtener el certificado, no obtener cierta nota o no acabar cierta cantidad de módulos
3. Enfoques **híbridos**: mezclan las dos categorías anteriores.

2.3.2. Modelado de las características del estudiante

Los sistemas de predicción de abandono utilizan información de los estudiantes extraída de la plataforma de aprendizaje. Esto incluye información de muchos tipos, como la información demográfica del registro y la matrícula, los datos de tareas y de exámenes, los datos de interacción con la plataforma, los mensajes en los foros, etc. En Alhothali et al. [2022] se clasifica esta información dependiendo del momento en el que se generan:

- Las características **pre-curso** se refiere a la información que se genera durante la matrícula y el registro. Esto incluye información demográfica y académica del estudiante, e información del curso.
- Características **intra-curso** (*in-course*): es la información que se genera mediante la interacción del estudiante con la plataforma. Estas características se suelen denominar comportamiento de aprendizaje, e incluye interacción con la plataforma, mensajes en foros, interacción con elementos multimedia, etc.
- Características **post-curso**: información relacionada con los resultados finales.

Y, ¿de qué forma son utilizadas cada una de estas características en los sistemas de predicción? Las características post-curso no se deben utilizar como *predictors*, en todo caso deben ser utilizadas para definir la etiqueta al estar relacionadas con los resultados. Las otras dos se han utilizado de formas distintas en la bibliografía. Por ejemplo Deeva et al. [2022] y Şahin [2021] utilizan únicamente datos de comportamiento. En Panagiotakopoulos et al. [2021] combina datos demográficos, datos de evaluaciones parciales con datos de comportamiento como número de mensajes en el foro o tiempo dedicado al curso. Incluso hay estudios que utilizan únicamente características pre-curso y obtienen buenos resultados. Por ejemplo, Khoushehgir and Sulaimany [2023] utiliza teoría de grafos sobre una red bipartita con únicamente la información de matrícula. En definitiva, no existe una solución única.

El problema surge del tratamiento de las **características de comportamiento**. Esta información es obtenida de los registros (*logs*) de las plataformas. Estos registros son ficheros que almacenan las acciones que se llevan a cabo en la plataforma. Al menos contienen la siguiente información:

- *Timestamp* o momento temporal de la acción.
- Un identificador del estudiante que ha propiciado la acción y del curso en el que está matriculado.
- Evento que ha sucedido en la plataforma.

Cada plataforma utiliza su propia estructura de archivo y su propio conjunto de eventos. En la figura 2.4 se muestra un ejemplo de este tipo de archivos obtenido de Feng et al. [2019].

Se puede ver que se compone de un identificador de curso-estudiante, el momento temporal de la acción, la fuente donde se origina la acción, el evento que sucede y el objeto al que se accede. Todos los eventos de este ejemplo son de acceso a recursos: `navigate` lo hace a otra sección del curso, `access` accede a recursos que no son vídeos o tareas y `problem` accede a tareas.

enrollment_id	time	source	event	object
1	2014-06-14T09:38:29	server	navigate	Oj6eQgzrdqBMlaCtaq1IkY6zruSrb71b
1	2014-06-14T09:38:39	server	access	3T6XwoiMKgol57cm29Rjy8FXVFcIomxl
1	2014-06-14T09:38:39	server	access	qxvBNYTfiRkNcCvM0hcGwG6hvHdQwnd4
1	2014-06-14T09:38:48	server	access	2cmZrZW2h6II91itO3e89FGcABLWhf3W
1	2014-06-14T09:41:49	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:41:50	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:42:28	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf
1	2014-06-14T09:42:30	browser	problem	RMtgC2bTAqEeftenUUyia504wsyzeZWf

Figura 2.4: Ejemplo de datos de comportamiento almacenados en archivos de registro. Fuente: Conjunto de datos KDD Cup 2015

Lo más importante de este tipo de características es que generan una serie temporal ligada a cada estudiante (conocida en la bibliografía como *clickstream*) y que para que sea procesada por los algoritmos de aprendizaje automático es necesario modelarla. Sin embargo, tal y como mencionan en Dalipi et al. [2018], no existe un estándar para hacerlo. Por ejemplo, en Li et al. [2016] generan una matriz para cada tipo de evento con los conteos semanales de cada alumno. Qiu et al. [2019] también utiliza una matriz, pero en este caso una por estudiante y conteos de ciertos eventos en una ventana temporal de longitud variable. En Panagiotakopoulos et al. [2021] utilizan los conteos de los eventos de comportamiento en cada fase y los unen en un vector unidimensional junto con otras características demográficas.

2.3.3. Algoritmos

En Prenkaj et al. [2021] se citan los principales algoritmos para predicción de abandono utilizados en la bibliografía para predicción de abandono. Entre los algoritmos de aprendizaje automático se suele usar:

- **Modelos de regresión**: son un método para analizar la relación entre dos o más variables mediante una función. Se consideran más un modelo estadístico que un algoritmo de aprendizaje automático. Dentro de los modelos de regresión se encuentran múltiples tipos, dependiendo de las suposiciones de la relación: lineal, logística, *Poisson*, polinómica entre otros.

- **Árboles de decisión**: son un tipo de modelos que tratan de definir los patrones en los datos mediante un conjunto de nodos y aristas. En cada nodo se marca una condición que divide el conjunto de datos y se asocia con otros nodos mediante aristas. La estructura resultante no incluye ciclos y dispone de un único nodo inicial por lo que

toma una forma de árbol. Existen múltiples algoritmos para crear un árbol de decisión, algunos de los más típicos son ID3 o C4.5.

- ***Naive Bayes***: son una familia de modelos que se fundamentan en el Teorema de Bayes. Se denominan *naive*, *naif* o ingenuos debido a que toman una serie de suposiciones que simplifican mucho el problema. Pese a que no son realistas suelen funcionar bastante bien.

- **Máquinas Vectores Soporte (SVM)**: son un tipo de modelo de *Machine Learning* que tratan de buscar el hiperplano que separa dos conjuntos de datos maximizando la distancia entre ellos. Este tipo de modelos permite el uso de las funciones *kernel* para transformar el espacio de los datos y conseguir separar conjuntos de datos más complejos.

- **Técnicas de *ensemble***: son un conjunto de técnicas para combinar varios modelos más pequeños. Entre estas posibles técnicas se puede destacar el *bagging*, un método que combina los modelos en paralelo; o el *boosting*, que lo hace de forma secuencial. Existen modelos muy típicos que utilizan estas técnicas como ***Random Forest*** o ***Adaboost***.

Dentro del área del *Deep Learning*, se pueden destacar:

- **Redes neuronales convolucionales (CNN)**: red neuronal que se ha usado principalmente en clasificación de imágenes. Este tipo de redes primero detectan características de bajo nivel y las combina para formar características más complejas. Para conseguirlo combina dos tipos de capas ocultas: las capas convolucionales buscan patrones pequeños mientras que las capas de *pooling* las agrupan seleccionando un subconjunto de ellas [James et al., 2013].

- **Redes Neuronales Recurrentes (RNN)**: tienen en cuenta distintas versiones de un mismo conjunto de datos. Son útiles para procesar datos en secuencia. Es decir, tienen en cuenta cierta correlación temporal. Para su funcionamiento hacen que los datos de salida de la red en cierto instante t se usen como entrada de la red en el instante $t + 1$. Este tipo de arquitecturas tenían el problema de que no tenían una memoria a largo plazo, que se ha solventado con arquitecturas más complejas como LSTM o GRU. Especialmente las primeras también han sido muy utilizadas en el problema de predicción de abandono.

2.3.4. Métricas de evaluación

Para analizar cuantitativamente el rendimiento de un algoritmo se requiere de métricas de evaluación. En Alhothali et al. [2022] mencionan que las principales métricas utilizadas en

el problema de predicción de abandono son: exactitud (o *accuracy*), especificidad, precisión, sensibilidad (exhaustividad o *recall*), medida F1 y AUC (*Area Under de ROC Curve*).

Una matriz de confusión recoge los resultados de una predicción en función de 4 medidas: positivos verdaderos, falsos positivos, negativos verdaderos y falsos negativos. Por sus siglas en inglés se suele usar TP (*true positive*), FP (*false positive*), TN (*true negative*) y FN (*false negative*). A continuación, se van a definir las métricas usadas en este problema mediante estas medidas excepto AUC, cuya definición se puede encontrar en Faraggi and Reiser [2002] y no se va a utilizar en este estudio.

La **exactitud** define la tasa global de aciertos sobre el total de predicciones:

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN}$$

La **especificidad** define la proporción de negativos predichos por el modelo:

$$\text{especificidad} = \frac{TP}{TP + FN}$$

La **precisión** define la tasa de acierto entre todas las predicciones positivas:

$$\text{Precision} = \frac{TP}{TP + FP}$$

La **exhaustividad** o **sensibilidad** define la proporción de positivos predichos por el modelo:

$$\text{Exhaustividad} = \frac{TP}{TP + FN}$$

La **medida F1** es una medida de resumen entre precisión y exhaustividad, y se define como media armónica de las dos:

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{exhaustividad}}{\text{precision} + \text{exhaustividad}}$$

2.4. Trabajos previos

Puesto que no se han encontrado trabajos que realicen trabajos similares a este, se han explorado trabajos previos en dos vías. Por un lado, aquellos que tratan de resumir el estado del arte de la predicción de abandono en MOOC mediante otras técnicas. Y por el otro lado, aquellos que utilizan la técnica del meta-análisis en tareas que utilizan aprendizaje automático. En los siguientes apartados se mencionan estos trabajos.

2.4.1. Resúmenes del estado del arte de la predicción de abandono

Como se ha mencionado anteriormente, el número de publicaciones en esta línea ha crecido considerablemente en los últimos años, lo que ha dado lugar que ya se hayan realizado algunas revisiones. En este trabajo se han encontrado cuatro:

1. En Dalipi et al. [2018] se realiza una revisión que incluye 25 estudios de predicción de abandono en MOOC. Se fijan principalmente en los algoritmos y *predictors* utilizados. En el artículo proponen algunas posibles causas de la alta tasa de abandono, y mencionan algunos retos pendientes y recomendaciones. Un dato a tener en cuenta es que es una revisión narrativa, no existe un procedimiento reproducible para la obtención de los estudios. Concluye mencionando algunos retos a los que se enfrenta esta línea: datos insuficientes, gestionar grandes cantidades de datos no estructurados, variabilidad en los datos, clases desbalanceadas, disponibilidad de conjuntos de datos públicos, falta de un estándar de creación y representación de los datos *clickstream*, y otros retos asociados al calendario del estudiante.

2. En Moreno-Marcos et al. [2019] se realiza una revisión sistemática en la que se incluyen 88 artículos sobre predicciones en MOOC (no sólo abandono). En el análisis se fijan en: las características de los MOOC mencionados en cada artículo, la tarea de predicción que realizan (la gran mayoría tratan de predecir el abandono), las *predictors*, los modelos y las métricas que utilizan. Llegan a la conclusión de que es urgente que se incorporen nuevas características de los estudiantes y algoritmos de aprendizaje automático, y que se deben explorar nuevas tareas de predicción y su relación con las actuales.

3. En Prenkaj et al. [2021] realizan una revisión narrativa de predicción de abandono en varios tipos de cursos en línea, entre los que se incluyen los MOOC. El artículo se centra en analizar el campo de conocimiento, los datos utilizados, el modelado de los datos, los algoritmos y las métricas de evaluación de cada estudio. Concluye que se necesitan estándares de evaluación, que se deben explorar los modelos de *deep learning* para predicción temprana y que los modelos deben considerar la duración de las fases de los cursos.

4. En Alhothali et al. [2022] realizan una revisión sistemática, esta vez en predicción de resultados en varios tipos de cursos. Hace un análisis de algunas características de los estudios muy similares a los anteriores. Entre estas están el tipo de curso, la plataforma en la que se publica, conjunto de datos, el tipo de *predictors*, algoritmos y métricas que utiliza, y los resultados que tratan predecir. Concluye que:

    - No hay consenso de definición de abandono.

- La mayoría de los MOOC revisados no obligan a participar en los cursos. Por lo que hay una gran disparidad entre los alumnos que entran por curiosidad y aquellos con intención de acabarlo.
- La mayoría de los estudios proponen soluciones de predicción a final de curso. Sólo algunos proponen la predicción temprana y publican el momento exacto de predicción. Esto dificulta la comparativa.
- Las representaciones del estudiante producidas por el tipo de modelado son muy dispersas e inútiles para comparar comportamientos de usuario.
- Limitaciones en los datos como instancias multi-valuadas o clases desbalanceadas.
- Escasa calidad de los datos de estudiantes que sólo se inscriben por curiosidad.
- Limitaciones éticas.

2.4.2. Meta-análisis para evaluar el rendimiento de tareas de aprendizaje automático

Puesto que no existen meta-análisis en predicción de abandono, se han explorado algunos artículos que utilizan esta técnica pare evaluar el rendimiento general de sistemas de aprendizaje automático.

Como se ha mencionado, el meta-análisis es una técnica muy común en medicina, por lo que también podremos encontrar estudios de este tipo aplicados al uso de técnicas de aprendizaje automático para resolver problemas médicos. Un ejemplo se puede ver en Nindrea et al. [2018]. En este estudio utilizan el meta-análisis con estudios que utilizan aprendizaje automático para calcular el riesgo de cáncer de mama. Utiliza la sensibilidad y la especificidad en un tipo de meta-análisis denominado de diagnóstico. Otro ejemplo puede ser Araújo et al. [2023] que lo aplica en predicción de intoxicaciones en tratamientos de cáncer de cabeza y cuello mediante aprendizaje automático. En este caso, utilizan AUC como medida de efecto.

Fuera del ámbito de la medicina, también se han encontrado algunos trabajos, aunque son mucho más escasos. En Azeem et al. [2019] analizan 15 estudios en detección de *code smell*. Prueba tanto el modelo de efecto fijo como de efectos variables con la medida F1 como medida de efecto. En Ayitey Junior et al. [2023] lo utilizan en predicciones de mercado *Forex*. Y en Ahmadi et al. [2022] para monitorización/predicción de aguas subterráneas. Estos dos últimos casos están menos relacionados con este trabajo al ser problemas de regresión.

Capítulo 3

Materiales y métodos

Tal y como se comentó en el capítulo anterior, la técnica del meta-análisis requiere que previamente se realice una revisión sistemática. Por lo tanto, este capítulo está dividido en dos grandes secciones. En la sección 3.1 se explica cómo se ha realizado la revisión sistemática y se comentan los datos obtenidos. Posteriormente, en la sección 3.2 se describen las técnicas y herramientas que se han utilizado para llevar a cabo el meta-análisis y se definen los experimentos realizados.

3.1. Revisión sistemática

En una revisión sistemática es vital definir el proceso de obtención de estudios de tal forma que sea reproducible. Básicamente, este proceso consta de dos pasos: primero conseguir toda la documentación relacionada con el tema en cuestión y después ir filtrándola hasta conseguir un conjunto de medidas representativas de una determinada evidencia. Son los denominados criterios de selección y exclusión de estudios, y serán explicados detalladamente en los dos siguientes apartados.

En este trabajo se ha seguido la declaración PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyse*s) [Page et al., 2021b]. En ella se define una serie de elementos mínimos que se deben incluir al publicar una revisión sistemática o un meta-análisis. Principalmente se compone de un diagrama de flujo que sintetiza el proceso de obtención de artículos y de una lista de requisitos a cumplir en la publicación. Al final de esta sección se muestra el diagrama completo.

3.1.1. Criterios de selección

Para este trabajo se ha decidido utilizar estudios que se encuentren en bases de datos científicas. Concretamente, se ha utilizado *Scopus*[1], *Web of Science*[2] e *IEEE Xplore*[3]. Este tipo de bases de datos almacenan referencias a distintos tipos de documentos científicos junto con información de la publicación. Una consulta a una de estas bases de datos se suele componer principalmente de términos textuales que son buscados en las distintas secciones del artículo científico. También permiten utilizar otros filtros como el año, el tipo de la publicación, la sección del artículo en la que se busca, etc. Varios criterios de búsqueda se pueden combinar mediante operadores booleanos en una misma consulta.

Para definir la consulta se ha refinado la propuesta de Alhothali et al. [2022]. Se han planteado tres conceptos relacionados con los objetivos de este meta-análisis. Para cada concepto se propone una lista de términos clave que lo podrían referenciar, que se unen utilizando el operador `OR` en la búsqueda. Los tres conceptos tenidos en cuenta son:

1. «**Abandono**» que se compone de los sinónimos «*Dropout*», «*Retention*», «*Completion*», «*Attrition*», «*Withdrawal*».

2. «**MOOC**» con el conjunto de sinónimos «*Online learning*», «MOOC», «*Online course*» y «*Online Education*».

3. «**Aprendizaje Automático**» con sus sinónimos «*Classification*», «*Prediction*», «*Machine Learning*», «*Predictive model*» y «*Deep learning*»

Estos términos se han buscado en el título y resumen de cada artículo. En el caso de *IEEE Xplore*, no permite ser tan concreto por lo que se han buscado en todos los metadatos del estudio, que incluye título, resumen y etiquetas.

Todas las consultas se realizaron el día 2/5/2023, por lo que en este trabajo se han incluido todos los artículos hasta esa fecha. Se obtuvieron un total de 729 artículos. De los cuales 401 procedían *Scopus*, 212 de *Web of Science* y 116 de *IEEE Xplore*. En este proceso se ha podido ver que gran parte se encuentran duplicados en varias fuentes, y que casi todos los artículos considerados se pueden encontrar buscando únicamente en *Scopus*.

Estas bases de datos están implementadas de tal forma que la consulta se puede definir en formato de texto, de esta forma es fácil publicar la búsqueda asegurando su reproducibilidad. A continuación, se muestran las consultas realizadas en cada una de las fuentes.

Consulta realizada en Scopus

```
1    TITLE-ABS ((
2    ("Dropout" OR "Retention" OR "Completion" OR "Attrition" OR "
     Withdrawal")
```

[1] https://www.scopus.com/
[2] https://www.webofscience.com/
[3] https://ieeexplore.ieee.org/

```
3     AND
4     ("Online learning" OR "MOOC" OR "Online course" OR "Online
    Education")
5     AND
6     ("Classification" OR "Prediction" OR "Machine Learning" OR "
    Predictive model" OR "Deep learning")
7     ))
```

Consulta realizada en WoS

```
1 (
2     (TI=("Dropout" OR "Retention" OR "Completion" OR "Attrition" OR "
    Withdrawal")
3     OR
4     AB=( "Dropout" OR "Retention" OR "Completion" OR "Attrition" OR "
    Withdrawal" ))
5  AND
6     (TI=("Online learning" OR "MOOC" OR "Online course" OR "Online
    Education")
7     OR
8     AB=("Online learning" OR "MOOC" OR "Online course" OR "Online
    Education"))
9  AND
10    (TI=( "Classification" OR "Prediction" OR "Machine Learning" OR "
    Predictive model" OR "Deep learning" )
11    OR
12    AB=( "Classification" OR "Prediction" OR "Machine Learning" OR "
    Predictive model" OR "Deep learning" ))
13 )
```

Consulta realizada en IEEE Xplore

```
1     (
2     ("All Metadata":"Dropout" OR "All Metadata":"Retention" OR "All
    Metadata":"Completion" OR "All Metadata":"Attrition" OR "All
    Metadata":"Withdrawal" )
3     AND
4     ("All Metadata":"Online learning" OR "All Metadata":"MOOC" OR "
    All Metadata":"Online course" OR "All Metadata":"Online Education"
     ) AND
5     ("All Metadata":"Classification" OR "All Metadata":"Prediction"
    OR "All Metadata":"Machine Learning" OR "All Metadata":"Predictive
     model" OR "All Metadata":"Deep learning" ))
```

3.1.2. Criterios de exclusión

Se han excluido aquellos artículos que no cumpliesen alguna de las siguientes características:

- No esté duplicado.
- Se haya publicado en una revista con revisión por pares.
- Pueda ser descargado por los medios disponibles.
- Trate de predicción de abandono en MOOC con aprendizaje automático.
- Incluya matriz de confusión.

Esto criterios se han aplicado secuencialmente en distintas fases:

1. En la fase de **identificación** se obtuvieron 729 estudios mediante las consultas definidas en la sección anterior. De los cuales, en primer lugar, se eliminaron 286 artículos duplicados. Posteriormente, se excluyeron 273 artículos que no habían sido publicados en revistas con revisión por pares.
2. En la fase de **cribado** se excluyeron 159 artículos de los 170 artículos resultantes la fase anterior. Se realizó en tres pasos:
   a) Lectura de título y resumen: se descartaron 124 artículos que no tenían relación con el tema del meta-análisis.
   b) Descarga del artículo completo: 2 de los artículos de los 46 considerados no se pudieron obtener.
   c) Lectura del artículo completo: se descartaron 2 artículos que no realizaban predicción de abandono, 29 que no tenían matriz de confusión (o no podía ser calculada) y 2 en los que la predicción no se realizaba en MOOC.

Tras aplicar estos criterios de inclusión y exclusión, quedaron 11 artículos. Un resumen de todo este proceso se puede ver en el diagrama de flujo PRISMA de la figura 3.1. Además de la matriz de confusión se han extraído algunas **características** de cada estudio:

- La **definición de abandono** utilizada mediante la categorización de Nagrecha et al. [2017] expuesta en la sección 2.3.1:
  - Falta de participación (8 artículos).
  - Objetivos de aprendizaje no cumplidos (3 artículos).
- El **conjunto de datos** utilizado. 1 artículo ha utilizado un conjunto de datos propio, mientras que 10 utilizan conjuntos de datos públicos. Entre estos se encuentran:

  - XuentangX[4]: conjunto de datos publicado por la plataforma educativa homónima (1 artículos).
  - KDD Cup 2015[5]: subconjunto de datos del anterior que fue utilizado para una competición en predicción de abandono (6 artículos).
  - OULAD[6]: conjunto de datos de la *Open University*. Un único artículo utiliza estos datos (1 artículo).
  - Standford[7]: conjunto de datos de un MOOC de mecánica cuántica de la *Standford University* a través de *Open edX* (1 artículo).
  - HarvardX[8]: conjunto de datos de 16 MOOC publicados en edX por la *Harvard University* y *Massachusetts Institute of Technology* (1 artículo).

- El **modelado del estudiante** utilizado. Se ha utilizado la categorización de Prenkaj et al. [2021]:
  - Modelado plano: los que utilizan sólo características pre-curso o estadísticas globales de características intra-estudio (2 artículos).
  - Modelado de secuencia: si utilizan estadísticas parciales de características intra-curso en distintos momentos del curso (9 artículos).
- El tipo de **algoritmo** utilizado. Se ha separado entre:
  - Estudios con algoritmos de Aprendizaje Automático (5 artículos).
  - Estudios con algoritmo de *Deep Learning.* (6 artículos)

Un resumen de los datos obtenidos se muestra en la tabla 3.1.

Antes de acabar este apartado, es necesario comentar algunas de las decisiones tomadas a lo largo de este proceso:

- Pese a la posibilidad de introducir o incrementar el sesgo de publicación, se ha buscado únicamente en revistas con revisión por pares por dos motivos. En primer lugar para buscar artículos de mayor calidad y segundo, para adaptar la cantidad de artículos considerados a la carga de trabajo prevista en este TFM.
- El proceso de eliminación de duplicados se ha llevado a cabo en dos pasos. Primero se ha utilizado el gestor de referencias Zotero y el Identificador del Objeto Digital o *Digital Object Identifier* (DOI). Puesto que no todos los estudios incluían identificador,

[4]No disponible actualmente
[5]`http://moocdata.cn/challenges/kdd-cup-2015`
[6]`https://analyse.kmi.open.ac.uk/open_dataset`
[7]`https://datastage.stanford.edu/`
[8]`https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147`

se ha tenido que realizar un segundo paso mediante la comprobación automatizada por título.

- En el inicio del proceso se valoraron distintas medidas de efecto para su estudio, por lo que se consideró que los artículos deberían tener matriz de confusión, o debería poder ser calculada con las métricas que se facilitaban en el estudio. Disponer de la matriz de confusión permitiría posponer la decisión acerca de las medidas a utilizar en el meta-análisis.

- En la mayoría de artículos revisados se realizan varios experimentos. Por lo tanto se requerían criterios para decidir la medida de efecto a escoger. Las principales diferencias entre experimentos dentro de un mismo estudio dependen de la técnica que utilizan (como por ejemplo, el modelo, la transformación de los datos, etc.), pero en algunos presentan resultados en distintos momentos del curso. Para seleccionar un único resultado por estudio se ha utilizado el criterio del «mejor entre los más tempranos». Es decir, primero se ha escogido el o los experimentos que utilizaban menor cantidad de datos temporales. En caso de que existieran varios se ha escogido el mejor entre ellos, que se ha tomado como aquel que publicase una medida F1 más alta.

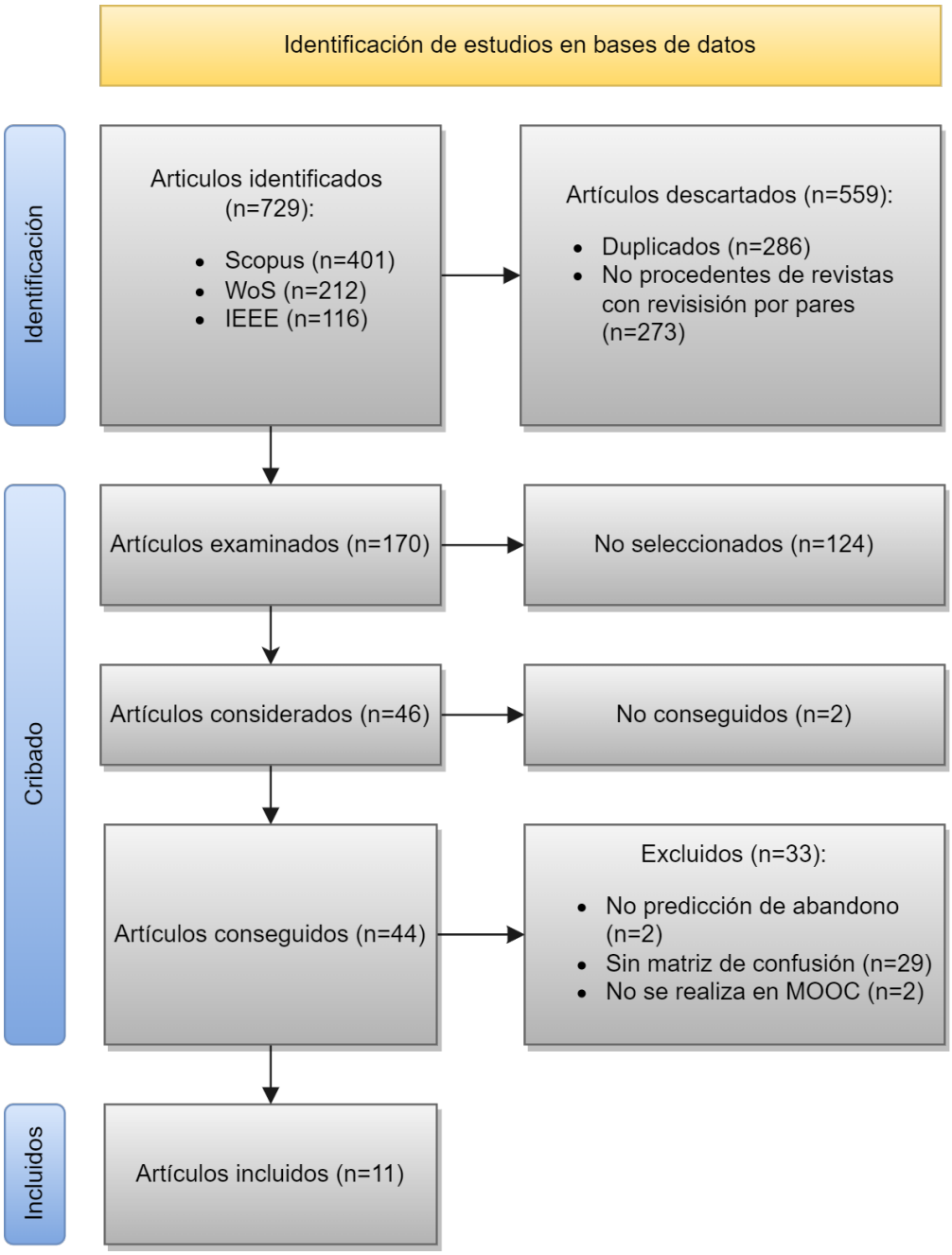

Figura 3.1: Diagrama de flujo PRISMA

Estudio	Conjunto de datos	Abandono	Modelado	Modelo	TN	FP	FN	TP
Jiang and Li [2017]	XuetangX	Participación	Secuencia	ML	55	0	16	169
Qiu et al. [2018]	KDD Cup 2015	Participación	Secuencia	ML	1123	1373	1309	8250
Mourdi et al. [2019]	Standford	Participación	Secuencia	DL	599	128	14	2844
Qiu et al. [2019]	KDD Cup 2015	Participación	Secuencia	DL	2973	3268	3066	20829
Wen et al. [2020]	KDD Cup 2015	Participación	Secuencia	DL	1408	1088	402	9156
Fu et al. [2021]	KDD Cup 2015	Participación	Secuencia	DL	2608	2384	2581	16536
Adnan et al. [2021]	OULAD	Aprendizaje	Secuencia	DL	2133	439	47	2270
Dass et al. [2021]	Propio	Aprendizaje	Secuencia	ML	8296	1587	889	8993
Chi et al. [2023]	HarvardX	Aprendizaje	Plano	ML	361	385	114	5183
Kumar et al. [2023]	KDD Cup 2015	Participación	Secuencia	DL	2071	1644	614	13770
Khoushehgir and Sulaimany [2023]	KDD Cup 2015	Participación	Plano	ML	438	4555	361	18755

Tabla 3.1: Estudios seleccionados para el meta-análisis.

3.2. Meta-análisis

Para llevar a cabo el análisis de los datos se han realizado varios experimentos. En total se han llevado a cabo 20 experimentos repartidos en **dos medidas de efecto**: sensibilidad y especificidad. Ambas son métricas de evaluación de pruebas de diagnóstico típicas que están definidas en la sección 2.3.4. El conjunto de experimentos se pueden dividir en 3 bloques:

1. Realizar un meta-análisis global utilizando varias técnicas. En total se han utilizado dos enfoques estadísticos, dos modelos y dos transformaciones.

2. Estudiar la relación entre variables de los estudios y sus resultados. Se han estudiado un total de 4 variables.

3. Estudiar el impacto de determinados estudios en el meta-análisis.

En todos los experimentos se considera que las características son lo suficientemente dispares para que se deba utilizar el **modelo de efectos aleatorios**. En las siguientes secciones se describen el conjunto de técnicas, modelos y tecnologías utilizadas.

3.2.1. Meta-análisis de proporciones

Las medidas de efecto definidas para el meta-análisis tienen la característica de ser proporciones. En el caso de la sensibilidad es la proporción de aciertos en los casos positivos, y la especificidad en los negativos. Por lo tanto, el meta-análisis debe llevarse a cabo siguiendo el enfoque del meta-análisis de proporciones. A partir de ahora, cuando se quiera referir a la proporción independientemente de una medida u otra se utilizará la siguiente notación.

Una proporción está definida por $p = \frac{y}{n}$. Donde $p \in [0, 1]$ es el valor de la proporción, y el número de eventos y n el tamaño muestral.

Existen dos posibles modelos para llevar a cabo este meta-análisis: el modelo Normal-Normal y el Binomial-Normal (a partir de aquí N-N y B-N respectivamente). En el anterior capítulo se dio una pequeña introducción al modelo N-N mediante la ecuación 2.1, que es el utilizado tradicionalmente en cualquier tipo de meta-análisis. Sin embargo, ahora es necesario dar una visión más genérica. Las expresiones que definen un meta-análisis de proporciones son:

$$\hat{\theta}_k | \theta_k \sim p(\theta_k)$$
$$\theta_k \sim \mathcal{N}(\theta, \tau^2)$$

Donde $\hat{\theta}_k$ es la medida de efecto observada en cada estudio, θ_k la medida de efecto real de cada estudio, que sigue una distribución Normal con media en θ (la medida global) y varianza τ^2. Estas expresiones recuerdan mucho a las de la ecuación 2.1, pero, en esta ocasión se ha dado una distribución genérica $p(\theta_k)$ para $\hat{\theta}_k$.

El modelo Normal-Normal

El modelo Normal-Normal tiene en cuenta el valor de la proporción de cada estudio ($\hat{p}_k$) como medida de efecto observada ($\hat{p}_k$). Tal medida se asume que sigue una distribución Normal con media en θ_k, y varianza σ_k^2 (la denominada intra-estudios). En consecuencia, la expresión para este modelo queda del siguiente modo:

$$\hat{p}_k \sim \mathcal{N}(\theta_k, \sigma_k)$$
$$\theta_k \sim \mathcal{N}(\theta, \tau^2)$$

Este enfoque es el que utiliza el método más clásico del meta-análisis: el Inverso de la Varianza [Borenstein, 2013]. Este método asigna a cada estudio un peso w_k^*, y el cálculo de la medida de efecto global viene dado por:

$$\hat{\theta} = \frac{\sum_{k=1}^{K} \hat{\theta}_k w_k^*}{\sum_{k=1}^{K} w_k^*}$$

En el modelo de efectos aleatorios, los pesos se calculan mediante el inverso de la suma de varianzas:

$$w_k^* = \frac{1}{s_k^2 + \tau^2}$$

Donde s_k^2 es la varianza intra-estudios estimada, que puede ser calculada mediante la varianza de la proporción, dada por:

$$\mathrm{Var}[p] = \frac{p(1-p)}{n} \tag{3.1}$$

Existen varios métodos para estimar la varianza inter-estudios $\hat{\tau}^2$. En este trabajo se ha utilizado la técnica de máxima verosimilitud restringida [Corbeil and Searle, 1976], un método iterativo de optimización.

El modelo Binomial-Normal

Detrás de una proporción subyace un número n de variables aleatorias binarias con distribución de Bernuilli en las que y veces ha tenido éxito. El modelo Binomial-Normal tiene en cuenta este hecho y propone que el número de aciertos observados en un estudio ($\hat{y}_k$) sigue una distribución binomial con el numero de casos de la proporción (n_k) y la medida de efecto real de cada estudio θ_k como parámetros. Para que este modelo funcione correctamente es necesario realizar una transformación *logit*. El modelo queda del siguiente modo:

$$\begin{aligned} \hat{y}_k &\sim \text{Bin}(n_k, \theta_k) \\ \text{logit}(\theta_k) &\sim \mathcal{N}(\theta^{LO}, \tau^2) \end{aligned} \tag{3.2}$$

La transformación logit está definida por:

$$\text{logit}(p) = \log \frac{p}{1-p} \tag{3.3}$$

En la práctica, los modelos mixtos lineales generalizados (GLMM por sus siglas en inglés) se consideran el método más común para utilizar este modelo. Son modelos de regresión que incorporan las dos fuentes de error (efecto fijo y efectos aleatorios), normalmente estimados mediante máxima verosimilitud. Este modelo está definido en Schwarzer et al. [2019] y Stijnen et al. [2010]. A diferencia del Inverso de la varianza, este meta-análisis no utiliza pesos.

Transformaciones

Como se ha visto en el modelo B-N, transformando las variables surgen propiedades que que pueden facilitar el análisis. Además de la transformación *logit*, existen otros tipos. Un ejemplo es la transformación *arcoseno-raíz* (a partir de ahora transformación *arcoseno*). La cual suele dar varianzas más estables debido a que dependen sólo del tamaño muestral, que suele ser un valor fijo [Lin and Xu, 2020]. Esta transformación está definida por:

$$\theta_k^{AS} = \arcsin \sqrt{p_k}$$

3.2.2. Modelos Bayesianos

En Borenstein [2013] se menciona que los modelos bayesianos pueden resultar útiles si se dispone de pocos estudios. En este trabajo se han realizado experimentos con versiones bayesianas de los dos modelos (véase Salpeter et al. [2009], Jackson et al. [2018] y Negrín-Hernández et al. [2021]). Para implementarlo se ha utilizado software de simulación MCMC.

Tal y como se vio en el capítulo anterior, los modelos bayesianos requieren que se especifique un *prior* para cada parámetro. Esto supone especificar una distribución para θ y τ^2. Se han usados *priors* débilmente informativos en ambos casos. En el caso del efecto agrupado θ se ha asumido que sigue una distribución uniforme entre 0 y 1. Mientras que se ha usado una distribución gamma-inversa(0.1; 0.1) para la varianza τ^2. Según Al Amer et al. [2021] este prior mejora la estabilidad y convergencia. Por lo tanto, en el caso bayesiano, los modelos N-N y B-N se completan con las distribuciones:

$$\theta \sim \text{Unif}(0; 1)$$
$$\tau^2 \sim \text{Gamma-Inv}(0.1; 0.1)$$

Por último, para evaluar que la simulación ha funcionado correctamente se han utilizado dos métricas [Gelman et al., 1995]: $\hat{R}$ y *Effective Sample Size* o ESS. La primera de ellas es una medida de convergencia de la simulación, y debe estar siempre muy próxima a 1 en todos los parámetros (depende del autor, pero normalmente por debajo de 1.01). Por otro lado, ESS es una medida de auto-correlación de las cadenas, se aconseja que esté por encima de 100.

3.2.3. Análisis de subgrupos

En ocasiones los resultados de un meta-análisis indican que la heterogeneidad es demasiado alta. Un posible enfoque ante esta situación consiste en buscar patrones segmentando los estudios en función de una o varias características cualitativas. Es lo que se suele denominar análisis de subgrupos. Mediante esta técnica es posible establecer una relación entre el valor de la característica y el de la medida de efecto. Consta de dos pasos:

1. Calcular el efecto de cada grupo como si fuera un meta-análisis por separado. Cada grupo consta de su efecto agrupado $\hat{\theta}_g$ y su heterogeneidad $\hat{\tau^2}_g$.

2. Comparar los resultados de cada grupo con un test estadístico que comprueba si las diferencias entre grupos se deben a la varianza intra-grupos o inter-grupos. En realidad se trata de un test Q, pero en este caso el p-valor del test es menor de 0.05 cuando al menos dos grupos son significativamente diferentes. Para evitar confusiones se utilizará test χ^2 o test de diferencias de subgrupos para este test, y test Q o test de diferencias entre estudios para el otro.

3.2.4. Análisis de influyentes

Otra de las técnicas que se han usado para estudiar la heterogeneidad es el análisis de influyentes. Se denominan estudios influyentes aquellos que tienen un gran impacto en la medida de efecto agrupada o en la heterogeneidad. Hay varias técnicas para identificar estudios influyentes, todas ellas están basadas en el método *Leave-One-Out* [Harrer et al., 2021]. Este método vuelve a realizar el meta-análisis tantas veces como estudios se dispongan, pero dejando uno fuera en cada iteración. Existen varios métodos para comprobar la influencia de los estudios, en este caso el análisis se limitará a estudiar el efecto en la correspondiente medida de efecto y en los índices de heterogeneidad I^2 y τ^2.

Este método puede ayudar a detectar errores o características de los estudios que no se habían visto inicialmente. Sin embargo, eliminar estudios del análisis sólo por el hecho de

ser influyentes supondrá una pérdida de poder estadístico. Es importante tener en cuenta que un meta-análisis debe abarcar todos los estudios posibles.

3.2.5. Paquetes software utilizados

La parte de procesamiento y análisis de datos se ha realizado mediante el lenguaje de programación R en un entorno *Jupyter Notebook*. Para el análisis se han usado dos paquetes principalmente. El paquete `meta` ha sido utilizado para implementar los modelos con enfoque frecuentista y para los modelos bayesianos se ha utilizado el paquete `RStan`. A continuación, se describen un poco más cada uno de ellos y mencionan las funciones utilizadas.

meta

El paquete `meta` [Balduzzi et al., 2019] es una librería muy popular para el lenguaje de programación R que incluye múltiples modelos, métodos y elementos de apoyo para realizar meta-análisis de forma sencilla.

En este trabajo se ha utilizado la función `metaprop` para llevar a cabo el meta-análisis de proporciones. A continuación, se describen los principales parámetros que se han utilizado:

- `event`: número de eventos o numerador de la proporción. En el caso de la sensibilidad son los TP y en la especificidad TN.
- `n`: número de observaciones o el denominador. En caso de sensibilidad y especificidad son los casos positivos(N^+) y casos negativos (N^-) totales respectivamente.
- `method`: método utilizado para el meta-análisis. Puede tomar dos valores: `Inverse` define que se debe usar el método de Inverso de la varianza y `GLMM`, como su propio nombre indica, GLMM.
- `sm`: define transformación a realizar. En este trabajo se han utilizado los valores `"PRAW"` para no realizar transformaciones, `"PAS"` para la transformación arcoseno y `"PLOGIT"` para la *logit*.
- `subgroup`: realiza análisis de subgrupos segmentando los estudios. El valor del parámetro debe ser un vector con los valores de la variable para cada estudio.

El paquete `meta` también incluye la función `metainf`, que es utilizada para realizar análisis de influyentes. Esta función recibe como parámetro un objeto `meta` resultado de realizar el meta-análisis global y automáticamente muestra los resultados de llevar a cabo la técnica.

RStan

`RStan` [Stan Development Team, 2023] es la interfaz en `R` de `Stan`, una de las principales plataformas de modelado estadístico. Está escrita en `C++` e incluye simulación de modelos bayesianos mediante MCMC. Es uno de los paquetes más utilizados por la comunidad bayesiana por lo que dispone de una amplia documentación. Este es el motivo principal por el que se ha escogido para este trabajo.

El elemento fundamental de RStan es el archivo en el que se define el modelo, que al menos está compuesto por tres bloques:

- `data`: declara las variables, tipo de dato, restricciones y dimensiones de los datos de entrada.
- `parameters`: declara los parámetros del modelo.
- `model`: define las distribuciones de probabilidad.

En el anexo A se incluye el código de los archivos `stan` de los dos modelos utilizados. Si se observa se podrá ver que existen otros bloques no esenciales como `generated quantities` para generar valores después de la ejecución, o `transformed parameters` para transformar parámetros.

Una vez se ha definido el modelo la simulación se ejecuta mediante la función `stan`. En ella es necesario especificar los parámetros de simulación: modelo, datos de entrada, parámetros de MCMC como número de cadenas e iteraciones, etc. El resultado se compone principalmente de una tabla con los valores de cada parámetro, y los resultados de convergencia del modelo.

Todos los experimentos de este trabajo levados a cabo mediante modelos bayesianos han utilizado la misma configuración de simulación: 4 cadenas de 2000 iteraciones cada una. En todos se ha comprobado que la autocorrelación y convergencia se encuentre en niveles correctos adecuado de $\hat{R}$ y ESS.

3.3. Resumen de experimentos realizados

En total se han realizado 20 experimentos, divididos en 10 configuraciones para las dos medidas de efecto seleccionadas.

En el primer bloque se consideran todos los estudios de forma conjunta. Los experimentos 1, 2 y 3 utilizan un enfoque frecuentista. Se ha implementado el modelo N-N mediante inverso de la varianza, con y sin transformación arcoseno; y el modelo B-N con transformación *logit* mediante GLMM. Los experimentos 4 y 5 replican los experimentos 1 y 3 desde una perspectiva bayesiana.

Posteriormente, en los experimentos 6 a 9 se realiza un análisis de subgrupos teniendo en cuenta distintas variables. Se han segmentado los estudios en base a el conjunto de datos, la definición de abandono, el modelo, y el modelado del estudiante.

Por último, en el experimento 10 se han analizado los estudios más influyentes mediante *Leave-One-Out*.

Tanto los experimentos de análisis de subgrupos como los de análisis de influyentes utilizan el modelo N-N frecuentista sin transformaciones. En la tabla 3.3 se muestra un resumen las características de los experimentos.

Tabla 3.2: Resumen de los experimentos realizados

Exp.	Tipo de análisis	Enfoque	Modelo	Detalles
1	Global	Frec.	N-N	Sin transformaciones
2	Global	Frec.	N-N	Transformación arcoseno
3	Global	Frec.	B-N	Transformación logit
4	Global	Bayes	N-N	Sin transformaciones
5	Global	Bayes	B-N	Transformación logit
6	Subgrupos	Frec.	N-N	Agrupado por definición de abandono
7	Subgrupos	Frec.	N-N	Agrupado por conjunto de datos
8	Subgrupos	Frec.	N-N	Agrupado por tipo de algoritmo
9	Subgrupos	Frec.	N-N	Agrupado por modelado del estudiante
10	Influyentes	Frec.	N-N	Sin transformaciones ni agrupaciones

Capítulo 4

Resultados

A lo largo de este capítulo se presentan los resultados de los 20 experimentos realizados. Se divide en dos secciones, una para cada medida de efecto (sensibilidad y especificidad). En el inicio de cada sección se realiza una comprobación del riesgo de sesgo mediante un gráfico de embudo. Todos los experimentos se acompañan del correspondiente *forest plot* con excepción del experimento 10, cuyos resultados se presentan en forma de tabla. En caso de los experimentos 4 y 5, correspondientes a los modelos con enfoque bayesiano, se incluye una gráfica de densidad de los parámetros globales del modelo.

4.1. Sensibilidad

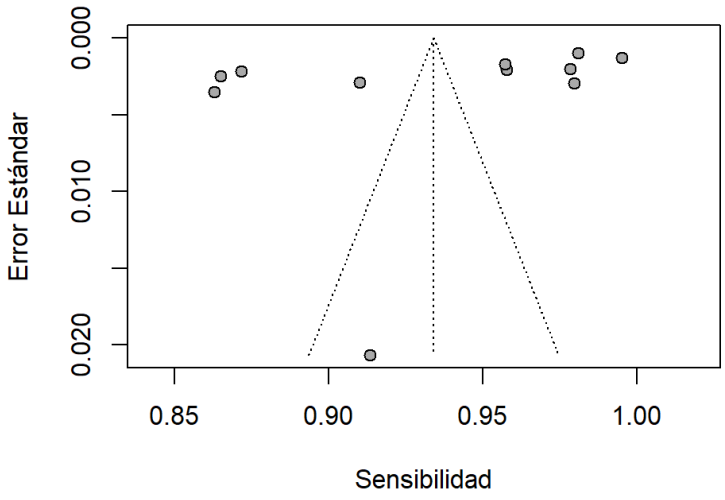

Figura 4.1: Gráfico de embudo para el meta-análisis de sensibilidad.

El meta-análisis de sensibilidad se compone de 11 estudios. En general todos los estudios obtienen un valor bastante alto, entre 0.86 y 1. La precisión de las medidas también es bastante alta en general. Obtienen un error menor de 0.005, excepto uno de los estudios

[Jiang and Li, 2017], que se encuentra por encima de 0.020. El gráfico de embudo (figura 4.1) no muestra una asimetría clara, ya que el estudio con menos precisión no permite ver con claridad la distribución del resto de estudios. No se ha considerado un motivo para excluir este estudio de los experimentos, posponiendo para el experimento 10 esta decisión. A continuación, se muestran los resultados de cada uno de los experimentos.

4.1.1. Experimento 1

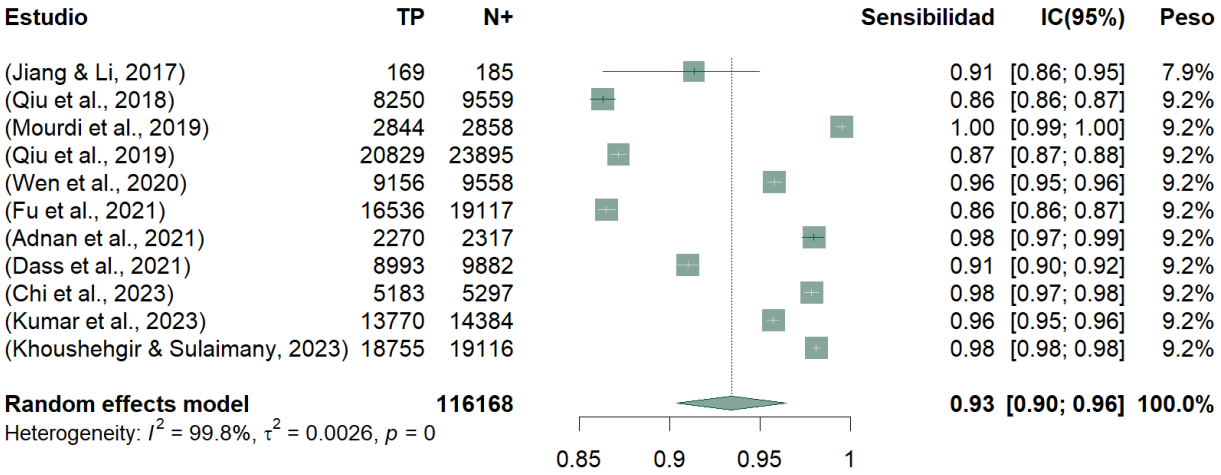

Figura 4.2: *Forest plot* del experimento 1 para meta-análisis de sensibilidad

Para el experimento 1 se ha utilizado el esquema más clásico de meta-análisis, que utiliza el modelo N-N sin transformaciones mediante el método de Inverso de la Varianza. En la figura 4.2 se muestra el *forest plot* correspondiente. Se ha obtenido un valor de 0.93 para el efecto agrupado, con un intervalo de confianza del 95 % entre 0.90 y 0.96. Los pesos obtenidos por el método son muy similares, esto es debido a la alta heterogeneidad del meta-análisis.

Pese a que el valor de τ^2 obtenido puede parecer muy pequeño (0.0026), es muy grande en relación a las varianzas intra-estudio calculadas, que se encuentran por debajo de 10^{-5}. Esto conlleva a un valor de I^2 muy alto. En la figura se puede ver que se ha obtenido un valor de 99.8 % , lo que indica que casi toda la variabilidad entre los estudios no puede ser explicada por el error muestral. De igual manera, el test Q de heterogeneidad obtiene un p-valor de 0, lo que indica que existen diferencias claramente significativas entre los estudios.

4.1.2. Experimento 2

En el experimento 2 se ha utilizado la transformación *arcoseno* en la medida de efecto. Los resultados obtenidos son muy similares a los del experimento anterior, tal como se muestra en la figura 4.3. El valor de la sensibilidad agrupada que se ha obtenido es de 0.94 con intervalo de $[0.91; 0.97]$, lo que supone un aumento de 0.01 con respecto al experimento 1.

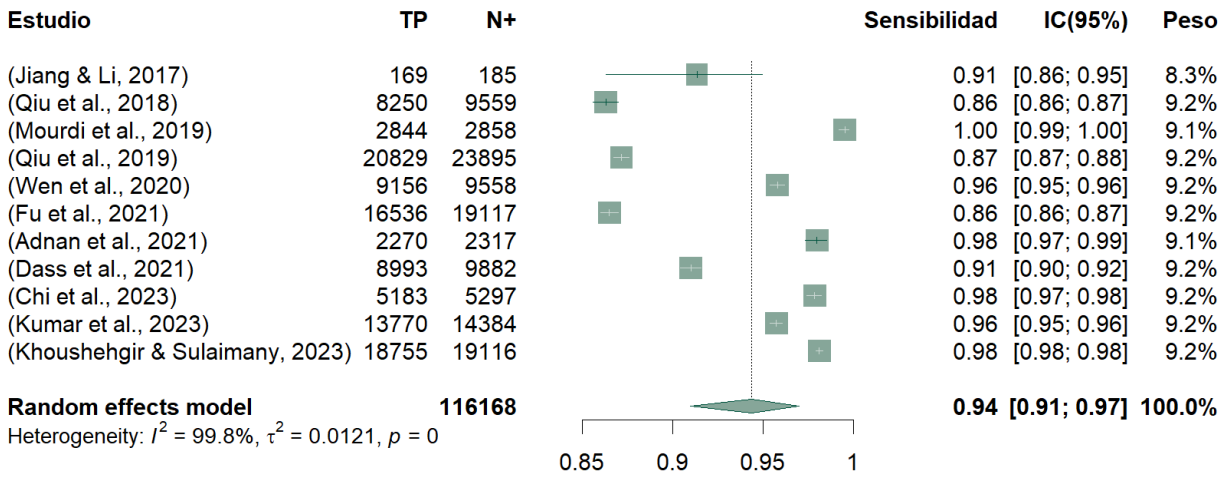

Figura 4.3: *Forest plot* del experimento 2 para la sensibilidad

Al igual que en el anterior experimento, los pesos de todos los estudios son muy similares, oscilando entre 8.3 % y 9.2 %.

Tampoco se han apreciado diferencias en la heterogeneidad respecto al experimento 1. De nuevo se obtiene un valor de 99.8 % para I^2 y el test Q encuentra diferencias significativas entre estudios ($p = 0$). Sin embargo, sí se aprecian diferencias en el valor τ^2, que es 0.121. Este valor es casi 5 veces el del experimento 1. Esto es debido a que se encuentra en la escala de la transformación.

4.1.3. Experimento 3

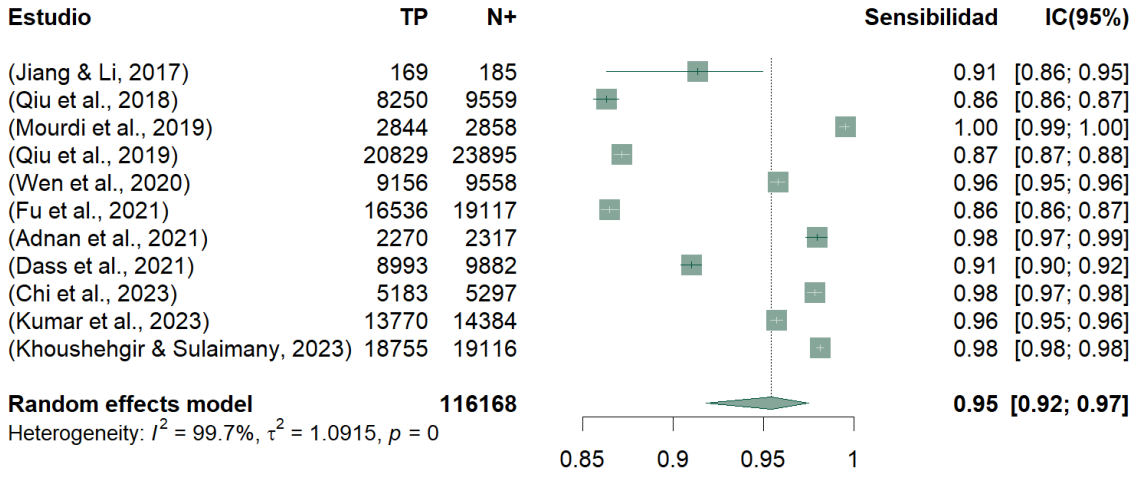

Figura 4.4: *Forest plot* del experimento 3 para la sensibilidad

El experimento 3, tal y como se ha explicado en el capítulo anterior, tiene en cuenta la estructura binaria intrínseca a la proporción mediante el modelo B-N y la transformación *logit*. Se ha llevado a cabo desde el enfoque frecuentista de los GLMM. El *forest plot* obtenido

se puede ver en la figura 4.4. El valor del efecto agrupado es 0.95 con un intervalo de confianza de $[0.92; 0.97]$. Estos valores están levemente por encima de los obtenidos en los dos experimentos anteriores. Es interesante observar que en este caso los intervalos de la sensibilidad agrupada no son simétricos a diferencia de los experimentos anteriores. En el gráfico no se presentan los pesos de cada estudio ya que este método no los utiliza.

La heterogeneidad se ha reducido levemente con este modelo. En el gráfico, se puede ver como dos de las medidas están muy próximas a la medida global. El valor de I^2 es de 99.7 % , aunque el p-valor del test de diferencias sigue siendo 0. El valor de τ^2 es claramente superior a los otros experimentos debido a que se encuentra en escala *logit*.

4.1.4. Experimento 4

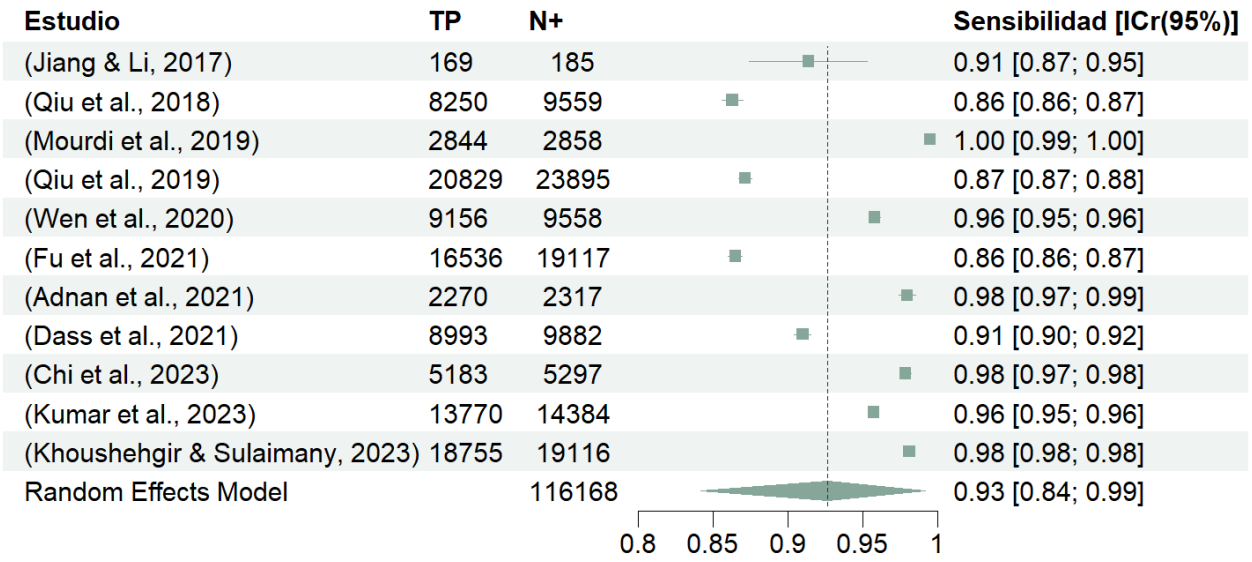

Figura 4.5: *Forest plot* del experimento 4 para la sensibilidad

El experimento 4 utiliza un enfoque bayesiano para el meta-análisis mediante el modelo N-N, el mismo del experimento 1. El *forest plot* del modelo se muestra en la figura 4.5.

El efecto agrupado obtenido en el meta-análisis es de $0.93[0.84, 0.99]$. El valor medio es similar al del experimento 1 pero los intervalos son bastante más amplios. Para medir la heterogeneidad se ha utilizado únicamente la varianza inter-estudios que ha obtenido un valor de 0.0244, por lo que es mucho más alta que la obtenida en el experimento 1. Estas diferencias con el enfoque frecuentista pueden ser causadas por los *priors* utilizados.

Para explorar más en profundidad los parámetros se hace uso de su distribución a posteriori. En la figura 4.6 se muestra la distribución para la varianza inter-estudios τ^2 y la sensibilidad agrupada θ, junto con su valor puntual e intervalo de credibilidad del 95 % (zona coloreada). En ambos casos se puede ver que sigue una distribución unimodal aunque con

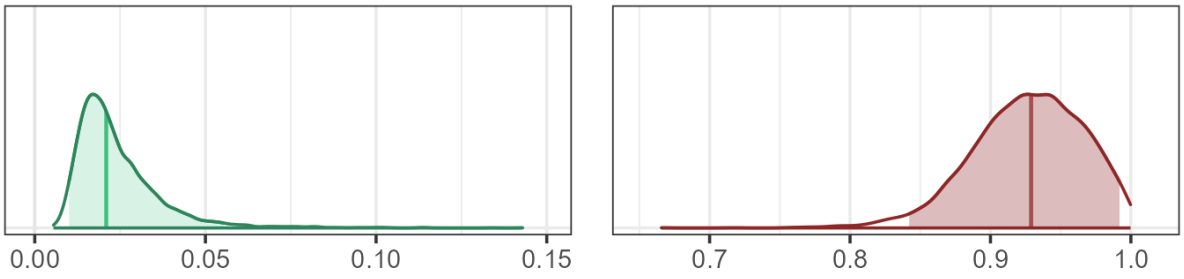

Figura 4.6: Distribución a posteriori de la varianza inter-estudios τ^2 (izquierda) y la medida de efecto agrupada θ (derecha) del experimento 4 de la sensibilidad.

ciertas diferencias. En el caso de θ la distribución se asemeja a una Gaussiana con la media y la moda en el mismo punto. Mientras que en la distribución de τ^2, media y moda están ligeramente separadas, y tiene una gran cola derecha.

4.1.5. Experimento 5

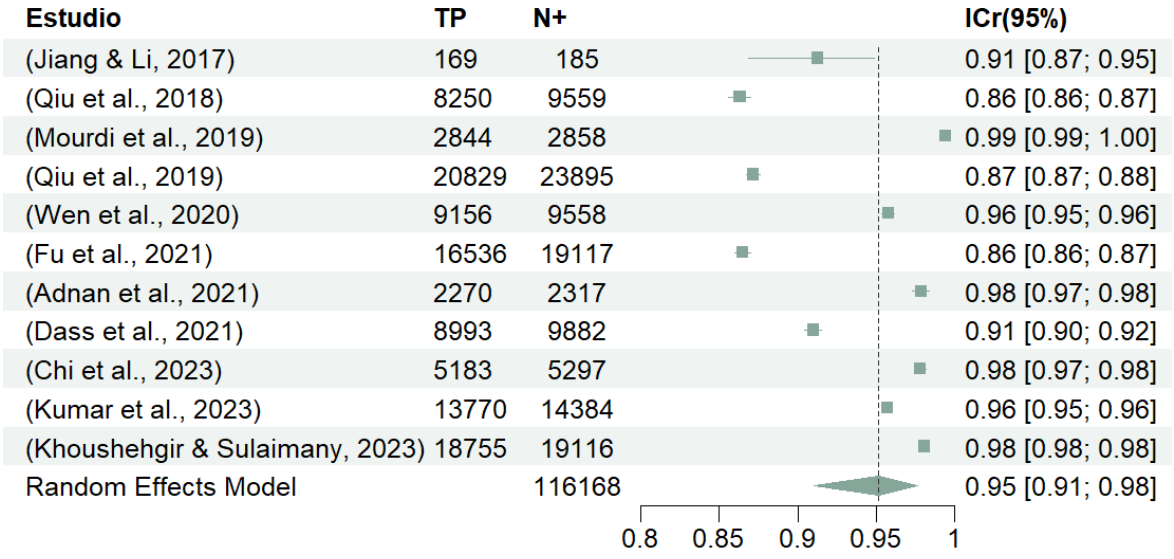

Figura 4.7: Diagrama de efectos del experimento 5 para la sensibilidad

El experimento 5 utiliza el modelo B-N con transformación *logit* con el enfoque bayesiano. El *forest plot* del modelo se muestra en la figura 4.7.

Este modelo ha conseguido un efecto agrupado de $0.95[0.91; 0.98]$. A diferencia del experimento 4, no sólo el valor medio es semejante al modelo frecuentista; si no también lo son los intervalos. Al igual que en el modelo anterior se ha obtenido un valor puntual de la varianza inter-estudios algo superior al modelo frecuentista (en escala *logit* $\tau^2 = 1.317$).

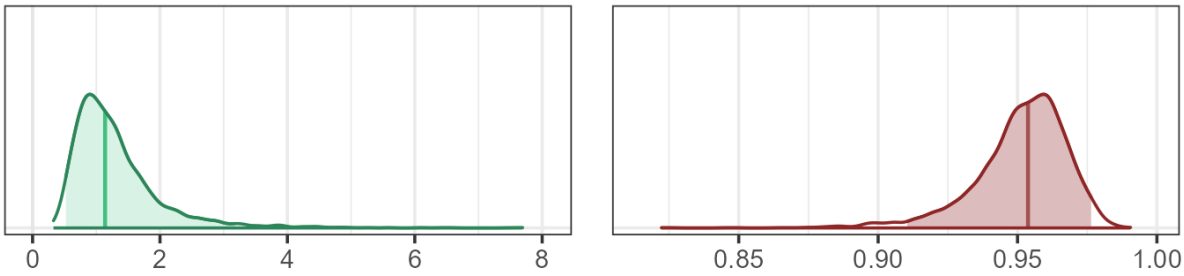

Figura 4.8: Distribución de la varianza inter-estudios τ^2 (izquierda) y la medida de efecto agrupada θ (derecha) del experimento 5 de la sensibilidad.

Para el análisis de las distribuciones a posteriori de los dos parámetros globales se puede observar la figura 4.8. De nuevo, ambos parámetros siguen una distribución unimodal. La distribución de la varianza inter-estudios es similar al experimento anterior, con la media ligeramente desplazada a la derecha de la moda y una larga cola derecha. No obstante, la medida de efecto agrupado presenta cierta asimetría y una larga cola izquierda a diferencia del modelo N-N.

4.1.6. Experimento 6

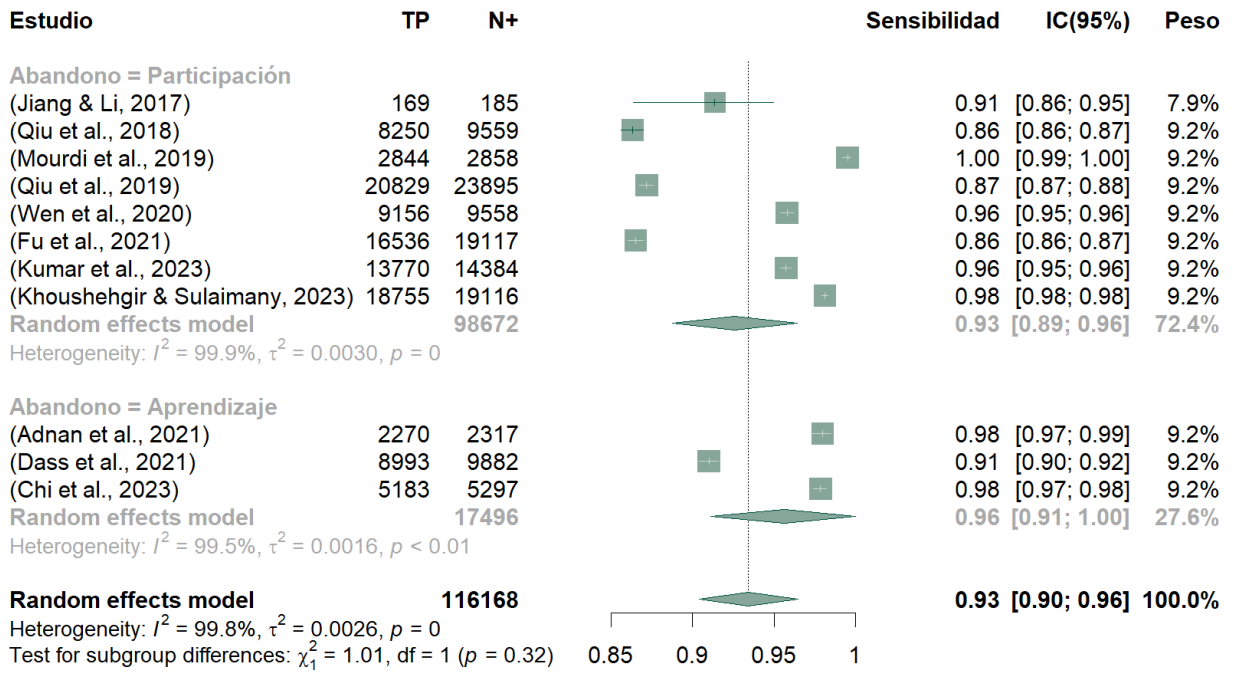

Figura 4.9: *Forest plot* del experimento 6 para la sensibilidad

Utilizando análisis de subgrupos, este experimento realiza el meta-análisis segmentando los estudios en función de la definición de abandono que haya utilizado. Se han considerado

aquellos que utilizan la definición basada en participación ($G_{Participacion}$) y en objetivos de aprendizaje ($G_{Aprendizaje}$). Los resultados se incluyen en la figura 4.9.

En $G_{Participacion}$ el efecto agrupado es de $0.93[0.89; 0.96]$. Incluye 8 estudios cuyas medidas de efecto se encuentran entre 0.86 y 1.00. Obtiene un valor de I^2 de 99.9 % y de τ_g^2 de 0.003. El p-valor del test de diferencias encuentra diferencias significativas entre estudios. Todos los índices indican que la heterogeneidad en este grupo es superior a la del análisis global.

En el grupo de $G_{Aprendizaje}$ se obtiene una sensibilidad agrupada de $0.96(0.91, 1.00)$, una medida 0.05 más alta que el otro grupo. La heterogeneidad se ha visto reducida respecto al global. El índice I^2 se ve reducido hasta 99.5 % y el de τ_g^2 hasta 0.0016. El p-valor del test de diferencias sigue siendo significativo. Pese a los pocos estudios que tiene este grupo la diferencias siguen siendo altas.

El test χ^2 para diferencias entre subgrupos muestra que no existen diferencias significativas entre los dos subgrupos con un p-valor de 0.32. En el gráfico se puede ver que el efecto global recae dentro del intervalo del efecto estimado para ambos grupos.

4.1.7. Experimento 7

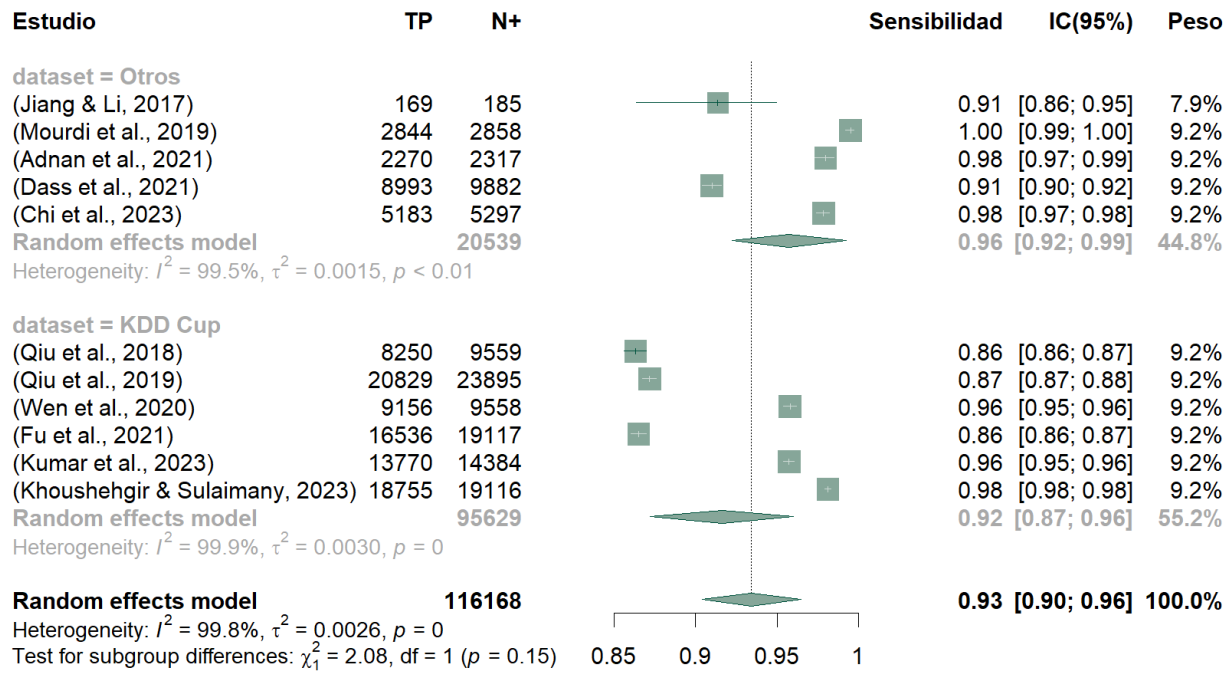

Figura 4.10: *Forest plot* del experimento 7 para la sensibilidad

En el experimento 7 se ha realizado el análisis de subgrupos agrupando los estudios en función del conjunto de datos utilizado. Puesto que la mayoría de estudios usan el conjunto *KDD Cup 2015* y el resto utilizan otros conjuntos, la agrupación se ha realizado en dos grupos. Por un lado los 6 estudios que utilizan *KDD Cup 2015* (G_{KDD}) y por otro lado el

resto de estudios (G_{otros}). Al igual que el resto de análisis de subgrupos, para llevar a cabo este análisis se ha utilizado el método de Inverso de la Varianza sin transformaciones. En la figura 4.10 se muestran los resultados de este experimento.

El efecto agrupado de G_{otros} está levemente por encima del global: $0.96[0.92; 0.99]$. La varianza inter-estudios se ha reducido respecto al análisis global, con un valor de 0.0015 para τ_g^2. El índice I^2 también se ha reducido hasta 99.5 %, pero el test Q sigue encontrado diferencias significativas. En G_{KDD} se obtiene una sensibilidad inferior, con un valor de $0.92[0.87; 0.96]$. La heterogeneidad obtenida es superior a la global, con un valor de I^2 de 99.9 % y τ_g^2 de 0.003.

El test χ^2 de diferencias de subgrupos no encuentra diferencias significativas ($p = 0.15$).

4.1.8. Experimento 8

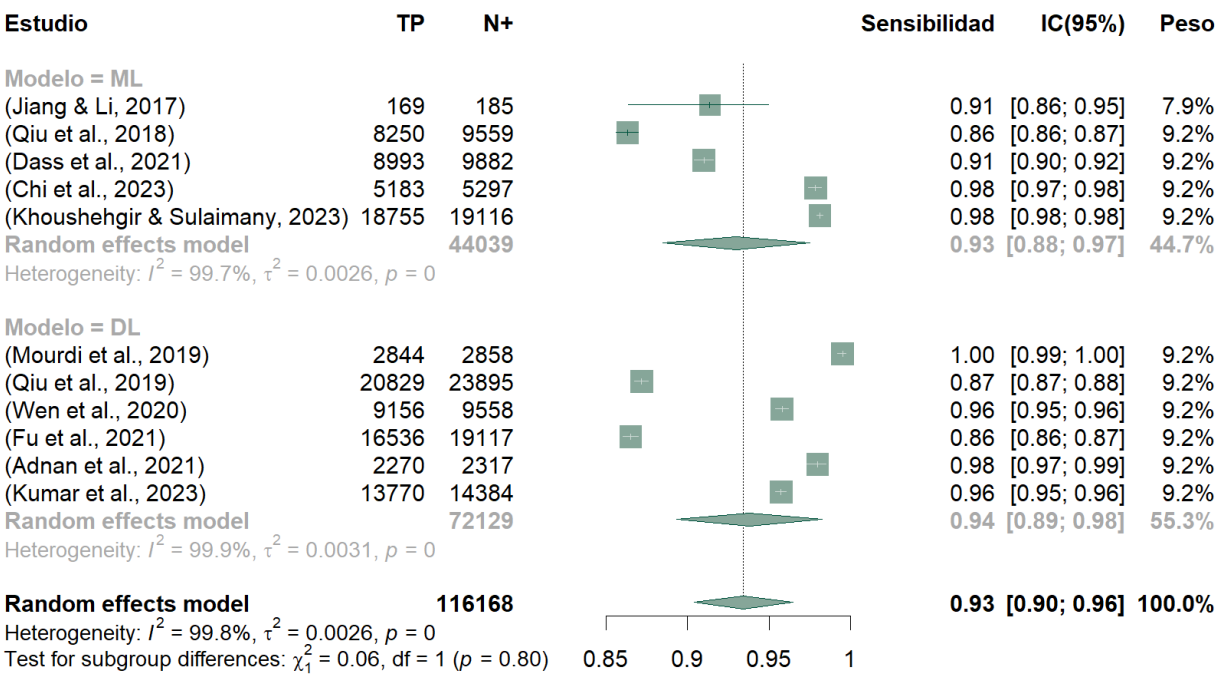

Figura 4.11: *Forest plot* del experimento 8 para la sensibilidad

La tercera variable que se ha utilizado para el análisis de subgrupos es el tipo de algoritmo utilizado. Se han definido dos grupos, aquellos que utilizan *Deep learning* (G_{DL}) y los que utilizan modelos de aprendizaje automático (G_{ML}), grupos compuestos por 6 y 5 estudios respectivamente. La figura 4.11 muestra el *forest plot* de este experimento.

A primera vista no se aprecian apenas diferencias entre los dos grupos. Se puede ver mucho solapamiento entre la medida global y la de los dos grupos. El efecto agrupado en G_{ML} es de $0.93(0.88; 0.97)$ y en G_{DL} es de $0.94[0.89; 0.98]$. En cuanto a la heterogeneidad,

el índice I^2 es del 99.7 % y 99.9 % en G_{ML} y G_{DL} respectivamente; y el valor de τ_g^2 es 0.0087 y 0.0069. El test Q encuentra diferencias significativas en ambos grupos.

Con estos valores resulta obvio que no se encuentren diferencias significativas entre los dos grupos mediante el test χ^2 ($p = 0.87$).

4.1.9. Experimento 9

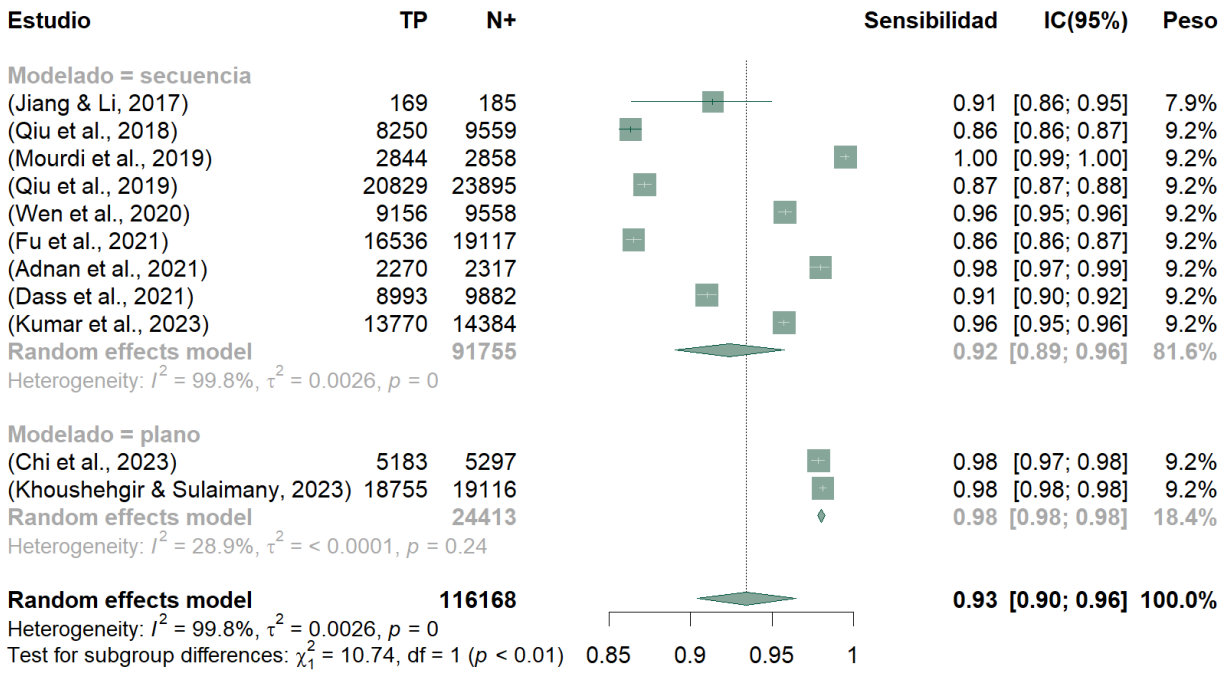

Figura 4.12: *Forest plot* del experimento 9 para la sensibilidad

Por último, se ha realizado el análisis de subgrupos en función del tipo de modelado del estudiante. El grupo G_{plano} se compone de 2 estudios y $G_{secuencia}$ de 9. En la figura 4.12 se muestran los resultados del meta-análisis, se puede ver que los resultados son muy distintos en los dos grupos.

La sensibilidad agrupada en $G_{secuencia}$ muy similar a la global con un valor medio de 0.92 e intervalos en $[0.89; 0.96]$. Los índices de heterogeneidad son exactamente iguales a los globales. Se obtiene un I^2 de 99.8 %, τ_g^2 de 0.0026, y el test Q obtiene un p-valor de 0.

Sin embargo, los dos estudios en G_{plano} tienen una medida muy similar. Lo que hace descender los valores de heterogeneidad en gran medida en este grupo. El valor de I^2 es 28.9 % y τ_g^2 está por debajo de 10^{-4}. El test Q no encuentra diferencias significativas ($p = 0.24$). La medida de efecto de este grupo es 0.98.

Visualmente se puede intuir que existen diferencias entre grupos, ya que la medida global no interseca con los intervalos de G_{plano} . El test χ^2 ha confirmado estas diferencias ($p < 0.01$).

4.1.10. Experimento 10

Por último, se ha realizado un análisis de influyentes mediante la técnica de *Leave One Out*. Esta técnica estudia el impacto de cada estudio en el meta-análisis excluyéndolo del conjunto.

En la tabla 4.1 se pueden ver la sensibilidad agrupada, el índice I^2 y la varianza inter-estudios de esta prueba. En general, los valores de la medida de efecto se encuentran entre 0.92 y 0.95, frente al 0.93 del análisis global. Por lo tanto, extraer un estudio del meta-análisis no afecta en gran medida a la sensibilidad.

Tabla 4.1: Resultados del análisis de influyentes en el meta-análisis de sensibilidad

Estudio excluido	Sensibilidad	τ^2	I^2
(Jiang & Li, 2017)	0.9360 [0.9031; 0.9689]	0.0028	99.8 %
(Qiu et al., 2018)	0.9414 [0.9117; 0.9712]	0.0023	99.8 %
(Mourdi et al., 2019)	0.9280 [0.8972; 0.9589]	0.0024	99.8 %
(Qiu et al., 2019)	0.9406 [0.9099; 0.9712]	0.0024	99.8 %
(Wen et al., 2020)	0.9318 [0.8986; 0.9649]	0.0028	99.8 %
(Fu et al., 2021)	0.9413 [0.9113; 0.9712]	0.0023	99.8 %
(Adnan et al., 2021)	0.9296 [0.8975; 0.9617]	0.0026	99.8 %
(Dass et al., 2021)	0.9366 [0.9035; 0.9697]	0.0028	99.8 %
(Chi et al., 2023)	0.9297 [0.8976; 0.9619]	0.0026	99.8 %
(Kumar et al., 2023)	0.9318 [0.8987; 0.9650]	0.0028	99.8 %
(Khoushehgir & Sulaimany, 2023)	0.9294 [0.8975; 0.9614]	0.0026	99.8 %
Global	0.9342 [0.9039; 0.9645]	0.0026	99.8 %

En cuanto a los valores de heterogeneidad, la varianza inter-estudios se mantiene estable entre 0.0023 y 0.0028, valores muy similares al análisis global. El valor de I^2 se mantiene también estable en la mayoría de extracciones, consiguiendo un valor de 99.8 %.

Con estos resultados no se puede considerar que ningún estudio tenga un impacto lo suficientemente alto como para considerar extraerlo del análisis. Aun así, se ha realizado una prueba más. Se ha comprobado cuantos estudios sería necesario extraer para que los índices de heterogeneidad alcanzasen valores aceptables. Para ello se ha ejecutado iterativamente esta técnica hasta que el p-valor del test Q esté por encima de 0.05. Se ha visto que se requieren eliminar 8 estudios del meta-análisis. El resultado es una sensibilidad agrupada de $0.957[0.954; 0.960]$, con un índice I^2 de 56.4 % y un p-valor de 0.1 en el test Q. Los estudios implicados en este meta-análisis son Jiang and Li [2017], Wen et al. [2020] y Kumar et al. [2023].

Al inicio de la sección se mencionaba Jiang and Li [2017] estaba dificultando la visualización del gráfico de embudo debido a su precisión. Se ha demostrado que este artículo es uno de los que más robustez aportan al meta-análisis y sus características no son distintas al resto por lo que se ha decidido no excluirlo. .

4.2. Especificidad

En este apartado se realizan los 10 experimentos con la especificidad como medida de efecto. Al igual que con la sensibilidad, antes se ha comprobado el riesgo de sesgo mediante un gráfico de embudo (figura 4.13). Como se puede ver la distribución de cada uno de los efectos es casi aleatoria. Los valores de especificidad prácticamente se distribuyen en todo el rango posible, aunque uno de los estudios se encuentra especialmente distanciado Khoushehgir and Sulaimany [2023]. Los errores estándar están todos por debajo de 0.02. Los tamaños muestrales de cada uno de los estudios son muy distintos, y más pequeños que en el anterior meta-análisis, abarcando desde 55 hasta 9883. En comparación con el meta-análisis de sensibilidad, en este análisis existen más estudios con tamaños muestrales pequeños, lo que conlleva una menor precisión.

Esta distribución indica que es muy poco probable que se puedan sacar conclusiones del meta-análisis, y en caso que así fuese, el sesgo de publicación sería muy alto.

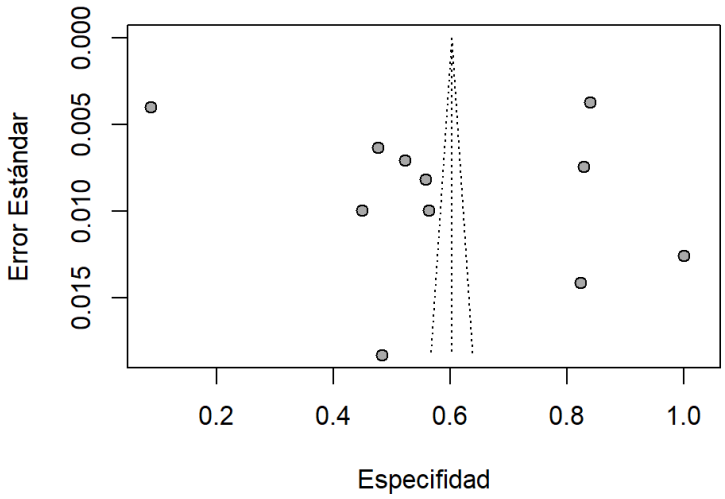

Figura 4.13: Gráfico de embudo de especificidad.

4.2.1. Experimento 1

El primer experimento realiza el meta-análisis mediante el método del Inverso de la Varianza sin transformaciones. Los resultados se puede ver en la figura 4.14.

El efecto agrupado obtenido es de 0.60 con unos intervalos de confianza del 95 % muy amplios, entre 0.45 y 0.75. Este método ha asignado el mismo peso a todos los estudios, esto es debido de nuevo a la alta heterogeneidad.

El índice I^2 está en 100 % y el test Q concluye que existen diferencias significativas entre estudios ($p = 0$). La varianza inter-estudios en este caso es de 0.0642, un orden de magnitud más grande que el meta-análisis de sensibilidad.

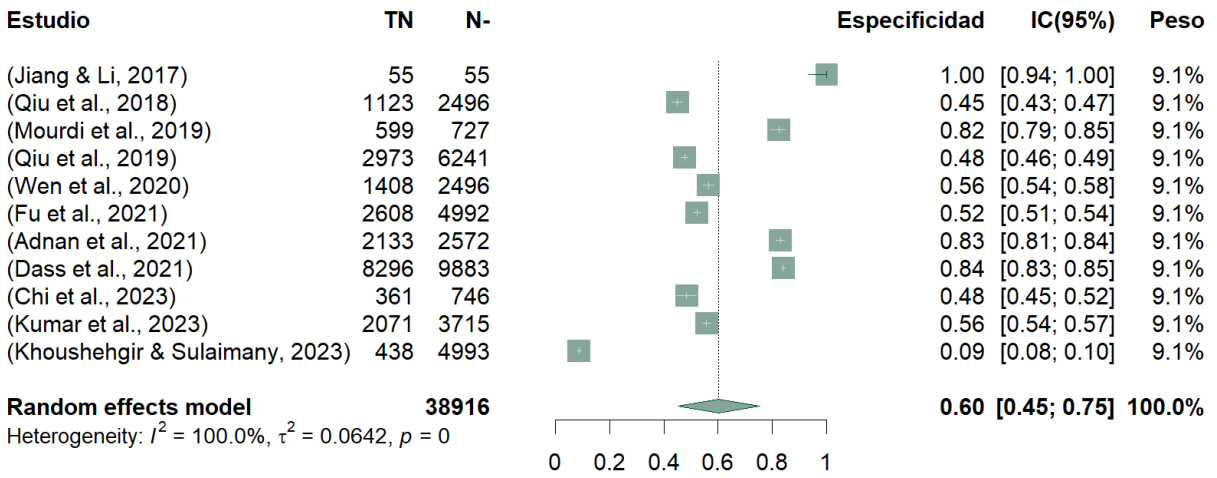

Figura 4.14: *Forest plot* del experimento 1 para la especificidad

4.2.2. Experimento 2

El segundo experimento realiza el meta-análisis mediante la transformación *arcoseno* y el método de Inverso de la Varianza. El *forest plot* de resultados se muestra en la figura 4.15.

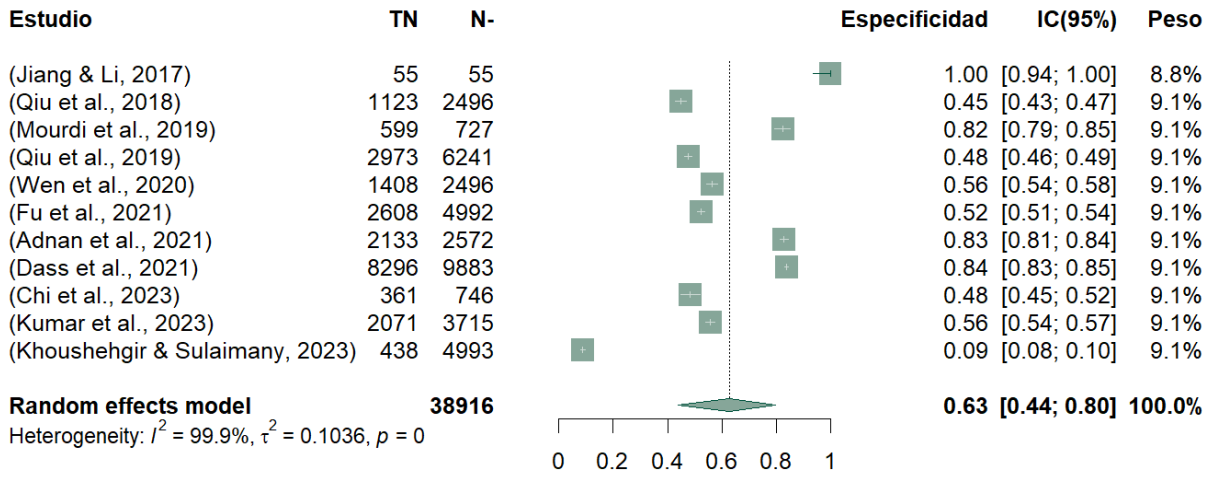

Figura 4.15: *Forest plot* del experimento 2 para la especificidad

El efecto agrupado obtenido es de 0.63 con intervalo de confianza $[0.44; 0.80]$. Esta estimación está levemente por encima a la del experimento 1, pero igual de imprecisa. Los pesos asignados son prácticamente los mismos otra vez.

Se puede ver la varianza inter-estudios es de 0.1036 (en escala transformada), de nuevo, un orden de magnitud por encima del experimento 1 de sensibilidad. Las otras métricas de heterogeneidad son similares a la del resto de experimentos. Se obtiene una $I^2 = 99.9\,\%$, levemente por debajo del experimento 1, y el test Q obtiene un p-valor de 0.

4.2.3. Experimento 3

En el experimento 3 se ha utilizado el modelo B-N mediante GLMM y transformación *logit*. Se recuerda que este método no adjudica pesos a los estudios. Los resultados se muestran en el *forest plot* de la figura 4.16.

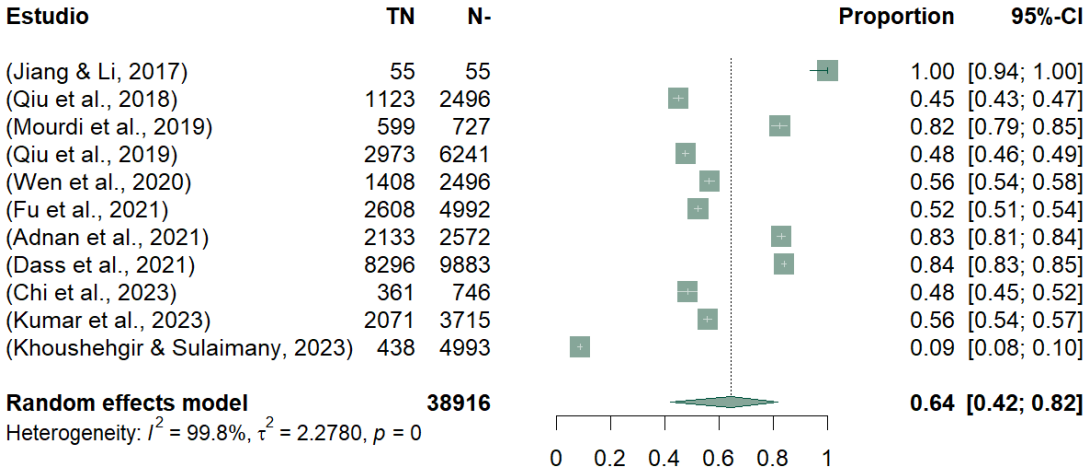

Figura 4.16: *Forest plot* del experimento 3 para la especificidad

La medida agrupada es de 0.64 con un intervalo de confianza del 95 % en $[0.42; 0.82]$. Un valor levemente por encima del valor obtenido en los experimentos anteriores.

Se obtiene un índice I^2 que concluye que el 99.8 % de la heterogeneidad está ocasionado por la varianza inter-estudios, que en este caso obtiene un valor de 2.278, un valor muy superior al obtenido en la sensibilidad. El test Q obtiene un *p-valor* de 0, al igual que en los experimentos anteriores.

4.2.4. Experimento 4

En este experimento se describen los resultados del modelo N-N bayesiano con los *priors* débilmente informativos mencionados en el capítulo anterior. Los resultados completos se muestran en el *forest plot* de la figura 4.17.

El experimento obtiene una especificidad agrupada de 0.60 con un intervalo de credibilidad del 95 % en $[0.41; 0.78]$. Una medida muy similar al modelo N-N frecuentista, aunque con intervalos más amplios. Este fenómeno también se pudo observar en experimento 4 del meta-análisis de sensibilidad. Se ha obtenido un valor puntual de τ^2 de 0.0938. Este valor es un 50 % más alto que el obtenido en el experimento 1.

En la figura 4.18 se muestran la distribución a posteriori de la varianza inter-estudios (izquierda) y la especificidad agrupada (derecha). En ambos casos tienen una forma similar al modelo N-N del meta-análisis de sensibilidad, aunque en este caso son más anchas y con colas más largas.

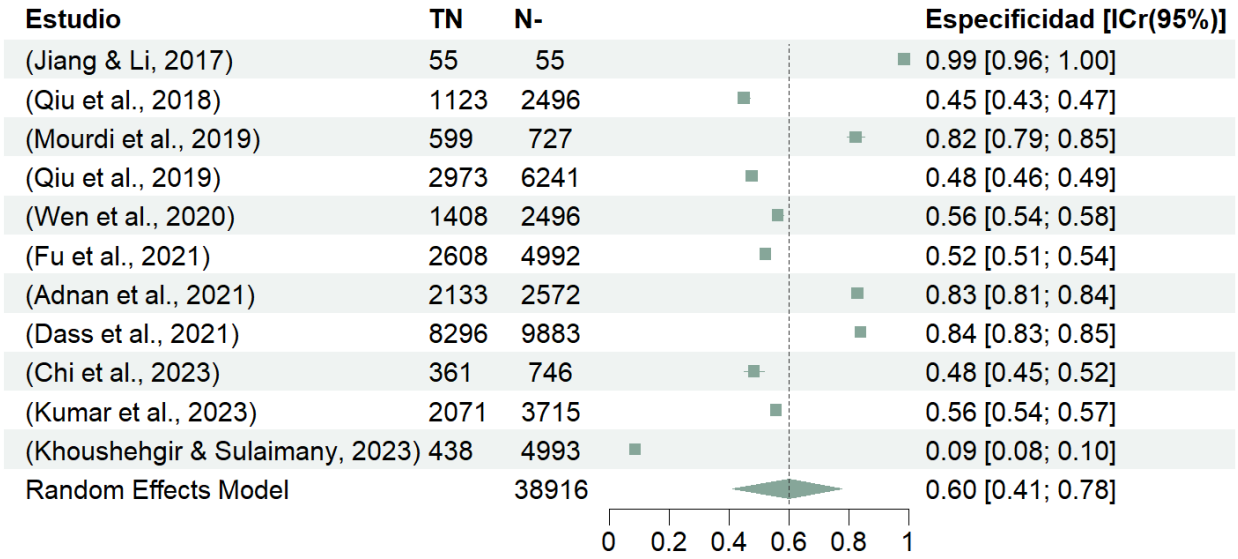

Figura 4.17: *Forest plot* del experimento 4 para la especificidad

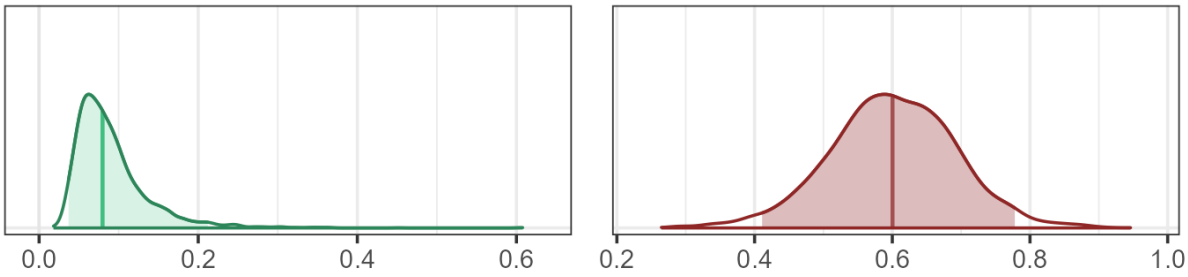

Figura 4.18: Distribución de la varianza inter-estudios τ^2 (izquierda) y la medida de efecto agrupada θ (derecha) del experimento 4 de la especificidad.

4.2.5. Experimento 5

En este apartado se muestran los resultados del modelo B-N bayesiano con la especificidad como medida de efecto agrupada. En el *forest plot* de la figura 4.19 se muestran los resultados.

Se ha obtenido un efecto agrupado de 0.63 con intervalo de credibilidad de $[0.42; 0.82]$, un valor ligeramente por debajo del efecto estimado por el semejante frecuentista. El intervalo de credibilidad estimado es exactamente igual al de confianza del modelo frecuentista. De igual manera, se estima un valor puntual de τ^2 de 2.337, prácticamente igual que el modelo B-N frecuentista.

Las distribuciones a posteriori de ambos parámetros se muestran en la figura 4.20. De nuevo, presenta formas similares a la del resto de modelos bayesianos.

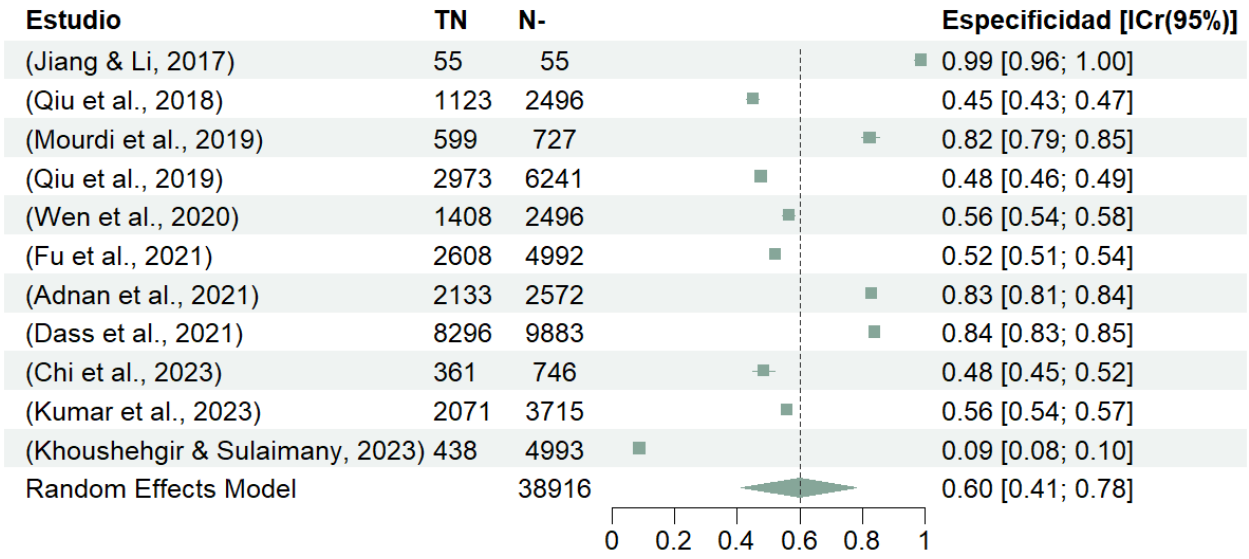

Figura 4.19: *Forest plot* del experimento 5 para la especificidad

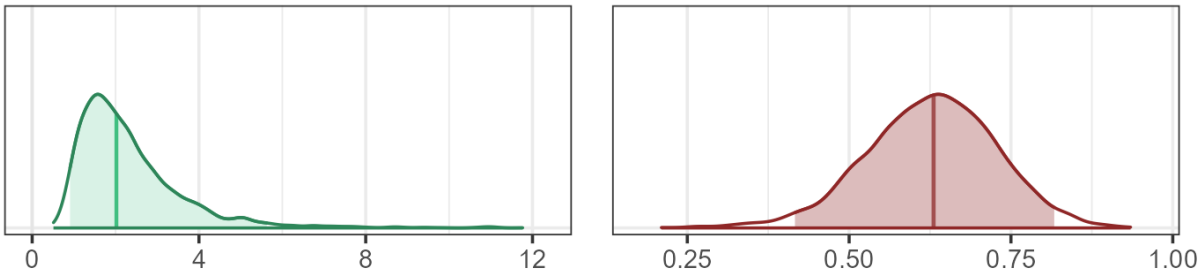

Figura 4.20: Distribución de la varianza inter-estudios τ^2 (izquierda) y la medida de efecto agrupada θ (derecha) del experimento 5 de la especificidad.

4.2.6. Experimento 6

El experimento 6 utiliza la definición de abandono como variable de segmentación para llevar a cabo el análisis de subgrupos. Al igual que en el resto de análisis de subgrupos, se utiliza el método de inverso de la varianza sin transformaciones y los resultados se muestran en la figura 4.21.

Se observa que existen diferencias entre las medidas de efecto de ambos grupos, aunque el solapamiento entre intervalos indica que no son significativas. $G_{participacion}$ obtiene un efecto de $0.56[0.37; 0.75]$. En $G_{aprendizaje}$ es de $0.72[0.49; 0.95]$. Los intervalos en ambos grupos son de una amplitud similar.

Los índices de heterogeneidad son ligeramente distintos en ambos grupos. $G_{participacion}$ obtiene un valor de I^2 del 99.4 % y τ_g^2 de 0.0723. En $G_{aprendizaje}$ es de 99.4 % y 0.0496

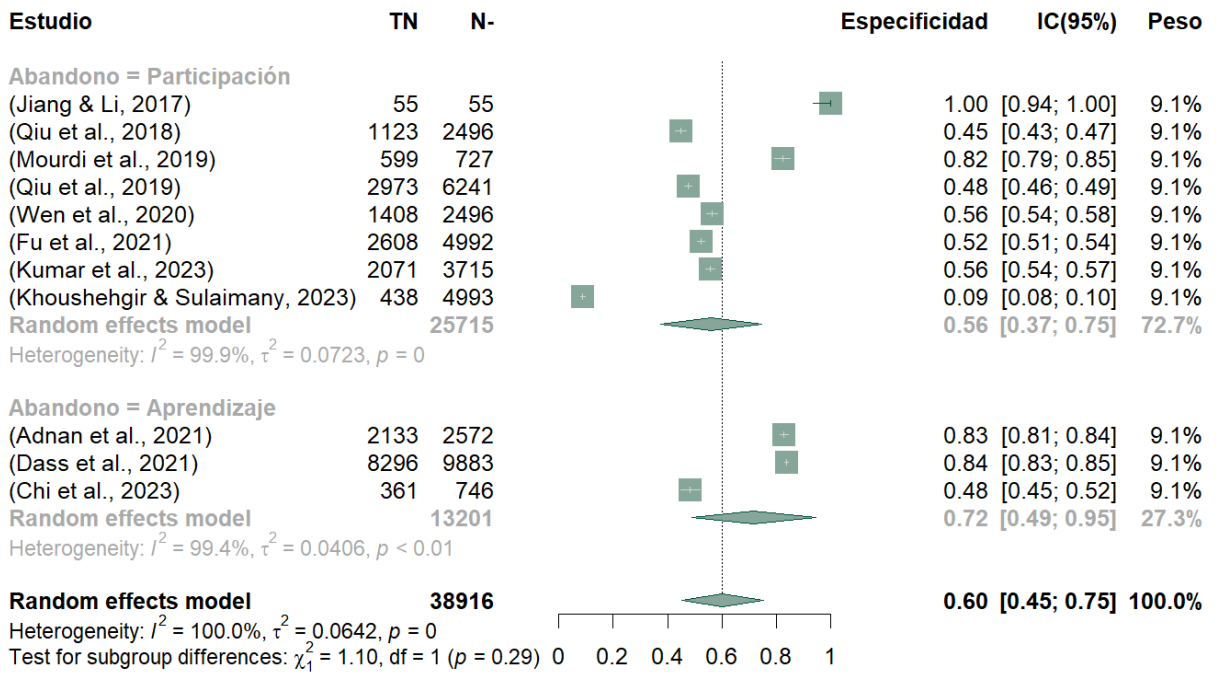

Figura 4.21: *Forest plot* del experimento 6 para la especificidad

respectivamente. Sin embargo, en los dos grupos se encuentran diferencias significativas entre estudios mediante el test Q.

El test χ^2 para diferencias de subgrupos concluye que no existen diferencias significativas entre ambos grupos ($p = 11$).

Se ha podido ver cierta semejanza con el experimento 6 de la sensibilidad. En ambos casos la medida de efecto de $G_{aprendizaje}$ es algo superior a la de $G_{participacion}$. Esto puede indicar que existe cierta relación entre el tipo de definición y el valor de la medida de efecto.

4.2.7. Experimento 7

El experimento 7 realiza el meta-análisis segmentando los estudios en función del conjunto de datos que han utilizado. Por un lado los 6 estudios que utilizan *KDD Cup 2015*, que forman G_{KDD}, y por el otro los 7 restantes (G_{otros}).

La especificidad agrupada obtenida por G_{otros} es $0.80[0.63; 0.96]$. Una medida por encima de la global, y con intervalos más estrechos. La heterogeneidad en este este grupo se ha reducido respecto al global ($I^2 = 99.3\,\%$ y $\tau_g^2 = 0.0353$) aunque se siguen encontrado diferencias significativas entre los estudios mediante el test Q ($p < 0.01$). En la figura se puede ver que la mayor parte de estudios están bastante próximos excepto Chi et al. [2023].

En G_{KDD} se obtiene una especificidad agrupada de $0.44[0.30; 0.59]$. La medida es más precisa que el global al igual que en G_{otros}. Aunque en este caso el valor puntual está por debajo del global. La heterogeneidad es similar también al otro grupo, con $I^2 = 99.9\,\%$ y $\tau_g^2 = 0.0323$, en ambos casos por debajo del global. Pero el test Q sigue encontrando

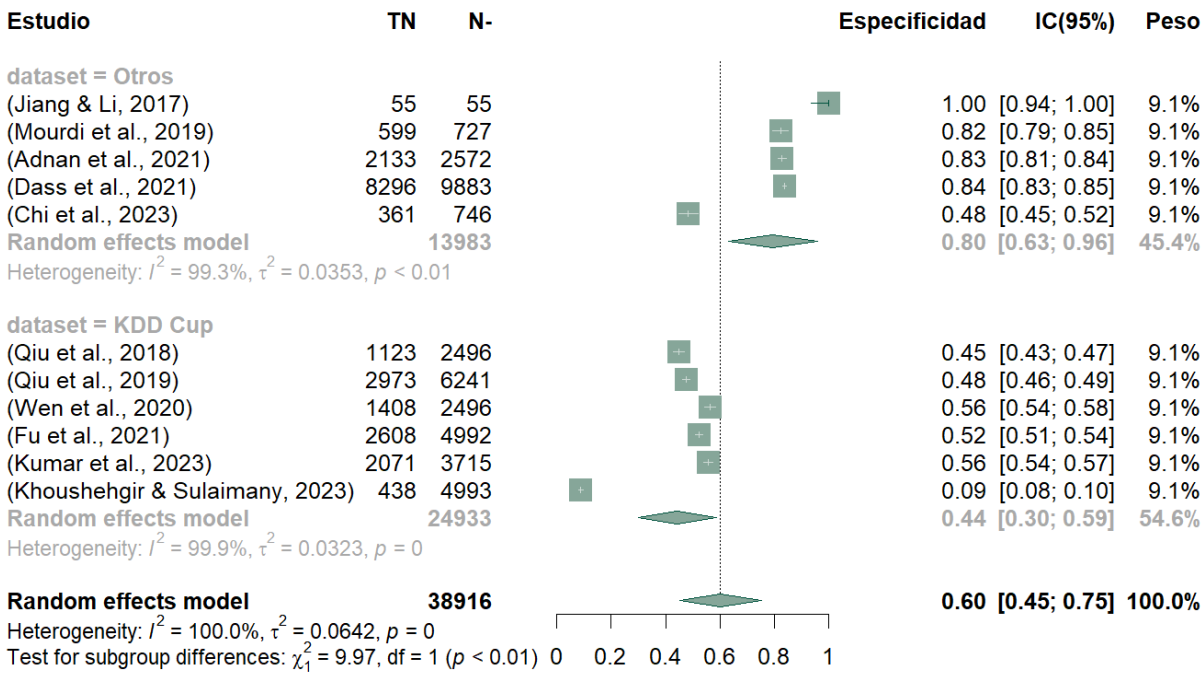

Figura 4.22: *Forest plot* del experimento 7 para la especificidad

diferencias significativas. En este grupo, la gran mayoría de medidas de efecto se encuentran muy próximas excepto Khoushehgir and Sulaimany [2023].

El test χ^2 para diferencias entre subgrupos concluye que sí existen diferencias significativas entre subgrupos ($p < 0.01$). Esto indica que existe una relación entre el tipo de conjunto de datos y la especificidad obtenida. Tal efecto podría ser generalizado para las dos medidas. Ya que en el experimento 7 las diferencias entre medidas de efecto agrupada eran similares, pese a que no se pudieron observar diferencias significativas.

4.2.8. Experimento 8

En este experimento se ha segmentado por el tipo de modelo utilizado para el análisis de subgrupos, dividiendo entre aquellos estudios que utilizan algoritmos de aprendizaje automático (G_{ML}) y los *Deep Learning* (G_{DL}). Ambos grupos están compuestos por 5 y 6 estudios respectivamente. En la figura 4.23 se muestra el *forest plot* con los resultados.

La medida de efecto es muy similar en ambos grupos, con 0.57 y 0.63 en G_{ML} y G_{DL} respectivamente. Aunque los intervalos son muy dispares: en caso de G_{DL} es $[0.50; 0.75]$, mucho más preciso que en G_{ML}, que obtiene $[0.26, 0.89]$.

En cuanto a la heterogeneidad, en ambos grupos el test Q no contempla similitudes entre los estudios ($p = 0$). El índice I^2 concluye que en G_{ML} el 100 % de la variabilidad está ocasionada por la varianza inter-estudios, y en G_{DL} el 99.7 %. El valor de τ_g^2 es de 0.1279 en el caso de G_{ML}, un valor muy por encima del global. Mientras en G_{ML} obtiene 0.243, un

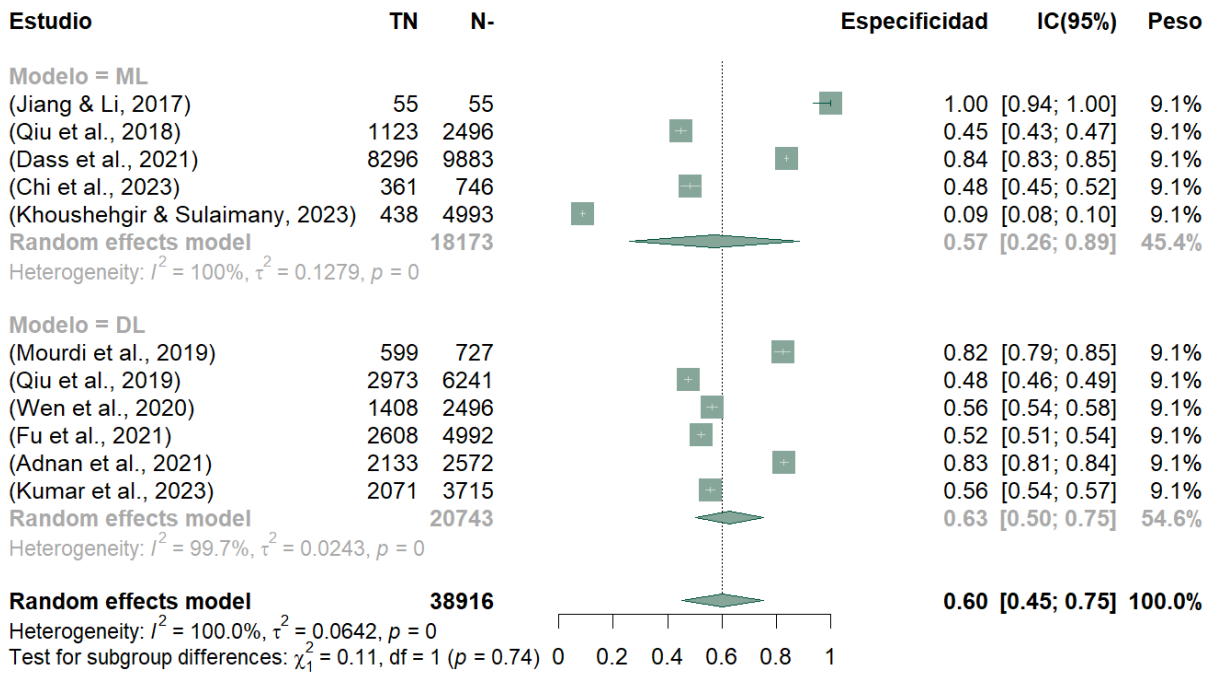

Figura 4.23: *Forest plot* del experimento 8 para la especificidad

valor inferior al global.

No se encuentran diferencias significativas entre ambos subgrupos mediante el test χ^2 ($p = 0.74$).

Se encontraron resultados similares en el meta-análisis de sensibilidad, con medidas de efecto agrupadas muy similares a la global. Esto es un indicio de que el tipo de modelo es independiente de su rendimiento.

4.2.9. Experimento 9

El último análisis de subgrupos para la especificidad se realiza sobre el tipo de modelado. Por un lado, G_{plano} utiliza el modelado plano y $G_{secuencia}$ el de secuencia.

En $G_{secuencia}$ se obtiene una especificidad agrupada de $0.71[0.57, 0.84]$. Un valor similar al global y algo más preciso. La heterogeneidad se ha reducido ligeramente. El valor de τ_g^2 ha visto reducido hasta 0.0396 e I^2 hasta 99.9 %. Aunque se siguen encontrando diferencias significativas entre estudios.

En el caso de G_{plano}, la medida de efecto es extremadamente imprecisa, se obtiene un valor de $0.50[0.03, 0.97]$. Las medidas de efecto de los dos estudios están casi 0.4 puntos alejadas entre sí. Esto provoca que el valor de τ_g^2 esté muy por encima del global, alcanzando un valor de 0.783 . El resto de índices de heterogeneidad vienen dados por un índice I^2 del 99.8 % y diferencias significativas en el test $Q(p < 0.01)$.

El test de diferencias entre grupos no encuentra diferencias significativas ($p = 0.41$).

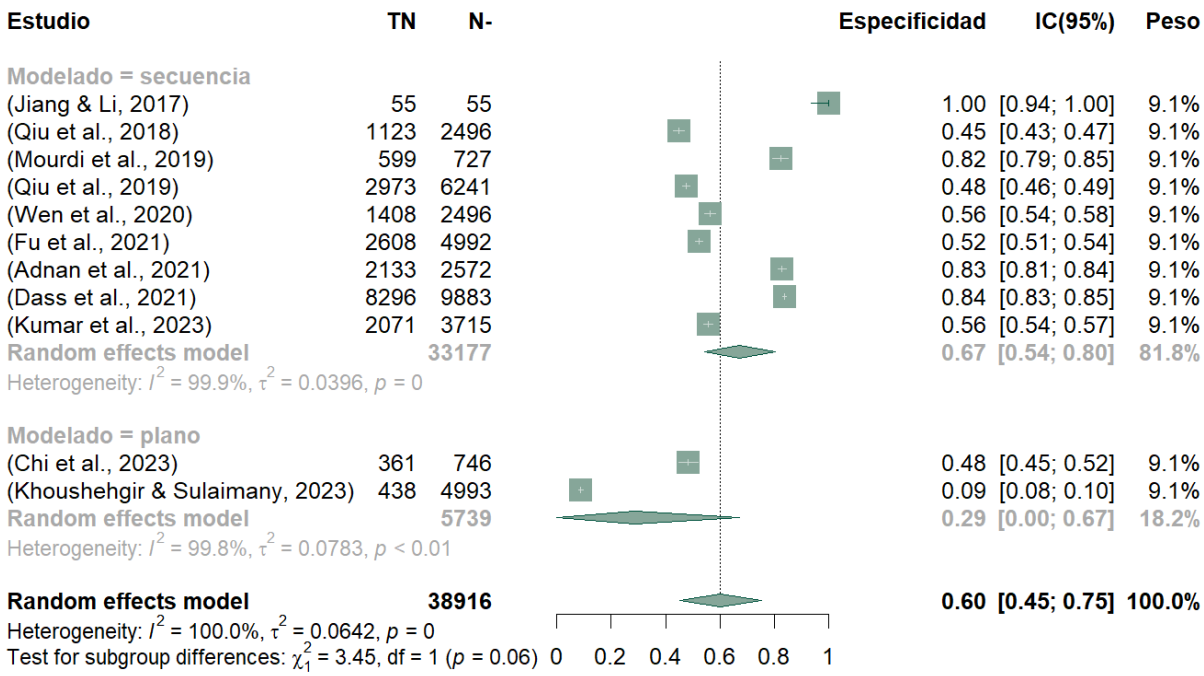

Figura 4.24: *Forest plot* del experimento 9 para la especificidad

En el experimento 9 de sensibilidad se encontraron diferencias significativas entre los dos grupos, principalmente debido a que los dos estudios de G_{plano} tenían medidas muy similares. En este caso es todo lo contrario, ambas medidas están muy alejadas. Es difícil de sacar patrones en este subgrupo debido a que tiene muy pocos estudios.

4.2.10. Experimento 10

En el último experimento se ha observado mediante la técnica de *Leave One Out* que uno de los estudios es extremadamente más influyente que el resto (tabla 4.2). Y es que dejar fuera a Khoushehgir and Sulaimany [2023] provoca que todas las medidas se vean muy afectadas. La especificidad agrupada alcanza el valor de 0.65, mientras que quitando cualquiera del resto de los estudios se mantiene en valores entre 0.58 y 0.61. De igual manera, despreciar este estudio logra que τ^2 baje desde 0.0642 hasta 0.0388. Además, el estudio mencionado logra que el índice I^2 se reduzca hasta 99.8 %, el único que hace que baje de 100 % junto con Dass et al. [2021], aunque este último no afecta tanto al resto de métricas.

La razón de la influencia de este estudio es que tiene un especificidad extremadamente baja y precisión muy alta: $0.09[0.08, 0.10]$. Es un valor muy dispar con el resto de estudios. Se han re-analizado los datos extraídos en búsqueda de errores y se ha podido comprobar que los datos utilizados son totalmente correctos. En el capítulo siguiente se tratará este tema.

Esta técnica ha sido ejecutada de forma iterativa y se ha comprobado que, al igual que

Tabla 4.2: Resultados del análisis de influyentes en el meta-análisis de sensibilidad

Estudio	Especificidad	τ^2	I^2
(Jiang & Li, 2017)	0.5634 [0.4217; 0.7051]	0.0522	100.0 %
(Qiu et al., 2018)	0.6184 [0.4561; 0.7807]	0.0685	100.0 %
(Mourdi et al., 2019)	0.5810 [0.4224; 0.7397]	0.0654	100.0 %
(Qiu et al., 2019)	0.6158 [0.4524; 0.7792]	0.0694	100.0 %
(Wen et al., 2020)	0.6070 [0.4415; 0.7725]	0.0712	100.0 %
(Fu et al., 2021)	0.6112 [0.4464; 0.7759]	0.0706	100.0 %
(Adnan et al., 2021)	0.5804 [0.4222; 0.7387]	0.0651	100.0 %
(Dass et al., 2021)	0.5794 [0.4218; 0.7370]	0.0645	99.9 %
(Chi et al., 2023)	0.6150 [0.4513; 0.7786]	0.0696	100.0 %
(Kumar et al., 2023)	0.6077 [0.4422; 0.7731]	0.0711	100.0 %
(Khoushehgir & Sulaimany, 2023)	0.6547 [0.5324; 0.7770]	0.0388	99.8 %
Global	0.6031 [0.4532; 0.07530]	0.0642	100.0 %

en el meta-análisis de sensibilidad, se requiere eliminar 8 estudios para que ver cambios significativos en la heterogeneidad. Se ha alcanzado un valor de 64.7 % para I^2 y un p-valor de 0.059 para el test Q. El valor de la especificidad agrupada en este meta-análisis es de $0.47[0.45, 0.49]$. Los estudios implicados son Qiu et al. [2018], Qiu et al. [2019] y Chi et al. [2023]. Tanto en el meta-análisis de sensibilidad como en este ha sido necesario eliminar 8 estudios para alcanzar niveles aceptables de heterogeneidad.

Capítulo 5

Discusión

En este capítulo se trata de responder las preguntas de investigación. En cada una de las siguientes secciones se incluye la respuesta correspondiente y las principales limitaciones que implica.

5.1. Pregunta de investigación 1

La primera pregunta de investigación decía: «¿En qué medida son útiles las técnicas de aprendizaje automático para la detección de abandono en MOOC?»

En este sentido, se han realizado varios experimentos considerando varios tipos de modelos y transformaciones, y se puede decir que existe un consenso en torno al valor de las medidas de efecto obtenidas. Además de un enfoque tradicional frecuentista, se ha probado un enfoque bayesiano también con resultados similares. Todos los experimentos han arrojado valores muy similares de las medidas de efecto, que en general son bastante altos, especialmente para la sensibilidad. Por lo tanto, se puede afirmar que las técnicas de aprendizaje automático son capaces de predecir el abandono en este tipo de cursos con un rendimiento realmente bueno. Sin embargo, esta afirmación debe ser tomada con cautela, especialmente por dos motivos:

- Los niveles de heterogeneidad son altos en todos los experimentos.
- Existen grandes diferencias entre las dos medidas de efecto analizadas.

En general, todos los estudios han obtenido valores extremadamente altos del índice I^2, siempre cercanos a 100 %; y p-valores muy próximos a 0 en el test Q, a pesar de que la varianza inter-estudios parecía aceptable. Estos **niveles de heterogeneidad** pueden provenir de la variabilidad entre las características de los estudios. En este trabajo se han extraído algunas características de cada estudio, y se ha visto que cada uno utiliza definiciones de abandono, algoritmos, modelados del estudiante, o conjuntos de datos distintos. Esto puede poner en compromiso uno de los requisitos del meta-análisis: que las medidas de efecto no sean comparables. Para comprobarlo, en este trabajo se ha realizado una primera aproximación

mediante la técnica de análisis de subgrupos y se ha podido ver que algunas características como la definición de abandono o el conjunto de datos utilizado son relevantes en el valor obtenido de la medida global. Este tema se tratará con más detenimiento en la siguiente sección. No obstante, hay que tener en cuenta que existen más características de los estudios que los hacen únicos, que no han sido analizadas y pueden aumentar la heterogeneidad.

Desde el punto de vista de Barker et al. [2021] los altos niveles de heterogeneidad no son un problema ya que es esperable que en un meta-análisis de proporciones la heterogeneidad sea alta. Esto es debido principalmente a los problemas del test Q y el índice I^2, tal y como se explica en el capítulo 2, estos índices tienen problemas con el número de estudios del meta-análisis y con la precisión de los estudios. En este trabajo se ha experimentado un indicio de este efecto al aplicar la técnica de análisis de influyentes. Se ha visto que ningún estudio tenía un impacto relevante en el análisis, pero que se podían alcanzar niveles aceptables de heterogeneidad reduciendo el número de estudios considerablemente. Ambas medidas de efecto coincidían en el número de estudios a excluir, pero no en los estudios.

Sin embargo, a pesar de que la principal causa de heterogeneidad sean la variabilidad de las características de los estudios y las propiedades del análisis de proporciones, es posible que exista cierta heterogeneidad causada por otras fuentes, concretamente:

1. El **criterio para seleccionar** una única medida dentro de cada estudio. Se escogió "el mejor resultado entre los más tempranos". El problema es que no existe una cantidad de tiempo mínima común. Por ejemplo, mientras que unos estudios utilizan los datos de una semana, otros utilizan tres. Por lo tanto, "el más temprano" se refiere a momentos temporales distintos dependiendo del estudio. Esto provoca un aumento de la heterogeneidad, ya que existe cierta correlación entre el momento de la predicción y el rendimiento del modelo [Deeva et al., 2022].

2. La **correlación entre las medidas de efecto**. Es conocido que entre las medidas sensibilidad y especificidad existe un compromiso provocado por el umbral de decisión [Kim et al., 2015]. Modificando este umbral hace que aumente una de ellas y se rebaje la otra. Esta correlación puede desembocar en un aumento de la heterogeneidad si no se tiene en cuenta en el modelo.

El otro motivo para poner en cuestión los resultados de este meta-análisis es que existen **grandes diferencias entre los resultados de sensibilidad y especificidad**. En general, la especificidad obtiene resultados más bajos y, sobre todo, más variables que la sensibilidad. En el meta-análisis de sensibilidad se ha obtenido una medida de efecto global entre 0.93 y 0.95 con intervalos de apenas ± 0.05, mientras que en el de especificidad se han sido desde 0.60 hasta 0.64 con intervalos que abarcan prácticamente todo el rango de valores posibles.

La sensibilidad, también conocida como exhaustividad, es una medida muy usada dentro de la ciencia de datos y en concreto en la predicción de abandono (tal y como comenta Prenkaj

et al. [2021]). No ocurre lo mismo con la especificidad, que en su lugar suele utilizarse la precisión. Esto tiene varios problemas. El primero es que si una medida no se publica, es posible que no se busque optimizarla por norma general. En segundo lugar, mientras que la dupla sensibilidad-especificidad tiene en cuenta todos los elementos de la matriz de confusión, al usarse precisión-exhaustividad se dejan de tener en cuenta los verdaderos negativos (TN) en la evaluación del algoritmo. Este es la razón por el que este trabajo utiliza especificidad.

Se desconoce el motivo por el cual se prefiere utilizar la precisión de forma casi generalizada en esta línea de investigación, pero es posible que esté relacionado con el impacto o coste asociado al error. Por ejemplo, tomar cierta medida de intervención en estudiantes sin riesgo de abandono tendrá un efecto negativo muy leve, frente al impacto positivo fuerte que puede tener sobre alumnos en riesgo de abandono.

Estos problemas se ven agravados en conjuntos de datos desbalanceados que, como se ha visto, es un problema común en este tipo de estudios, donde la tasa de abandono elevada de los MOOC hace que el número de casos positivos (abandonos) sea mucho más elevado que el de casos negativos (retención) [Dalipi et al., 2018, Prenkaj et al., 2021, Alhothali et al., 2022]. Este hecho hace que la especificidad sea mucho más variable, ya que decrece mucho más rápido con cada error en sistemas con sensibilidad alta y pocos ejemplos negativos. La alta variabilidad provoca que se vean casos muy extremos, como es el caso de Khoushehgir and Sulaimany [2023], uno de los más influyentes en el análisis de especificidad. Este estudio publicaba una precisión de 0.83, un valor que puede considerarse suficientemente bueno. Sin embargo, el 80 % de abandonos en el conjunto de datos ocasiona que su especificidad esté por debajo de 0.1.

5.2. Pregunta de investigación 2

Respecto a la segunda pregunta, se cuestionaba: «¿Existen diferencias de rendimiento entre sistemas con distintas características?»

Se ha podido comprobar que existen cierta correlación entre las características del estudio y el valor de la medida de efecto. Especialmente aquellas características que tienen relación con los datos, y no con el modelo. Para este trabajo se consideraron 4 características de cada estudio que se creía que podían afectar al rendimiento del sistema, y se ha confirmado en los resultados en algunas de ellas. Más concretamente:

- Cuando se dividen los estudios por **definición de abandono**, ambas medidas de efecto son más altas en aquellos estudios que utilizan la definición basada en objetivos de aprendizaje, aunque estas diferencias no son significativas.

- Sí se han visto diferencias significativas en la especificidad al segmentar por **conjunto de datos**. Aquellos que no utilizan datos de *KDD Cup 2015* obtienen mejores resultados, tendencia que se repite en la sensibilidad aunque sin la suficiente significancia.

- En el caso del **tipo de algoritmo** no se han visto diferencias entre estudios que utilizan aprendizaje automático y los que usan *Deep Learning*.

- Al dividir por **modelado del estudiante** se obtienen valores de efecto agrupado y de significancia muy distintos dependiendo de la medida de efecto.

Este tipo de análisis tiene la limitación importante de que la cantidad de estudios que se incluyen en cada grupo puede ser muy pequeña. Esto aumenta el riesgo de sesgo y el error asociado a la medida obtenida en el grupo. Un ejemplo de esto se puede encontrar en el grupo de modelado plano. Este grupo está formado únicamente por dos estudios y, tal y como se mencionaba antes, los resultados de las dos medidas de efecto son completamente distintos. Mientras que el análisis de sensibilidad conseguía una medida global muy precisa logrando valores muy aceptables de heterogeneidad, en el análisis de especificidad ocurría todo lo contrario, las dos medidas estaban muy separadas, provocando así unos índices muy altos de heterogeneidad con tan sólo dos estudios.

5.3. Pregunta de investigación 3

Por último, en la tercera pregunta de investigación se planteó: «¿Es viable utilizar la técnica del meta-análisis para evaluar el rendimiento global de técnicas de aprendizaje automático?»

Este trabajo resulta innovador al tratar de aplicar la técnica del meta-análisis para evaluar el rendimiento de sistemas de aprendizaje automático fuera del área de la medicina, y en especial en predicción de abandono en MOOC. Es importante recalcar que se ha logrado obtener una medida de efecto global y que las técnicas que se han utilizado están asentadas en la comunidad científica. No se ha encontrado ningún estudio que realice un meta-análisis similar al propuesto en este trabajo. Aunque Fahd et al. [2022] indica realizar un meta-análisis, al final únicamente elabora una distribución de las métricas de evaluación.

Se ha podido comprobar que los métodos tradicionales de meta-análisis no están lo suficientemente desarrollados para la evaluación del rendimiento de tareas que usan aprendizaje automático. Por lo que es importante continuar investigando hasta alcanzar un marco de trabajo bien asentado dentro de la ciencia de datos. En este punto, se han identificado 5 principales problemas o retos a los que se enfrenta:

1. El primero es la **medida de efecto** que se debería utilizar. En este trabajo se planteó usar la sensibilidad y especificidad, como es común en gran parte de las publicaciones en medicina. Se ha visto que utilizar este tipo de medidas puede ser una fuente de heterogeneidad por diversos motivos expuestos anteriormente. En Azeem et al. [2019] se mencionaba que lo ideal sería utilizar la AUC como medida única como medida de efecto pero no se suele publicar.

2. En segundo lugar se plantea el problema con las **medidas de heterogeneidad** tradicionales, las cuales no son adecuadas para este tipo de meta-análisis. Tal y como se ha contemplado en este capítulo, el test Q y el índice I^2 no son adecuados para meta-análisis basados en proporciones.

3. También existen dificultades para estimar el **sesgo de publicación**. Los gráficos de embudo no han dado evidencias claras de la presencia de sesgo de publicación, a pesar de que se introdujo de forma evidente al seleccionar únicamente estudios publicados en revistas con revisión por pares. Según Barker et al. [2021], ni este tipo de gráfico ni ningún otro método tradicional para evaluar el sesgo de publicación son válidos en un meta-análisis de proporciones.

4. Otro aspecto importante es la falta de **reproducibilidad** en los artículos. Como se ha comentado, se decidió excluir aquellos estudios que no habían sido publicados en revistas con revisión por pares. Se tomó esta decisión para tratar de aumentar la calidad de los estudios obtenidos y además, adecuar la carga de trabajo. Sin embargo, en la mayoría de artículos ha sido imposible obtener los datos necesarios para llevar a cabo el meta-análisis, y en los restantes han existido serias dificultades. Especialmente los datos necesarios para el cálculo del error asociado a la medida como el tamaño del conjunto de evaluación o la proporción de abandonos, datos que son fundamentales para garantizar la reproducibilidad de los artículos.

5. El último reto consiste en reducir la gran **variabilidad de características** que existe en los problemas de aprendizaje automático. Mientras que en medicina el conjunto de estudios suele ser bastante homogéneo debido a que suelen estar todas las variables controladas. En aprendizaje automático se utilizan datos históricos pasados que frecuentemente han sido obtenidos para otros propósitos, dando lugar a que cada estudio tenga características casi únicas. Esto puede atacar directamente al requisito de comparabilidad de los estudios, por ello, la definición de requisitos de exclusión, la extracción de características y el análisis de subgrupos cobran especial relevancia en este tipo de meta-análisis.

Capítulo 6

Conclusiones y trabajos futuros

A lo largo de este trabajo se ha realizado un análisis cuantitativo de la aplicación de técnicas de aprendizaje automático y *deep learning* para la predicción de abandono de estudiantes en MOOC. Se ha podido concluir, pese a los niveles altos de heterogeneidad, que estos sistemas resultan útiles para la detección de abandono, puesto que son capaces de detectar más del 90 % de estudiantes que abandonan en la mayoría de los estudios realizados. Aunque estos sistemas tienen una alta tasa de falsos negativos, dando lugar a errores al predecir los estudiantes que completan el curso.

Este trabajo no ha tenido en cuenta factores importantes en el análisis como puede ser el momento de la predicción del abandono. De igual manera, el tamaño de la muestra utilizada puede considerarse reducida. Estos dos factores, unidos a la reducción de la heterogeneidad, aumentarían considerablemente el poder estadístico de este meta-análisis. A continuación, se proponen algunas vías para futuros trabajos:

- Extender el meta-análisis teniendo en cuenta el momento de predicción como una variable. Un posible enfoque sería contemplar usar modelos de meta-regresión.
- Explorar el uso de otras métricas más representativas del campo como precisión y exhaustividad, o medidas únicas como F1 o AUC.
- También se podría tener en cuenta la correlación entre medidas mediante un análisis bivariante. El denominado meta-análisis de diagnóstico tiene en cuenta este aspecto entre otros.
- Explorar los intervalos de predicción como medida de heterogeneidad.
- Extender el análisis aplicando unos criterios de selección menos restrictivos, incorporando la denominada literatura gris.
- Explorar la heterogeneidad en grupos más reducidos utilizando variables de menor granularidad y variables múltiples en análisis de subgrupos.

- En la vía de los modelos bayesianos, se podría explorar el efecto de los *priors*. Un buen punto de partida sería utilizar las distribuciones a posteriori de este trabajo.

Bibliografía

Research and Markets. Massive Open Online Courses - Global Strategic Business Report, 2024. URL https://www.researchandmarkets.com/reports/5140372/massive-open-online-courses-global-strategic.

United States Securities and Exchange Commision. Coursera, INC.; form 10-k, 2022. URL https://www.sec.gov/ix?doc=/Archives/edgar/data/1651562/000095017023004143/cour-20221231.htm.

2U. 2U releases 2022 transparency & outcomes report, 2023. URL https://2u.com/newsroom/2022-transparency-outcomes-report/.

International Centre For Engineering Education (ICEE). The first global partner of ICEE — XuetangX exceeds 100 million users worldwide, 2022. URL http://www.icee-unesco.org/news/125#.

Times of India. With 3 crore enrolments, Swayam tops other eLearning platforms, 2023. ISSN 0971-8257. URL https://timesofindia.indiatimes.com/education/news/with-3-crore-enrolments-swayam-tops-other-elearning-platforms/articleshow/97810091.cms.

Dhawal Shah. By The Numbers: MOOCs in 2020, 2020. URL https://www.classcentral.com/report/mooc-stats-2020/.

Christian Gütl, Rocael Hernández Rizzardini, Vanessa Chang, and Miguel Morales. Attrition in mooc: Lessons learned from drop-out students. In *Learning Technology for Education in Cloud. MOOC and Big Data: Third International Workshop, LTEC 2014, Santiago, Chile, September 2-5, 2014. Proceedings 3*, pages 37–48. Springer, 2014.

Nikolaos Floratos, Teresa Guasch, and Anna Espasa. Recommendations on formative assessment and feedback practices for stronger engagement in moocs. *Open Praxis*, 7(2): 141–152, 2015.

Daniel F. O. Onah, Jane Sinclair, and Russell Boyatt. Dropout rates of massive open online courses : behavioural patterns. In L. Gómez Chova, A. López Martínez, and I. Candel Torres, editors, *EDULEARN14 Proceedings*, pages 5825–5834, Barcelona, Spain, 2014.

IATED Academy. ISBN 978-84-617-0557-3. URL https://wrap.warwick.ac.uk/65543/. ISSN: 2340-1117.

Tieyuan Liu, Qiong Wu, Liang Chang, and Tianlong Gu. A review of deep learning-based recommender system in e-learning environments. *Artificial Intelligence Review*, 55(8): 5953–5980, 2022.

Prakhar Bhardwaj, PK Gupta, Harsh Panwar, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, and Anubha Bhaik. Application of deep learning on student engagement in e-learning environments. *Computers & Electrical Engineering*, 93:107277, 2021.

Gábor Kőrösi and Richard Farkas. Mooc performance prediction by deep learning from raw clickstream data. In *Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4*, pages 474–485. Springer, 2020.

Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J Muñoz-Merino, Iria Estévez-Ayres, and Carlos Delgado Kloos. Sentiment analysis in moocs: A case study. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1489–1496. IEEE, 2018.

Pedro Manuel Moreno-Marcos, Pedro J. Muñoz-Merino, Jorge Maldonado-Mahauad, Mar Pérez-Sanagustín, Carlos Alario-Hoyos, and Carlos Delgado Kloos. Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. *Computers & Education*, 145:103728, February 2020. ISSN 0360-1315. doi: 10.1016/j.compedu.2019.103728. URL https://www.sciencedirect.com/science/article/pii/S0360131519302817.

Fisnik Dalipi, Ali Shariq Imran, and Zenun Kastrati. MOOC dropout prediction using machine learning techniques: Review and research challenges. In *2018 IEEE Global Engineering Education Conference (EDUCON)*, pages 1007–1014, April 2018. doi: 10.1109/EDUCON.2018.8363340. ISSN: 2165-9567.

Pedro Manuel Moreno-Marcos, Carlos Alario-Hoyos, Pedro J. Muñoz-Merino, and Carlos Delgado Kloos. Prediction in MOOCs: A Review and Future Research Directions. *IEEE Transactions on Learning Technologies*, 12(3):384–401, July 2019. ISSN 1939-1382. doi: 10.1109/TLT.2018.2856808. Conference Name: IEEE Transactions on Learning Technologies.

Muhammad Ilyas Azeem, Fabio Palomba, Lin Shi, and Qing Wang. Machine learning techniques for code smell detection: A systematic literature review and meta-analysis. *Information and Software Technology*, 108:115–138, April 2019. ISSN 0950-5849. doi: 10.1016/j.infsof.2018.12.009. URL https://www.sciencedirect.com/science/article/pii/S0950584918302623.

Michael Ayitey Junior, Peter Appiahene, Obed Appiah, and Christopher Ninfaakang Bombie. Forex market forecasting using machine learning: Systematic Literature Review and meta-analysis. *Journal of Big Data*, 10(1):9, January 2023. ISSN 2196-1115. doi: 10.1186/s40537-022-00676-2. URL https://doi.org/10.1186/s40537-022-00676-2.

Michael Borenstein, editor. *Introduction to meta-analysis*. Wiley, Chichester, nachdr. edition, 2013. ISBN 978-0-470-05724-7.

Mathias Harrer, Pim Cuijpers, Toshi A Furukawa, and David D Ebert. *Doing Meta-Analysis With R: A Hands-On Guide*. Chapman & Hall/CRC Press, Boca Raton, FL and London, 1st edition, 2021. ISBN 978-0-367-61007-4. URL https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074.

Gerta Rücker, Guido Schwarzer, James R. Carpenter, and Martin Schumacher. Undue reliance on I2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, 8(1):79, November 2008. ISSN 1471-2288. doi: 10.1186/1471-2288-8-79. URL https://doi.org/10.1186/1471-2288-8-79.

Matthew J. Page, Jonathan A. C. Sterne, Julian P. T. Higgins, and Matthias Egger. Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods*, 12(2):248–259, 2021a. ISSN 1759-2887. doi: 10.1002/jrsm.1468. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1468. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1468.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995. ISBN 0-429-25841-0.

Christopher G. Thompson and Brandie Semma. An alternative approach to frequentist meta-analysis: A demonstration of Bayesian meta-analysis in adolescent development research. *Journal of Adolescence*, 82:86–102, July 2020. ISSN 0140-1971. doi: 10.1016/j.adolescence.2020.05.001. URL https://www.sciencedirect.com/science/article/pii/S0140197120300713.

Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. .º'Reilly Media, Inc.", 2022.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Areej Alhothali, Maram Albsisi, Hussein Assalahi, and Tahani Aldosemani. Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review. *Sustainability*, 14(10):6199, January 2022. ISSN 2071-1050. doi: 10.3390/su14106199. URL https://www.mdpi.com/2071-1050/14/10/6199. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Bardh Prenkaj, Paola Velardi, Giovanni Stilo, Damiano Distante, and Stefano Faralli. A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. *ACM Computing Surveys*, 53(3):1–34, May 2021. ISSN 0360-0300, 1557-7341. doi: 10.1145/3388792. URL https://dl.acm.org/doi/10.1145/3388792.

Jing Chen, Jun Feng, Xia Sun, Nannan Wu, Zhengzheng Yang, and Sushing Chen. MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine. *Mathematical Problems in Engineering*, 2019:1–11, March 2019. ISSN 1024-123X, 1563-5147. doi: 10.1155/2019/8404653. URL https://www.hindawi.com/journals/mpe/2019/8404653/.

Cong Jin. MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interactive Learning Environments*, 31(2):714–732, February 2023. ISSN 1049-4820, 1744-5191. doi: 10.1080/10494820.2020.1802300. URL https://www.tandfonline.com/doi/full/10.1080/10494820.2020.1802300.

Zengxiao Chi, Shuo Zhang, and Lin Shi. Analysis and Prediction of MOOC Learners' Dropout Behavior. *Applied Sciences*, 13(2):1068, January 2023. ISSN 2076-3417. doi: 10.3390/app13021068. URL https://www.mdpi.com/2076-3417/13/2/1068.

Sheran Dass, Kevin Gary, and James Cunningham. Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model. *Information*, 12(11):476, November 2021. ISSN 2078-2489. doi: 10.3390/info12110476. URL https://www.mdpi.com/2078-2489/12/11/476.

S. Nagrecha, J.Z. Dillon, and N.V. Chawla. MOOC dropout prediction: Lessons learned from making pipelines interpretable. In *Int. World Wide Web Conf. , WWW Companion*, pages 351–359. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349147 (ISBN). doi: 10.1145/3041021.3054162. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060302143&doi=10.1145%2f3041021.3054162&partnerID=40&md5=2a25f6a2524afa3b17197da4c7444a08. Journal Abbreviation: Int. World Wide Web Conf. , WWW Companion.

Galina Deeva, Johannes De Smedt, and Jochen De Weerdt. Educational Sequence Mining for Dropout Prediction in MOOCs: Model Building, Evaluation, and Benchmarking.

IEEE Transactions on Learning Technologies, 15(6):720–735, December 2022. ISSN 1939-1382, 2372-0050. doi: 10.1109/TLT.2022.3215598. URL https://ieeexplore.ieee.org/document/9925094/.

Mehmet Şahin. A Comparative Analysis of Dropout Prediction in Massive Open Online Courses. *Arabian Journal for Science and Engineering*, 46(2):1845–1861, February 2021. ISSN 2193-567X, 2191-4281. doi: 10.1007/s13369-020-05127-9. URL http://link.springer.com/10.1007/s13369-020-05127-9.

Theodor Panagiotakopoulos, Sotiris Kotsiantis, Georgios Kostopoulos, Omiros Iatrellis, and Achilles Kameas. Early Dropout Prediction in MOOCs through Supervised Learning and Hyperparameter Optimization. *Electronics*, 10(14):1701, July 2021. ISSN 2079-9292. doi: 10.3390/electronics10141701. URL https://www.mdpi.com/2079-9292/10/14/1701.

Fatemeh Khoushehgir and Sadegh Sulaimany. Negative link prediction to reduce dropout in Massive Open Online Courses. *Education and Information Technologies*, January 2023. ISSN 1360-2357, 1573-7608. doi: 10.1007/s10639-023-11597-9. URL https://link.springer.com/10.1007/s10639-023-11597-9.

Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. Understanding Dropouts in MOOCs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):517–524, July 2019. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.3301517. URL https://ojs.aaai.org/index.php/AAAI/article/view/3825. Number: 01.

W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, and Z. Wu. Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. In *Proc Int Jt Conf Neural Networks*, volume 2016-October, pages 3130–3137. Institute of Electrical and Electronics Engineers Inc., 2016. ISBN 9781509006199 (ISBN). doi: 10.1109/IJCNN.2016.7727598. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85007203277&doi=10.1109%2fIJCNN.2016.7727598&partnerID=40&md5=22a63d21f776b3bcffdad606e35443d6. Journal Abbreviation: Proc Int Jt Conf Neural Networks.

Lin Qiu, Yanshen Liu, Quan Hu, and Yi Liu. Student dropout prediction in massive open online courses by convolutional neural networks. *Soft Computing*, 23(20):10287–10301, October 2019. ISSN 1433-7479. doi: 10.1007/s00500-018-3581-3. URL https://doi.org/10.1007/s00500-018-3581-3.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

David Faraggi and Benjamin Reiser. Estimation of the area under the roc curve. *Statistics in medicine*, 21(20):3093–3106, 2002.

Ricvan Dana Nindrea, Teguh Aryandono, Lutfan Lazuardi, and Iwan Dwiprahasto. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. *Asian Pacific Journal of Cancer Prevention*, 19(7):1747–1752, July 2018. ISSN 1513-7368. doi: 10.22034/APJCP.2018.19.7.1747. URL http://journal.waocp.org/article_65369.html. Publisher: West Asia Organization for Cancer Prevention (WAOCP), APOCP's West Asia Chapter.

Anna Luíza Damaceno Araújo, Matheus Cardoso Moraes, Maria Eduarda Pérez-de Oliveira, Viviane Mariano da Silva, Cristina Saldivia-Siracusa, Caique Mariano Pedroso, Marcio Ajudarte Lopes, Pablo Agustin Vargas, Sara Kochanny, Alexander Pearson, Syed Ali Khurram, Luiz Paulo Kowalski, Cesar Augusto Migliorati, and Alan Roger Santos-Silva. Machine learning for the prediction of toxicities from head and neck cancer treatment: A systematic review with meta-analysis. *Oral Oncology*, 140:106386, May 2023. ISSN 1368-8375. doi: 10.1016/j.oraloncology.2023.106386. URL https://www.sciencedirect.com/science/article/pii/S1368837523000817.

Arman Ahmadi, Mohammadali Olyaei, Zahra Heydari, Mohammad Emami, Amin Zeynolabedin, Arash Ghomlaghi, Andre Daccache, Graham E. Fogg, and Mojtaba Sadegh. Groundwater Level Modeling with Machine Learning: A Systematic Review and Meta-Analysis. *Water*, 14(6):949, January 2022. ISSN 2073-4441. doi: 10.3390/w14060949. URL https://www.mdpi.com/2073-4441/14/6/949. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021b.

FENG Jiang and WENTAO Li. Who Will Be the Next to Drop Out? Anticipating Dropouts in MOOCs with Multi-View Features. *International Journal of Performability Engineering*, 13(2):201, March 2017. ISSN 0973-1318. doi: 10.23940/ijpe.17.2.p201.mag. URL http://www.ijpe-online.com/EN/10.23940/ijpe.17.2.p201.mag.

Lin Qiu, Yanshen Liu, and Yi Liu. An Integrated Framework With Feature Selection for Dropout Prediction in Massive Open Online Courses. *IEEE Access*, 6:71474–71484, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2881275. URL https://ieeexplore.ieee.org/document/8534361/.

Youssef Mourdi, Mohamed Sadgal, Hamada El Kabtane, and Wafaa Berrada Fathi. A machine learning-based methodology to predict learners' dropout, success or failure in MOOCs. *International Journal of Web Information Systems*, 15(5):489–509, December 2019. ISSN

1744-0084, 1744-0084. doi: 10.1108/IJWIS-11-2018-0080. URL https://www.emerald.com/insight/content/doi/10.1108/IJWIS-11-2018-0080/full/html.

Yimin Wen, Ye Tian, Boxi Wen, Qing Zhou, Guoyong Cai, and Shaozhong Liu. Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs. *Tsinghua Science and Technology*, 25(3):336–347, June 2020. ISSN 1007-0214. doi: 10.26599/TST.2019.9010013. URL https://ieeexplore.ieee.org/document/8858088/.

Qian Fu, Zhanghao Gao, Junyi Zhou, and Yafeng Zheng. CLSA: A novel deep learning model for MOOC dropout prediction. *Computers & Electrical Engineering*, 94:107315, September 2021. ISSN 0045-7906. doi: 10.1016/j.compeleceng.2021.107315. URL https://www.sciencedirect.com/science/article/pii/S0045790621002901.

Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir, and Sana Ullah Khan. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access*, 9:7519–7539, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3049446. URL https://ieeexplore.ieee.org/document/9314000/.

Gaurav Kumar, Amar Singh, and Ashok Sharma. Ensemble Deep Learning Network Model for Dropout Prediction in MOOCs. *International journal of electrical and computer engineering systems*, 14(2):187–196, February 2023. ISSN 18477003, 18476996. doi: 10.32985/ijeces.14.2.8. URL https://ijeces.ferit.hr/index.php/ijeces/article/view/1686.

Robert R Corbeil and Shayle R Searle. Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.

Guido Schwarzer, Hiam Chemaitelly, Laith J. Abu-Raddad, and Gerta Rücker. Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, 10(3):476–483, September 2019. ISSN 1759-2887. doi: 10.1002/jrsm.1348.

Theo Stijnen, Taye H. Hamza, and Pinar Özdemir. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*, 29(29):3046–3067, 2010. ISSN 1097-0258. doi: 10.1002/sim.4040. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4040. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4040.

Lifeng Lin and Chang Xu. Arcsine-based transformations for meta-analysis of proportions: Pros, cons, and alternatives. *Health Science Reports*, 3(3):e178, 2020.

Shelley R. Salpeter, Ji Cheng, Lehana Thabane, Nicholas S. Buckley, and Edwin E. Salpeter. Bayesian Meta-analysis of Hormone Therapy and Mortality in Younger Postmenopausal

Women. *The American Journal of Medicine*, 122(11):1016–1022.e1, November 2009. ISSN 0002-9343. doi: 10.1016/j.amjmed.2009.05.021. URL https://www.sciencedirect.com/science/article/pii/S0002934309006664.

Dan Jackson, Martin Law, Theo Stijnen, Wolfgang Viechtbauer, and Ian R. White. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine*, 37(7):1059–1085, 2018. ISSN 1097-0258. doi: 10.1002/sim.7588. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7588. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7588.

Miguel-Angel Negrín-Hernández, María Martel-Escobar, and Francisco-José Vázquez-Polo. Bayesian Meta-Analysis for Binary Data and Prior Distribution on Models. *International Journal of Environmental Research and Public Health*, 18(2):809, January 2021. ISSN 1661-7827. doi: 10.3390/ijerph18020809. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7832911/.

Fahad M. Al Amer, Christopher G. Thompson, and Lifeng Lin. Bayesian Methods for Meta-Analyses of Binary Outcomes: Implementations, Examples, and Impact of Priors. *International Journal of Environmental Research and Public Health*, 18(7):3492, March 2021. ISSN 1661-7827. doi: 10.3390/ijerph18073492. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8036799/.

Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. How to perform a meta-analysis with r: a practical tutorial. *BMJ Ment Health*, 2019.

Stan Development Team. RStan: the R interface to Stan, 2023. URL https://mc-stan.org/. R package version 2.32.3.

Timothy Hugh Barker, Celina Borges Migliavaca, Cinara Stein, Verônica Colpani, Maicon Falavigna, Edoardo Aromataris, and Zachary Munn. Conducting proportional meta-analysis in different types of systematic reviews: a guide for synthesisers of evidence. *BMC Medical Research Methodology*, 21:1–9, 2021.

Kyung Won Kim, Juneyoung Lee, Sang Hyun Choi, Jimi Huh, and Seong Ho Park. Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers-Part I. General Guidance and Tips. *Korean Journal of Radiology*, 16(6):1175–1187, 2015. ISSN 2005-8330. doi: 10.3348/kjr.2015.16.6.1175.

Kiran Fahd, Sitalakshmi Venkatraman, Shah J. Miah, and Khandakar Ahmed. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, 27(3):3743–3775, April 2022. ISSN 1573-7608. doi: 10.1007/s10639-021-10741-7. URL https://doi.org/10.1007/s10639-021-10741-7.

Apéndice A

Código de los modelos utilizados en RStan

A continuación se incluye el código de los dos modelos utilizados en `RStan`.

A.1. Modelo Normal-Normal

```
data {
 int <lower = 1> k; // Numero de estudios
 real <lower = 0, upper = 1>  x[k]; // Medida de efecto observada
 real <lower=0> sigma[k]; // Varianzas observadas
}

parameters {
 real <lower = 0, upper = 1> theta_i[k]; // Medida de efecto real
 real <lower = 0, upper = 1> theta; // media del parametro agrupado
 real <lower = 0> tau; // Heterogeneidad
}

model {
 x ~ normal(theta_i, sigma);
 theta_i ~ normal(theta, tau);

 // Priors
 theta ~ uniform(0, 1);
 tau^2 ~ inv_gamma(0.1, 0.1);
}

generated quantities{
   real <lower = 0> tausqr = tau^2;
```

```
}
```

A.2. Modelo Binomial-Normal

```

data {
 int <lower = 1> k; // Numero de estudios
 int <lower = 0>  x[k]; // Eventos observados al ser binomial debe
   ser de tipo entero
 int <lower=0> n[k]; // Casos
}

parameters {
 real<lower=0, upper=1> theta_i[k]; // Medida de efecto real
 real logtheta;
 real <lower = 0> tau; // Heterogeneidad
}

transformed parameters {
   real<lower=0, upper=1> theta = inv_logit(logtheta);
}

model {
 x ~ binomial(n, theta_i);
 logit(theta_i) ~ normal(logtheta, tau);

 // Priors
 theta ~ uniform(0, 1);
 tau^2 ~ inv_gamma(0.1, 0.1);
}

generated quantities{
   real <lower = 0> tausqr = tau^2;
}
```