Asignación de puntajes en exámenes estandarizados mediante el uso de redes neuronales y técnicas de equiparación psicométricas compatibles: Caso examen Saber 11 en Colombia

Ricardo René Duplat Durán

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2024

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Asignación de puntajes en exámenes estandarizados mediante el uso de redes neuronales y técnicas de equiparación psicométricas compatibles: Caso examen Saber 11 en Colombia

Ricardo René Duplat Durán

Trabajo Final presentado como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:
Luis Fernando Niño Vásquez, Ph.D.

Línea de Investigación:
Sistemas inteligentes
Grupo de Investigación:
Laboratorio de Investigación en Sistemas Inteligentes – LISI

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2024

Dedicatoria

A Paola, mi pareja y compañera de vida,

sin su apoyo, no habría logrado este proyecto.

A mi familia, por su apoyo y enseñanzas, por su formación y ejemplo.

A Marinita, quien despierta a la vida y a la que prometo con incondicionalidad acompañar sus pasos.

Declaración de obra original

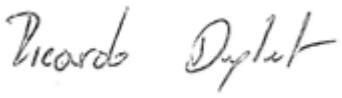
Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.



_____

Ricardo René Duplat Durán

Fecha: 28/01/2024

Agradecimientos

Al profesor Luis Fernando Niño, por su apoyo y compromiso en el desarrollo de este trabajo. De él, aprendí diferentes perspectivas y técnicas innovadoras en cursos que resultaron novedosas para mí, considerando mi formación como estadístico.

A los integrantes del grupo LISI, quienes a lo largo de esta maestría compartieron sus opiniones y consejos, enriqueciendo este trabajo.

Al Icfes, instituto que me brindó apoyo al permitirme utilizar los datos en los que se basa este trabajo, así como por ofrecer horarios flexibles que facilitaron mi asistencia a clases y la conclusión de todas las responsabilidades de la maestría.

A Paola, mi compañera de vida, cuyo apoyo me permitió sobrellevar las jornadas y días demandantes y extensos. Su motivación fue fundamental para superar los momentos difíciles.

A mis padres, quienes con su ejemplo me enseñaron sobre responsabilidad, compromiso y otros valores que han definido quién soy en mi vida adulta.

Resumen

Asignación de puntajes en exámenes estandarizados mediante el uso de redes neuronales y técnicas de equiparación psicométricas compatibles: Caso examen Saber 11 en Colombia

Los exámenes estandarizados son valiosas herramientas para evaluar de manera objetiva tanto las características cognitivas como no cognitivas de una población específica. Para construir escalas de medición que reflejen con precisión los constructos que estos exámenes buscan evaluar, se recurre comúnmente a la Teoría de Respuesta al Ítem (TRI), una técnica estadística. Sin embargo, la TRI presenta limitaciones cuando sus supuestos no se cumplen, comprometiendo la comparabilidad a lo largo del tiempo y entre subpoblaciones.

Este trabajo de grado se propone desarrollar una metodología innovadora que utiliza Redes Neuronales Artificiales (RNA), específicamente a través de AutoEncoders (AE), para preservar las ventajas de la TRI y aplicarla incluso cuando sus supuestos no se cumplen, buscando incluso mejorar la calidad de ajuste y pronóstico. La investigación se basa en el análisis del examen Saber 11 aplicado en los años 2018 y 2019, durante los calendarios A y B en el país.

Se obtuvieron resultados que en algunos casos superan el rendimiento de un modelo clásico de la TRI, como el modelo logístico de 2 parámetros (2PL). Esta metodología propuesta no solo busca subsanar las limitaciones de la TRI en ciertos contextos, sino que también busca optimizar la precisión en la asignación de puntajes en exámenes estandarizados mediante técnicas de equiparación compatibles con la psicometría. La aplicación de RNA, en particular a través de AE, emerge como una prometedora alternativa que contribuye al avance de la evaluación estandarizada, ofreciendo mayor flexibilidad y robustez en la medición de constructos educativos.

Palabras clave: Calificación de exámenes estandarizados, Teoría de Respuesta al Ítem (TRI), Redes Neuronales Artificiales (RNA), AutoEncoders (AE), Psicometría, Modelo logístico de 2 parámetros (2PL), Equiparación de puntajes.

Abstract

Scoring Assignment in Standardized Exams through the Use of Neural Networks and Psychometric Equating Techniques. Case Study: Saber 11 Exam in Colombia

Standardized exams are valuable tools for objectively assessing both cognitive and non-cognitive characteristics of a specific population. To construct measurement scales that accurately reflect the constructs these exams aim to evaluate, the Item Response Theory (IRT), a statistical technique, is commonly employed. However, IRT has limitations when its assumptions are not met, compromising comparability over time and among subpopulations.

This thesis aims to develop an innovative methodology using Artificial Neural Networks (ANNs), specifically through AutoEncoders (AE), to preserve the advantages of IRT and apply it even when its assumptions are not met, seeking to enhance the quality of fit and forecasting.

The research is based on the analysis of the Saber 11 exam administered in 2018 and 2019, during schedules A and B in the country. Results were obtained that, in some cases, outperform the performance of a classical IRT model, such as the 2-parameter logistic model (2PL). This proposed methodology not only aims to address the limitations of IRT in certain contexts but also seeks to optimize accuracy in score assignment in standardized exams through equating techniques compatible with psychometrics. The application of ANN, particularly through AE, emerges as a promising alternative contributing to the advancement of standardized assessment, offering greater flexibility and robustness in measuring educational constructs.

Key words*:* Standardized exam scoring, Item Response Theory (IRT), Artificial Neural Networks (ANNs), AutoEncoders (AE), Psychometrics, 2-parameter logistic model (2PL), Score equating.

Este Trabajo Final de maestría fue calificado en marzo de 2024 por el siguiente evaluador:

Álvaro Mauricio Montenegro Díaz, Ph.D.
Profesor Facultad de Ciencias
Universidad Nacional de Colombia

Contenido

Lista de gráficas

Lista de Figuras

Lista de tablas

Lista de abreviaturas

Abreviatura	Término
AE	AutoEncoder
VAE	AutoEncoder Variacional
AC	Examen Saber 11
TRI	Teoría de respuesta al ítem
SL	Stoking- Lord
2PL	Modelo logístico de dos parámetros
3PL	Modelo logístico de tres parámetros
ML	Machine Learning
IA	Inteligencia artificial
Icfes	Instituto Colombiano para la evaluación de la educación
RNA	Red neuronal Artificial
MA	Matemáticas
CN	Ciencias Naturales
LC	Lectura Crítica
SC	Sociales y Ciudadanas
PRC	Porcentaje

Introducción

Para medir la calidad de la educación en una población objetivo es útil el uso de exámenes estandarizados, puesto que estos permiten el análisis de los resultados de forma directa y una de sus características es que disminuyen la subjetividad de las valoraciones, de tal forma que la habilidad de los evaluados es el único factor que influye en su puntaje. No obstante, existen limitaciones derivadas de los métodos actualmente utilizados para la generación de dichas calificaciones, como son: los tiempos de procesamiento, los supuestos que se usan en el ajuste de los modelos estadísticos usuales para estimar dichas calificaciones (los cuales en ocasiones son difíciles de obtener) y los recursos económicos invertidos para entregar resultados de alta calidad, lo cual es deseable para poder generar información útil y veraz para el desarrollo de políticas públicas, entre otras iniciativas, que mejoren la calidad de la educación a diferentes niveles y que permitan tomar medidas correctivas precisas y oportunas.

Con el fin de disminuir estas limitaciones y teniendo en cuenta el auge y efectividad del uso de redes neuronales en procesos de medición y predicción, la idea de incursionar en modelos que hagan uso de ellas, se hace cada día más necesario y útil, pues, una vez ajustados los parámetros de las redes, se posibilita la generación de un modelo con altos niveles de predicción, en menor tiempo y con menor cantidad de supuestos a verificar.

Incursionar en el desarrollo de nuevos modelos implica también exponer los empleados en la actualidad en Colombia y en el mundo en general. Siendo así, el Instituto Colombiano para la Evaluación de la Educación (ICFES) emplea modelos psicométricos para la medición de las habilidades y de los factores asociados al aprendizaje de los estudiantes de Colombia (Icfes, 2021). Por otro lado, pruebas internacionales como PISA o TALIS también usan modelos psicométricos para estimar la habilidad de los evaluados, de tal forma que se garantiza la comparabilidad entre las diferentes economías, es decir las localizaciones geográficas donde se aplican dichos exámenes, las cuales pueden ser ciudades o países (PISA, 2019).

El presente documento tiene como objetivo presentar el trabajo de grado desarrollado para optar al título de Magíster en ingeniería de sistemas y computación en modalidad de

profundización. Desde esta perspectiva se plantea un orden lógico de los siguientes capítulos del documento; 1. Estado del arte, donde se realiza un abordaje teórico de los temas a desarrollar, 2. Marco de Investigación, que incluye planteamiento del problema, Justificación y Objetivo general y específicos, 3. Metodología, en el que se esclarece el tipo de estudio y la ruta metodológica para el desarrollo del presente trabajo, 4. Resultados, 5. Discusión y análisis de resultados y finalmente 6. Conclusiones y Trabajos futuros.

1. Estado del arte

El presente estado del arte abarca los antecedentes en la asignación de puntajes de exámenes estandarizados, así como los casos en que las RNA han sido utilizadas en la solución de problemas relacionados con este tema. También se ofrece una descripción sintética de varios conceptos tratados a lo largo del documento, para así poder entregar un panorama a lectores no familiarizados con el área de la psicometría o de las RNA. A su vez, se ofrece un contexto general del examen Saber 11, del uso de ML en la evaluación de exámenes estandarizados y cómo ha evolucionado este último en la psicometría. Por último, se presentan estudios en los que ya se ha propuesto el uso de los AE en la asignación de puntajes como punto de referencia para este trabajo.

1.1 Antecedentes

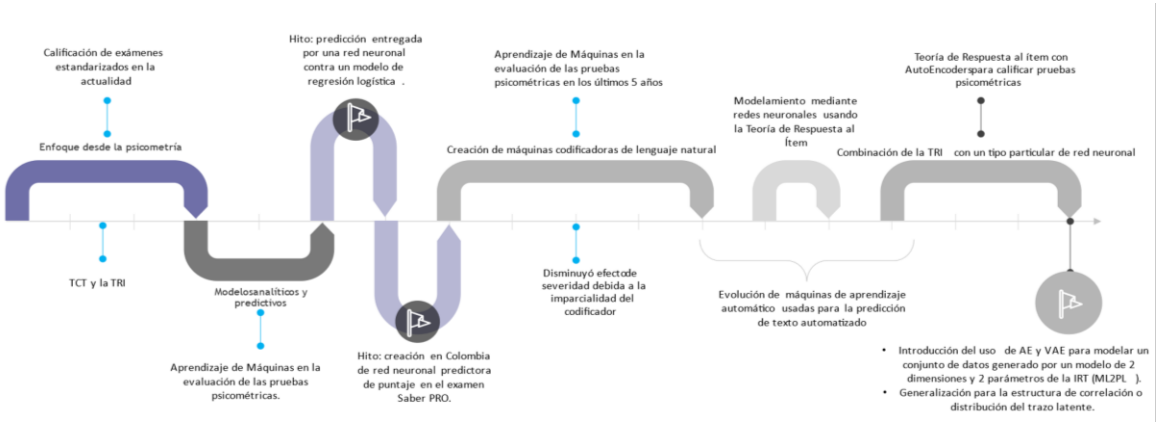
Para desplegar los antecedentes del trabajo se deben abordar dos conceptos centrales: psicometría y aprendizaje de máquinas. Aunque estos conceptos son áreas del conocimiento bastante amplias, es necesario centrarse en algunos conceptos fundamentales. La *Figura 1* constituye una línea de tiempo del desarrollo del estado del arte, indicando algunos hitos identificados dentro de la búsqueda documental.

1.2 Construcción de antecedentes: búsqueda bibliográfica

Para la construcción de antecedentes se debe tener claridad acerca de la necesidad y pertinencia de la información que se recaba, de este modo, la búsqueda documental fue guiada por el tema amplio y los objetivos planteados para este trabajo. Se establecieron parámetros para tener en cuenta dentro de la búsqueda, como periodos temporales, tipología documental ligada al aporte que representa para el desarrollo del contenido y, en particular, se realizó una selección de términos y ecuaciones de búsqueda que arrojaran los contenidos más aportantes.

Figura 1

Línea de tiempo del desarrollo del estado del arte. Elaboración propia, basada en búsqueda de antecedentes.



En ese sentido, es importante destacar el uso de cinco preguntas orientadoras que se presentan posteriormente, y que determinaron la ecuación de búsqueda usando las palabras claves que se describen a continuación: 1) se resolvió usando la ecuación "psychometry" and "Item Response Theory", 2) tuvo como ecuación de búsqueda "psychometry" and "Machine Learning", 3) consistía en la misma ecuación que 2), pero restringiendo la búsqueda a documentos publicados desde 2018, 4) fue resuelta mediante "Neural Network" AND "Item Response Theory" y 5) agregando a la ecuación anterior la palabra clave "AutoEncoders".

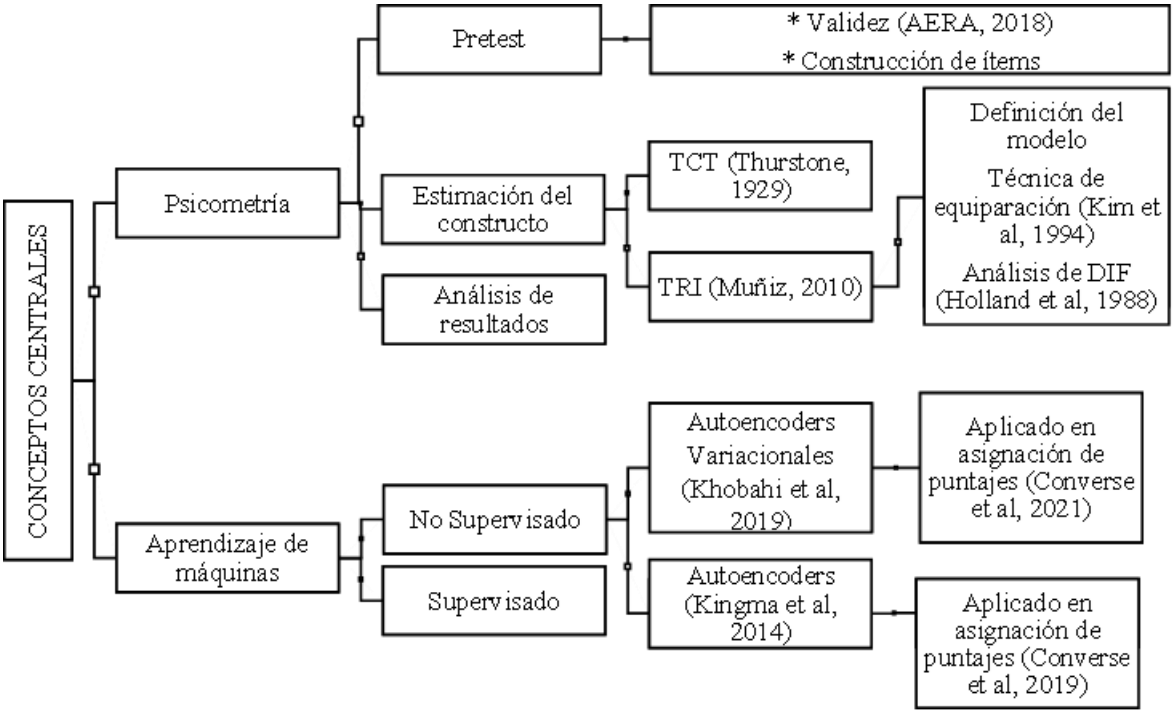
Cada ecuación de búsqueda produjo un conjunto de artículos y libros, de los cuales se seleccionaron aquellos que tenían mayor número de citaciones y los que, a su vez, contenían en el título *abstract* o párrafos introductorios con algunas de las palabras clave, anteriormente citadas. En esta primera selección de artículos se procedió a leer los resúmenes (*abstracts*), para así seleccionar los documentos en los que se podría ahondar por su pertinencia y aporte. De este conjunto final, se destacan los 3 artículos que se presentan en la última pregunta orientadora, con los cuales se puede ver la evolución de la calificación mediante redes neuronales de los exámenes estandarizados.

1.3 Descripción sintética de los conceptos

A continuación, se definen una serie de conceptos claves para ampliar, clarificar y entender los temas tratados en el presente documento. Cada una de estas definiciones pertenecen principalmente a dos grandes áreas: la psicometría y el aprendizaje de máquinas, lo que permite esclarecer el enfoque teórico orientador. La *Figura 2* muestra cómo se agrupan los conceptos centrales que se describirán posteriormente, los cuales sirven como contexto general de la temática a tratar.

Figura 2

Mapa de los conceptos centrales del estado del arte. Elaboración propia.



Psicometría: Es un conjunto de métodos, técnicas y teorías implicadas en la medición de variables que no pueden ser observadas directamente, por ejemplo, la inteligencia (Martínez et al, 2014).

Constructo: Es el concepto que pretende medir el instrumento, el cual debe estar desarrollado bajo un marco teórico que lo sustente (Martínez et al, 2014).

Tipos de prueba: i) Ejecución máxima: evalúan el mayor rendimiento que puede alcanzar el evaluado con respecto al constructo a evaluar. Por ejemplo, pruebas de inteligencia, conocimientos, competencias etc. (Muñiz, 2010). ii) Ejecución típica: evalúan características ordinarias del evaluado, más no el máximo de su potencial como sucede con las anteriores. Por ejemplo, pruebas de actitudes, opiniones, etc. (Muñiz, 2010).

Ítem: Se define como ítem a cada una de las preguntas que son aplicadas en las diferentes pruebas estandarizadas. De forma general, se puede entender que un ítem es la mínima parte de una prueba que recibe puntaje. La construcción deficiente de los mismos incidirá en las propiedades métricas finales del instrumento de medida y en la validez de las inferencias que se hagan a partir de las puntuaciones. En particular, los ítems dicotómicos son aquellos que tienen dos o más opciones de respuesta, pero solo una de ellas se puntúa. Ejemplos de estos ítems son los que se responden con "sí" o "no" o los de selección múltiple con única respuesta.

Validez: La validez se entiende como el grado en que un instrumento cuenta con un soporte teórico y suficientes evidencias que respaldan las conclusiones dadas a partir de los puntajes de los evaluados en una prueba (AERA, 2018).

Confiabilidad: Corresponde al grado con que los puntajes de una medición se encuentran libres de error aleatorio de medida. Las posibles fuentes de error que influyen en la confiabilidad son el tiempo de aplicación, la calidad del instrumento, la calidad de los ítems y la parcialidad de los evaluadores (Virla, 2010).

Teoría Clásica del Test: Esta teoría se basa en que la puntuación empírica de un evaluado está compuesta por su puntuación verdadera y un error de medida. Fue propuesta inicialmente por los trabajos de Spearman (1904, 1907) (Mutch, 2005) y Thurstone (1926-1927) (Thurstone, 1929).

Teoría de respuesta al ítem (TRI): Los modelos de TRI se aplican con mayor frecuencia en pruebas con ítems que se puntúan como correctos o incorrectos, generalmente codificados como 1

y 0, respectivamente. Se asume que una respuesta correcta indica un nivel más alto de competencia que una respuesta incorrecta.

Los planteamientos de la TRI se fundamentan en el postulado de que la ejecución de una persona en una prueba puede predecirse y explicarse por un conjunto de factores personales llamados, en conjunto, “habilidad” (trazo latente) y en el hecho de que la relación entre la ejecución de la persona evaluada y la habilidad que la soporta puede describirse por una función monótona creciente, representada por una función logística (Muñiz, 2010).

Rasgo o trazo Latente: Un rasgo latente es una característica que no puede ser medida directamente. Por ejemplo, la inteligencia, competencia en matemáticas o neuroticismo (Cortada, 2005).

Correlación Biserial: El coeficiente de correlación biserial puntual es una medida estadística utilizada para medir la relación lineal entre una variable dicotómica y una con escala de intervalo o razón (Muñiz, 2010).

Aprendizaje de máquinas: Es una parte de la Inteligencia Artificial que busca construir métodos que "aprendan" de los datos para el mejoramiento del desempeño de los algoritmos en algún tipo de tarea (Mitchell, 1997).

Redes Neuronales Artificiales (RNA): Las RNA son herramientas de modelamiento computacional que se han aplicado en diferentes disciplinas para tratar problemas complejos del mundo real. Se pueden definir como estructuras densamente comprimidas de elementos (neuronas) interconectados entre sí (Basheer et al, 2001).

AutoEncoders (AE): Son un tipo de RNA no supervisados (Kramer, 1991) que generan una representación de los datos originales en un espacio latente, es decir, un trazo latente multidimensional (Kingma et al, 2014).

AutoEncoders Variacionales (VAE): Son un tipo de RNA que sigue una idea similar a los AE, sin embargo, estos en realidad pertenecen a la familia de métodos variacionales bayesianos (Stone

2013). La diferencia conceptual entre los VAE y los AE radica en que el espacio latente de los primeros sigue una función de distribución que es una mixtura de distribuciones diferentes, en lugar de tan solo un vector fijo (Khobahi et al, 2019).

1.4 Calificación de exámenes estandarizados

En la actualidad existen principalmente dos enfoques para entregar puntajes bajo la teoría psicométrica: la TCT y la TRI. La TCT tiene como ventajas que requiere tamaños de muestra pequeños, consta de un modelo matemático, es más simple y la estimación de sus parámetros es fácil. Por otro lado, las desventajas consisten en que los resultados no son invariantes ante la población en que se aplica el test, los resultados no son comparables ante diferentes versiones del examen o del tiempo de aplicación de este (Muñiz, 2018). La *tabla 1* muestra una breve comparación de las propiedades de cada modelo.

Tabla 1

Comparación entre TCT y TRI

Atributo	Teoría Clásica del Test	Teoría de Respuesta al Ítem
Tamaño de muestra (Sanin, 2017)	No influye	Grande
Comparabilidad temporal (Kim 2006)	No	Si
Análisis comportamiento diferencial (Kim, 1994)	No	Si
Tiempo computacional (Hambleton, 1993)	Mínimo	Alto
Modelos (Londregan, 2021)	Único	1PL, 2PL, 3PL, MRG, MFR…

En cuanto la TRI, su enfoque se basa en las propiedades de las tareas o preguntas más que las del test global. La gran contribución de la TRI se centra en la posibilidad de obtener mediciones que no dependen de los instrumentos, los evaluados o el tiempo de aplicación del test. Así mismo, la TRI está compuesta de modelos que dependen del tipo de dato a tratar: para respuesta dicotómica están los modelos de 1PL o de Rasch, los cuales consideran que la probabilidad de contestar un ítem depende de la habilidad de la persona y la dificultad del ítem (Rasch, 1960). El modelo 2PL es una extensión del anterior, añadiendo el parámetro de discriminación, el cual indica la velocidad en que cambia la probabilidad de contestar correctamente el ítem con respecto a cambios de la habilidad.

En Colombia, el ICFES utiliza este modelo para calificar los módulos específicos de los exámenes Saber TyT y PRO (Icfes, 2020). El modelo 3PL añade el parámetro de pseudo-azar, la cual es una cota de probabilidad mínima de contestar correctamente el ítem sin importar la habilidad del evaluado. Este modelo es utilizado en Colombia para calificar las pruebas genéricas del examen Saber 11 (Icfes, 2020). Thissen, D. describe al detalle cada uno de estos modelos (Thissen et al, 2001). Aunque existen generalizaciones de estos modelos para el caso multidimensional, es decir que el trazo latente que mide la prueba no es unidimensional, estos modelos resultan complejos de estimar (Hartig et al, 2009).

Además de la aplicación de los modelos, deben realizarse varios análisis extra para poder afirmar que los resultados de la calificación son válidos y confiables. Entre estos análisis se encuentran: 1) Análisis de comportamiento diferencial del ítem (DIF), como el procedimiento de Stoking-Lord, el cual se realiza para garantizar que la probabilidad de contestar el ítem no depende del tipo de población, es decir, uno de supuesto de la TRI, que al no cumplirse, implica estimaciones sesgadas para subpoblaciones (Holland et al, 1988). 2) Análisis Factoriales para comprobar la unidimensionalidad de la prueba (Bro et al, 2014). 3) Análisis de ítem, el cual consiste en evaluar el comportamiento psicométrico de cada ítem que compone la prueba mediante la información que otorga en la estimación de la habilidad (Samajima, 1994). 4) Análisis multigrupo, que consiste en comprobar si el supuesto de que la población sigue una única distribución de probabilidad, que si no se cumple, implica sesgo no medible en las subpoblaciones o incluso la no convergencia de los modelos (Bolt et al, 2004). 5) Depuración de la población para la estimación de parámetros, las cuales son un conjunto de reglas que descartan a algunos evaluados para eliminar ruido, como el análisis de copia (Jara et al, 2010). Muchos de estos métodos de detección de copia hacen uso de modelos nominales de la TRI, los cuales también tienen una variedad de supuestos estadísticos que también deben ser comprobados para la validez de las conclusiones que deriven de su uso (Ostini et al, 2005). 6) Equiparación, que es un procedimiento que garantiza la comparación a través de exámenes aplicados en diferentes momentos del tiempo o conformados por diferentes conjuntos de preguntas. Para ello existen diferentes técnicas, como lo son la técnica de Stoking-Lord (Kim et al, 1994) o la fijación de parámetros de ítem (Kim, 2006).

1.5 Examen saber 11

El examen Saber 11 tiene como objetivo evaluar a estudiantes de grado 11 de bachillerato. Simultáneamente, se ofrece la posibilidad de presentar el examen a personas que deseen repetirlo de forma individual o a aquellos que busquen realizar el proceso de validación del bachillerato. En este trabajo, se dispone únicamente de información de estudiantes que estuvieron presentes durante toda la aplicación del examen.

Este examen se realiza dos veces al año. La primera aplicación está dirigida a colegios con calendario B, cuyo año escolar inicia en septiembre y finaliza en junio. En cuanto a la segunda aplicación, esta se lleva a cabo en colegios y establecimientos educativos que inician en febrero y concluyen en noviembre (Ministerio de Educación Nacional, s.f.).

El examen consta de cinco pruebas: Matemáticas, Lectura Crítica, Sociales y Ciudadanas, Ciencias Naturales e Inglés. Cada prueba contiene un número diferente de ítems de medición, utilizados en el proceso de asignación de puntajes, e ítems pilotos que no se emplean en la asignación de puntajes, pero se aplican en modalidad de medición en subsiguientes aplicaciones. Además, la escala de estas pruebas se define a través de un puntaje de media 50 y desviación estándar de 10, con un puntaje mínimo de 0 y máximo de 100 (ICFES, Examen Saber 11, s.f.).

Los resultados históricos del examen revelan que en la primera aplicación se evalúan menos de 23 mil estudiantes, mientras que en la segunda aplicación este número asciende a alrededor de 540 mil. Asimismo, los resultados promedio por prueba son mayores en el calendario B que en el calendario A (ICFES, Resultados Agregados Saber 11, s.f.).

Por otro lado, el diseño de construcción de las pruebas del examen Saber 11 se basa en el Diseño en Bloques Incompletos (BIBs), el cual implica la creación de diversas versiones del examen o formas mediante la combinación de un número determinado de bloques de ítems. Estos bloques se generan a partir de un subconjunto de todos los ítems posibles para la aplicación, asegurando que la intersección entre cualquier par de bloques de ítems sea nula, y que la unión de los bloques contenga todos los ítems posibles para la aplicación. A modo de ejemplo, si una prueba de una aplicación consta de 100 ítems, se podrían construir 10 bloques, cada uno compuesto por 10 ítems.

Además, se espera que cada una de las formas del examen compartan por lo menos un bloque de ítems (Instrumentos, s.f.), sin embargo, esto no aplica para la prueba de inglés, la cual sigue un diseño diferente al de las demás pruebas. Por esta razón, y teniendo en cuenta que esta prueba es opcional para cierta población (estudiantes con discapacidad no motriz y algunas etnias específicas del país), esta prueba no se tendrá en cuenta en el presente trabajo.

1.6 Aprendizaje de máquinas en el análisis de resultados de pruebas psicométricas

El aprendizaje de máquinas para la evaluación de las habilidades de las personas lleva siendo usado desde hace relativamente bastante tiempo. Stevens (2006) plantea varios modelos analíticos para medir la forma en que estudiantes de educación media y universitaria crean estrategias para resolución de problemas de ciencias en páginas web. En primera medida hace uso de la TRI para estimar la habilidad de los evaluados y, por otro lado, crea un modelo de clasificación mediante redes neuronales tomando como entrada las acciones de los estudiantes para resolver cada problema (Stevens, 2006).

Por otro lado, una tarea fundamental a la hora de crear un test consiste en la escogencia de aquellos ítems de un banco disponible, los cuales deben tener ciertas características dependiendo de si es un test orientado a la medición general de la población o aprobatorio. Cuando el banco de ítems es demasiado grande, resulta difícil armar uno o varios tests con la mayor información posible entregada por los ítems. El-Alfy y Abdel-Aal (2008) desarrollaron una red neuronal que realiza esta tarea maximizando la información del test con un número mínimo de ítems que componen el test (El-Alfy et al, 2008).

Otro de los usos del aprendizaje de máquinas consiste en la detección de copia durante la aplicación del examen. Amir A. desarrolla técnicas de inteligencia artificial y aprendizaje de máquinas para reconocer rostros durante exámenes electrónicos para, de esta forma, identificar fraude cometido por algunos evaluados (Amir, 2022).

Además del uso de las redes neuronales para tareas de la evaluación estandarizada, también se han usado para modelar y predecir el puntaje de los estudiantes en pruebas internacionales y

nacionales. Por ejemplo, Bozak A. y Aybek E. compararon la predicción entregada por una red neuronal contra un modelo de regresión logística para predecir los puntajes de estudiantes turcos de 15 años en las pruebas PISA, encontrando que la red neuronal es significativamente mejor que el modelo estadístico (Bozak et al, 2020). A nivel nacional, García J. y Sánchez P. crearon una red neuronal con el fin de predecir el puntaje en el examen Saber PRO, el cual es dirigido a medir la habilidad de estudiantes universitarios en Razonamiento Cuantitativo, Inglés, Competencias Ciudadanas, Lectura Crítica y Comunicación escrita (García-González et al, 2019).

1.7 Evolución del Aprendizaje de Máquinas en la psicometría

Algunas pruebas psicométricas consisten en evaluar las competencias en cuanto a la escritura, para ello se suelen desarrollar pruebas en donde se ofrece un estímulo o tema y el evaluado debe elaborar un ensayo. Cuando este tipo de pruebas es aplicado a un número grande de evaluados, suele ser muy costoso contratar codificadores que asignen un puntaje a cada texto, con el fin de ajustar un modelo de TRI que logre medir la habilidad de la población. En este sentido, se han desarrollado máquinas que codifican automáticamente los textos mediante el procesamiento del lenguaje natural, quitando así un posible efecto de severidad debida a la imparcialidad del codificador y reducción de costos. La tabla 2 muestra cómo han evolucionado las máquinas de aprendizaje automático que se han usado en diferentes contextos para la predicción de texto automatizado.

Tabla 2

RNA para procesamiento de texto con memoria. Elaboración propia basada en bibliografía.

RNA	Año
LSTM (S Hochreiter, J Schmidhuber)	1997
GAN (GoogFellow et al)	2014
GPT1 (Radford et al)	2018
BERT (Sun et al)	2018
Trasnformers (Lin et al)	2018
GPT2 (Radford et al)	2019
GPT3 (Mann et al)	2021
GPT4 (OpenAI)	2023

Los tests estandarizados no solamente están diseñados para medir constructos cognitivos, también se utilizan para emociones o incluso enfermedades mentales. En este sentido, el aprendizaje de máquinas también ha sido utilizado para la medición de algunas condiciones como la demencia mediante tests psicométricos, los cuales no son necesariamente diligenciados por el evaluado sino por psicólogos. Dunn T. realizó un estudio transversal con patrones detectados por cuidadores en pacientes con demencia y deterioro cognitivo leve, aplicando técnicas de aprendizaje de máquinas (Dunn et al, 2022).

En esta misma línea, Vakadkar K. desarrolló una red neuronal aplicada a datos recolectados mediante una prueba psicométrica con el fin de detectar autismo en niños (Vakadkar et al, 2021). Como estos artículos, se pueden encontrar por lo menos 20 artículos más desde 2018 en Scopus que muestran cómo se puede usar el aprendizaje de máquinas en evaluaciones psicométricas cognitivas y no cognitivas.

1.8 Inclusión de la Teoría de Respuesta al Ítem en las RNA

Una técnica de entrenamiento, propuesta por Lalor et al, usa la TRI para acortar el tiempo en que una red neuronal profunda tarda en ajustarse para modelar datos de respuesta binaria. Básicamente, se entrena la red que usa un subconjunto especial de los datos con ciertas propiedades psicométricas óptimas, después se aplica esto a un conjunto de datos más general (Lalor et al, 2017). Similar a esta técnica, en el ICFES actualmente se está desarrollando una investigación en la cual se pretende usar los resultados de un modelo de Múltiples Facetas de Rasch (Linacre, 1994), con el fin de estimar la severidad de los codificadores de la prueba de Escritura de Saber 3579 (MinEducacion, 2022) y así entrenar una red neuronal que permita codificar automáticamente los textos. Este proceso es útil puesto que no todos los codificadores tienen el mismo nivel de severidad, por lo que el ajuste de un codificador automático no es sencillo.

En cuanto a las pruebas estandarizadas, Jung et al. realizaron una comparación entre una codificación automática en ítems de respuesta corta para la prueba TIMSS 2019 y las codificaciones realizadas por humanos llegando a una correlación de 0.91, además de proponer el uso de ciertas

técnicas de TRI como medidas de control de calidad a la codificación realizada por la RNA (Jung, 2022).

1.9 Uso de la Teoría de Respuesta al ítem con los AutoEncoders en la calificación de pruebas psicométricas

Anteriormente se expuso la forma en que se puede combinar la TRI con las redes neuronales en general. Ahora se presenta como se podría combinar la TRI con un tipo particular de red neuronal. En realidad, es un campo muy poco trabajado, puesto que realizar una búsqueda que combine la TRI con este tipo de red solo se encuentran 3 artículos, los cuales pertenecen a un mismo grupo de investigadores.

Los AutoEncoders (AE) consisten en un tipo específico de red neuronal que reducen la dimensionalidad de los datos de entrada en un número determinado de trazos latentes (codificación) y después se aumenta la dimensión tratando de recuperar los datos originales (decodificación) minimizando el error, el cual se mide mediante la diferencia entre el dato original y el estimado, así mismo, los AutoEncoders Variacionales (VAE) son una modificación de estos en que se permite que el trazo latente no sea un vector fijo, sino que está compuesto de una mezcla de distribuciones (Kingma et al, 2019).

En el estudio realizado por Curi et al se introdujo el uso de AE y VAE para modelar un conjunto de datos generado por un modelo de 2 dimensiones y 2 parámetros de la TRI (ML2PL) (Curi et al, 2019). Este fenómeno puede presentarse en algunos exámenes estandarizados, por ejemplo, cuando se busca evaluar el conocimiento matemático de un estudiante mediante preguntas que están formuladas utilizando un texto complejo. En este caso, el estudiante debe comprender el texto antes de poder plantearlo como un problema matemático, lo cual requiere habilidades tanto en lectura como en matemáticas. Estas dos habilidades son dimensiones interdependientes.

Cuando el examen estandarizado no está diseñado para medir un único trazo latente, sino muchos, que no necesariamente son muy correlacionados, los AE convencionales no tienen un poder predictivo alto. En este caso, es mejor usar un VAE para modelar los datos. Lo anterior es mostrado por Converse, en el que ajustan una de estas redes neuronales haciendo uso de un modelo

de TRI multidimensional, con la ventaja de que los parámetros de la red son interpretables, lo cual no es usual en el campo de las redes neuronales (Converse et al, 2019). Converse también presenta un VAE cuando los trazos latentes están fuertemente correlacionados en un trabajo posterior (Converse et al, 2021).

Una conclusión clara que se desprende de esta revisión de literatura es la posibilidad de utilizar redes neuronales artificiales (RNA) en la obtención de puntajes para exámenes estandarizados. Sin embargo, no se han encontrado artículos que describen los procedimientos para comparar resultados en la medición de un constructo mediante diferentes versiones de la prueba o en diferentes momentos del tiempo (equiparación), ni tampoco métodos para evaluar el comportamiento diferencial de los ítems.

2. Marco de investigación

En este capítulo, se aborda de manera integral el problema que conlleva actualmente la asignación de puntajes en exámenes estandarizados en el país. Además, se plantea la justificación del estudio, el objetivo general y objetivos específicos.

2.1 Planteamiento del problema

La aplicación de un examen estandarizado para evaluar pruebas cognitivas permite mejorar la confiabilidad y la validez de los resultados (Muñiz, J., 2018). Es importante que los resultados de estos exámenes sean comparables entre subpoblaciones y que no dependan del momento del tiempo en que sean aplicados (Phelps, 2011, Eignor, 2006 y Hambleton, 1991). Esto con el fin de poder realizar comparaciones para lograr identificar los aspectos a mejorar en las políticas públicas educativas del país.

En la actualidad, los resultados de los exámenes estandarizados son modelados a través de la TRI, la cual está compuesta de modelos estadísticos usados en la psicometría (ICFES, 2020). Estos modelos tienen varios inconvenientes que se expusieron en la primera pregunta orientadora del desarrollo del estado del arte: 1) Gran cantidad de supuestos estadísticos que en ocasiones no se cumplen (Doran et al, 1985 & Rios et al, 2021) y, 2) si la población sigue una distribución sesgada o existen subgrupos con distribuciones diferentes, los resultados presentaran sesgo no medible (Lord, 1983 & Warm, 1989).

Una posible alternativa a los modelos de TRI consiste en el uso de RNA que, una vez entrenadas, puedan reducir el tiempo computacional, así como eliminar el proceso de comprobación de supuestos teóricos. En cuanto al posible sesgo de la población, este puede ser controlado mediante el ajuste de un AutoEncoder variacional (AEV), el cual permite cualquier tipo de distribución para el trazo latente. Para evaluar el comportamiento del test en diferentes subpoblaciones, se puede considerar el uso de una mixtura distribucional del trazo latente del AEV (Converse, 2019 y 2021).

Teniendo en cuenta lo anterior, se plantea la siguiente pregunta de investigación: *¿Cómo puede implementarse el uso de las redes neuronales artificiales en la realización de análisis psicométricos para la obtención de calificaciones comparables entre las aplicaciones del examen saber 11 de dos años consecutivos (aplicaciones 2018 y 2019)?*

2.2 Justificación

La experiencia laboral en el ICFES ha dejado múltiples aprendizajes y el interés por indagar acerca de herramientas para optimizar los procesos y, del mismo modo, propiciar mejores resultados en las tareas y funciones que cumple el instituto. Siendo la institución del estado que se encarga de la evaluación de la educación en Colombia en todos sus niveles y, a su vez, de propiciar investigaciones sobre los factores que inciden en la calidad educativa, con el fin de mejorarla; optimizar los procesos que llevan a esto, es promover la garantía de mejores tomas de decisiones a nivel de políticas públicas que se basen en datos fidedignos y oportunos, la transparencia en el acceso a la educación, el reconocimiento, tanto individual como colectivo y, en definitiva, vislumbrar panoramas más equitativos de la educación en Colombia.

Ha sido esta la motivación para emprender la búsqueda de alternativas que, en términos prácticos, busquen la optimización en los procesos de calificación del examen Saber 11, mediante el uso de redes neuronales y técnicas de equiparación psicométricas compatibles, lo que representa una alternativa que no implica el uso de la TCT, la cual presenta varias limitaciones, como se ha expuesto previamente. Algunos de los problemas y ventajas que podrían resolverse o presentarse son:

a) Disminuyen el tiempo de procesamiento, puesto que una vez entrenada la red neuronal, el tiempo es muy inferior al que tardan los modelos de Teoría de Respuesta al Ítem (TRI), modelo actualmente usado, lo que permite la obtención de resultados oportunos.

b) En problemas prácticos, en ocasiones no es posible garantizar los supuestos de la TRI, lo que invalida el uso de esta. Una alternativa entonces puede ser el ajuste mediante redes neuronales.

c) Los AEV permiten mayor libertad al momento de suponer un comportamiento distribucional del trazo latente. Como explican Lord y Warm, cuando la distribución de la habilidad medida a través de un examen estandarizado no es simétrica, genera sesgo en los resultados, lo que implica que una parte de la población tendría una medición errónea, esto se traduce como una limitante importante para la TRI (Lord, 1983; Warm, 1989).

Por lo anterior, es necesario encontrar un método que permita implementar el uso de redes neuronales en la psicometría, de tal forma que se obtengan todos los beneficios que ofrece la TRI clásica: comparabilidad a través del tiempo, entre versiones de exámenes y entre subpoblaciones.

2.3 Objetivo general

Desarrollar una metodología para la asignación de puntajes del examen Saber 11 en Colombia mediante el uso de redes neuronales y técnicas de equiparación psicométricas compatibles.

2.3.1 Objetivos específicos

- Construir un esquema de red neuronal artificial que estime cada uno de los constructos medidos en el examen Saber 11.
- Identificar la técnica de equiparación compatible con los resultados de la red neuronal artificial.
- Implementar la metodología del esquema con las aplicaciones del examen Saber 11 en 2018 y 2019.
- Evaluar el metodología de calificación propuesto para las aplicaciones del examen Saber 11 estudiadas.

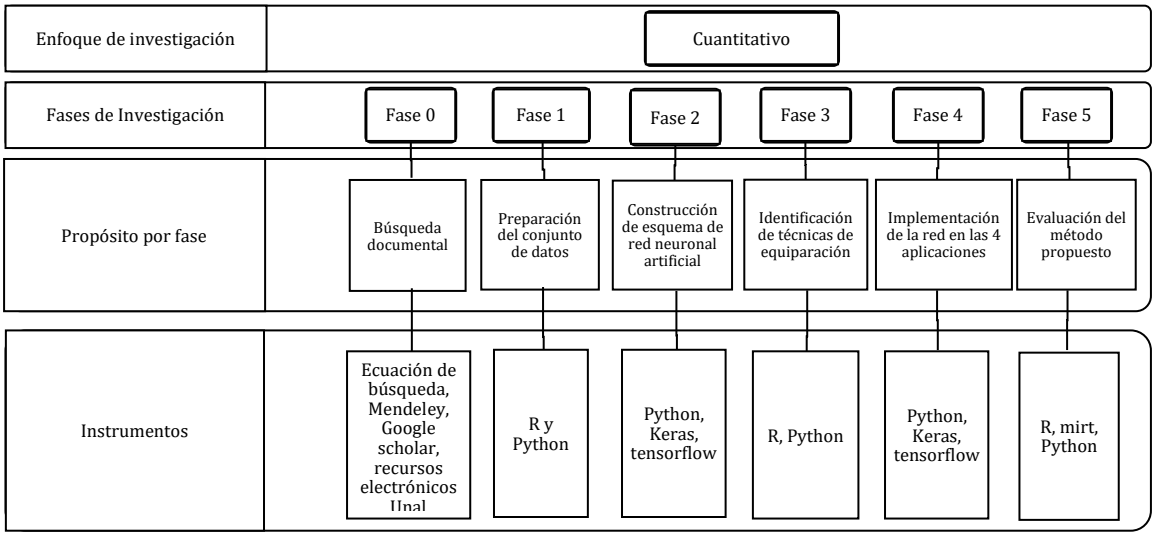
3. Metodología

Con respecto al tipo de investigación para desarrollar el trabajo de grado, el enfoque es de tipo cualitativo, exploratorio y propositivo, pues se busca proponer un método que utiliza herramientas previamente creadas como las RNA junto con la TRI de una forma novedosa.

La *figura* 3 ilustra las 5 fases que se identificaron para poder desarrollar la investigación: la fase 1, de tipo preparativa, en la cual se complementó la búsqueda documental, junto con la preparación del conjunto de datos a usar en cada una de las aplicaciones. En la fase 2, se definió la arquitectura del AutoEncoder. En la fase 3 se estableció la forma de aplicación de la técnica de equiparación aplicable de la TRI, que finalmente se determinó como la técnica de Stoking-Lord. En la fase 4 se aplicó lo definido en las dos fases anteriores, tanto en la aplicación inicial que se usó como línea base y a las siguientes aplicaciones en una misma escala. Por último, en la fase 5 se evaluó la técnica propuesta y se comparó con la metodología usual de la TRI para calificar el examen Saber 11, la cual fue mediante el ajuste de modelos 2PL.

Figura 3

Fases de la investigación. Elaboración propia.



3.1 Preparación y análisis exploratorio de los datos

Los pasos que se usaron para la preparación de los datos para cada prueba de cada aplicación fue el siguiente:

1. A partir del conjunto de datos original, se ajustó un modelo logístico de dos parámetros (2PL) mediante el paquete MIRT de R, con el objetivo de identificar y eliminar ítems cuya carga factorial fuera igual o inferior a 0. Este procedimiento excluye aquellos ítems que, debido a su construcción o a un error en la definición de la respuesta correcta, no están vinculados al constructo latente que se pretende medir o incluso que contribuyen de manera negativa.

Es importante destacar que este procedimiento representa un paso inicial en el análisis de ítems, siendo necesario realizar análisis adicionales, como análisis univariados por opción de respuesta, análisis de unidimensionalidad, análisis de comportamiento diferencial del ítem por subpoblaciones, análisis de estadísticas de TCT, entre otros, para la selección final de ítems en una prueba estandarizada, los cuales no se realizaron dado que no es el objetivo del documento.

2. Con el conjunto definitivo de ítems, se llevó a cabo una selección aleatoria de evaluados para constituir los conjuntos de entrenamiento, validación y prueba, asignando el 80%, 10% y 10% del total de evaluados, respectivamente.

3. Respecto al conjunto de entrenamiento, validación y prueba, se llevó a cabo la identificación y denominación de la forma que cada evaluado presentó. Este proceso permitió la construcción de una tabla específica para cada forma, incluyendo los ítems y los evaluados correspondientes, con el propósito de evitar la presencia de datos faltantes.

La *Tabla 3* presenta el número de ítems tanto en su versión original como en la versión final para cada prueba y aplicación, así como el número total de evaluados. A través de esta tabla, se observa que fue necesario eliminar solo un ítem para la prueba de Matemáticas en las aplicaciones de 2018-1 y 2018-2. En relación con las demás pruebas, no se registraron ítems eliminados en ninguna de las aplicaciones.

Además, se destaca que existe una variación significativa en el número de evaluados entre los diferentes calendarios. En particular, se observa que la primera aplicación de cada año,

correspondiente al calendario B, representa menos del 4% de los evaluados en comparación con la segunda aplicación, que corresponde al calendario A.

Tabla 3

Número de evaluados e ítems por prueba y aplicación.

APLICACIÓN	PRUEBA	EVALUADOS	ITEMS ORIGINAL	ITEMS FINAL
2018-1	CIENCIAS NATURALES	19065	105	105
2018-2	CIENCIAS NATURALES	549986	105	105
2019-1	CIENCIAS NATURALES	21167	105	105
2019-2	CIENCIAS NATURALES	533552	105	105
2018-1	LECTURA CRÍTICA	19225	73	73
2018-2	LECTURA CRÍTICA	552038	73	73
2019-1	LECTURA CRÍTICA	21254	73	73
2019-2	LECTURA CRÍTICA	535341	73	73
2018-1	MATEMÁTICAS	19165	89	88
2018-2	MATEMÁTICAS	552034	89	88
2019-1	MATEMÁTICAS	21262	89	89
2019-2	MATEMÁTICAS	535355	89	89
2018-1	SOCIALES Y CIUDADANAS	19184	89	89
2018-2	SOCIALES Y CIUDADANAS	551607	89	89
2019-1	SOCIALES Y CIUDADANAS	21240	89	89
2019-2	SOCIALES Y CIUDADANAS	534948	89	89

4. Una vez establecido el conjunto final de ítems para cada aplicación, se procedió al análisis de las diversas formas construidas para el examen. La *Tabla 4* exhibe el número de ítems compartidos por las distintas formas del examen de Matemáticas para el periodo 20181. En la diagonal principal se encuentran consignados los totales de ítems de cada una de las 14 formas, mientras que la intersección de la fila i-ésima y la columna j-ésima indica la cantidad de ítems compartidos entre la forma i y la forma j, con i, j = A,…,N.

El *Anexo 1.A* incluye todas las tablas que detallan por prueba y aplicación el número de ítems que comparten cada una de las formas. Al analizar estas tablas, se puede determinar que cada

forma está compuesta por 4 bloques de ítems, el cual varía en número según la prueba, y que cada pareja de formas puede compartir 0, 1, 2 o 3 bloques.

Tabla 4

Conteo de ítems comunes entre formas de la prueba de Matemáticas en las aplicaciones 2018-1.

FORMA	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
A	44	22	11	22	22	11	33	11	22	22	22	22	22	22
B	22	44	22	22	22	22	11	11	22	33	22	11	22	22
C	11	22	43	11	22	21	21	21	10	22	21	22	32	22
D	22	22	11	44	22	11	11	33	22	22	22	22	22	22
E	22	22	22	22	44	22	22	22	22	22	0	22	22	22
F	11	22	21	11	22	43	21	21	32	22	21	22	10	22
G	33	11	21	11	22	21	43	21	21	22	21	22	21	11
H	11	11	21	33	22	21	21	43	21	22	21	22	21	11
I	22	22	10	22	22	32	21	21	43	11	21	11	21	22
J	22	33	22	22	22	22	22	22	11	44	22	22	11	11
K	22	22	21	22	0	21	21	21	21	22	43	22	21	22
L	22	11	22	22	22	22	22	22	11	22	22	44	11	33
M	22	22	32	22	22	10	21	21	21	11	21	11	43	22
N	22	22	22	22	22	22	11	11	22	11	22	33	22	44

Además, es imperativo comparar el número de ítems compartidos entre dos aplicaciones consecutivas. Esto permitirá identificar el método de equiparación necesario para realizar comparaciones de puntajes cada vez que se aplique el examen, independientemente del calendario o del año de aplicación. La Tabla 5 presenta el número de ítems compartidos entre las formas de la prueba de Matemáticas en las aplicaciones 2018-1 y 2018-2.

En la *Tabla 5* y en aquellas contenidas en el *Anexo 1.B* , que abarcan todas las comparaciones entre aplicaciones consecutivas de todas las formas para todas las pruebas, se observa que el máximo número de ítems compartidos entre dos formas de aplicaciones consecutivas es aproximadamente tres cuartas partes de la longitud máxima de la prueba, lo cual es un hecho relevante al momento de definir la técnica de equiparación de la TRI compatible con la asignación de puntajes con AE.

Tabla 5.

Número de ítems compartidos entre las formas de la prueba de Matemáticas en las aplicaciones 2018-1 y 2018-2.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	24	13	5	23	15	16	25	3	13	12	4	16	12	15
B	12	11	4	22	15	16	15	12	1	11	16	5	23	26
C	4	13	17	1	6	7	14	5	14	11	16	7	12	6
D	25	14	3	34	24	13	22	15	14	23	12	13	23	24
E	5	5	8	13	17	7	3	18	5	13	16	7	13	17
F	1	12	13	12	13	2	11	13	11	22	22	2	21	13
G	13	12	15	12	15	5	24	3	23	22	14	16	11	4
H	14	13	13	23	24	2	21	15	24	33	22	13	22	13
I	12	23	12	23	12	12	21	13	11	22	21	1	32	23
J	12	0	4	22	26	5	15	12	12	22	16	16	12	15
K	22	21	11	22	12	12	33	0	21	21	11	12	21	12
L	15	5	6	13	15	4	12	6	15	13	2	15	1	4
M	15	24	16	12	5	17	24	5	14	11	15	6	23	16
N	15	16	6	13	4	15	12	6	4	2	2	4	12	15

3.2 Definición de la RNA y el método de equiparación dentro de la aplicación.

Considerando que cada prueba cuenta con múltiples formas en cada aplicación y que cada evaluado responde a una de estas formas, al combinar los evaluados de dos formas diferentes en una tabla única (con los evaluados en las filas y los ítems en las columnas), el resultado sería una tabla en la cual solo los ítems compartidos contendrían respuestas en todas las filas, mientras que las columnas restantes tendrían datos faltantes en aproximadamente la mitad de las filas. Teniendo en cuenta esta estructura, se describe el procedimiento definido para aplicar un AE con el fin de estimar la habilidad de los evaluados:

1. Aplicar un AE a cada una de las formas (A, B, …, M, N) de manera independiente.
2. Seleccionar una forma objetivo para determinar la escala de la aplicación, por ejemplo, la forma A.
3. Seleccionar parejas de formas de tal manera que la primera forma sea la forma objetivo y la segunda, cualquiera de las otras formas, con la condición de que compartan más de

una cuarta parte de los ítems de la primera forma. En el caso de las formas que no puedan emparejarse con la forma objetivo, seleccionar como primera forma aquella que comparta el mayor número de ítems con la forma objetivo y completar las parejas con el mismo criterio del número mínimo de ítems compartidos. La Tabla 6 muestra la selección de parejas para la prueba de Matemáticas en 20181.

4. Aplicar un AE a la combinación de cada pareja de formas únicamente con los ítems compartidos.
5. A través del Método de Equiparación de Stocking-Lord (SL), realizar un ajuste para transformar las habilidades de los evaluados con la Forma 2 (paso 1) a las habilidades estimadas mediante el AE de la pareja que la contiene (paso 3). Del mismo modo, utilizando nuevamente el SL, transformar las habilidades de los evaluados de la Forma 1, calculadas a partir del AE de la pareja (paso 3), a las habilidades del AE de la Forma 1 (paso 1). Es decir que, siendo la $Pareja = (F_1, F_2)$, las habilidades de los evaluados mediante la Forma 2 en escala de la forma objetivo, θ_2^*, estarían dadas por:

$$\theta_2^* = g(f(\theta_2, \theta_{12}), \theta_1),$$

donde θ_2 es la habilidad de los evaluados con la Forma 2 a partir del AE del paso 1, θ_{12} es la habilidad de los evaluados de la pareja de formas a partir del AE del paso 3, θ_1 es la habilidad de los evaluados de la Forma 1 con el AE del paso 1, $f(.)$ es la función que equipara las habilidades de la Forma 2 con el de la pareja y $g(.)$ es la función que equipara la habilidad de la pareja con la de la Forma 1. Cuando la pareja no contiene a la forma objetivo, la transformación está dada por

$$\theta_2^* = g(f(\theta_2, \theta_{12}), \theta_1^*),$$

donde θ_1^* representa las habilidades ya transformadas a la escala de la forma base.

La *Gráfica 1* muestra el histograma y la función de densidad de este proceso para la prueba de matemáticas 20181. En esta se puede observar que todas las escalas están en el mismo rango y no se aprecia alguna que resalte con respecto al comportamiento general de la población. El *Anexo 2* muestra los diagramas de densidad e histogramas para todas las pruebas en todas las aplicaciones.

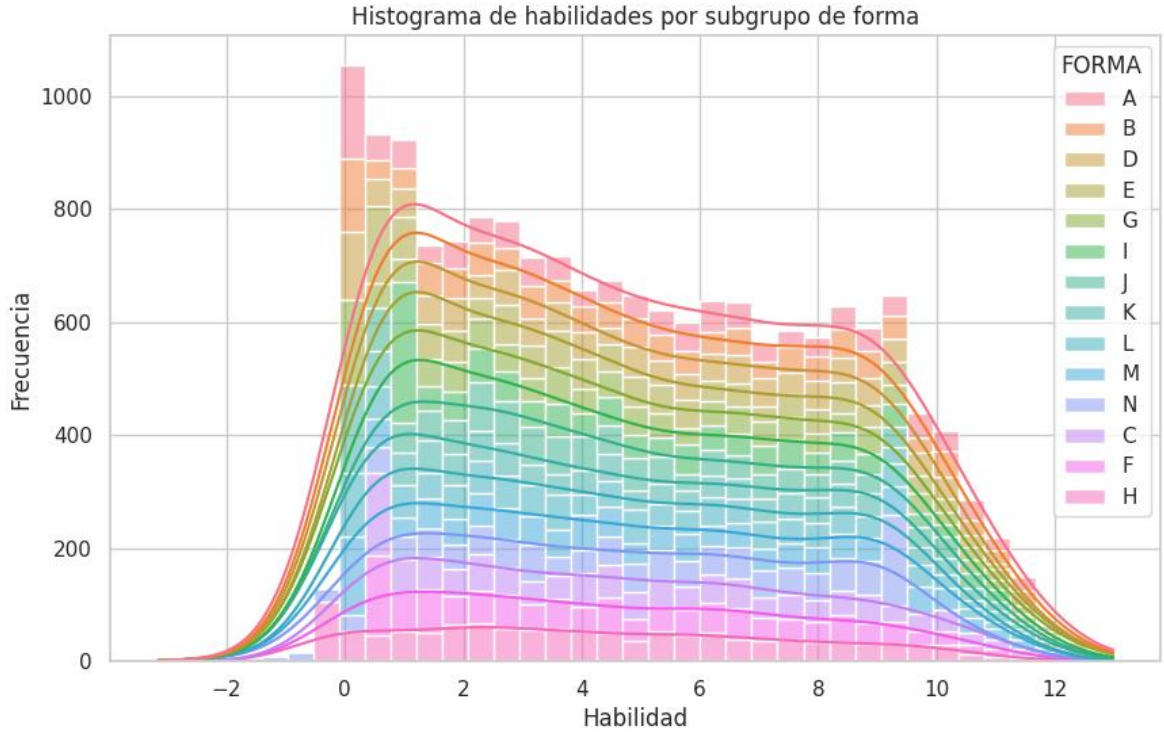
Tabla 6.

Selección de parejas para la prueba de Matemáticas en 20181.

ID	Forma 1	Forma 2
1	A	B
2	A	D
3	A	E
4	A	G
5	A	I
6	A	J
7	A	K
8	A	L
9	A	M
10	A	N
11	G	C
12	G	F
13	G	H

Gráfica 1.

Histograma y función de densidad para la prueba de matemáticas 20181.



Ahora bien, el AE definido para aplicar en los pasos 1 y 3 del algoritmo está dado por:

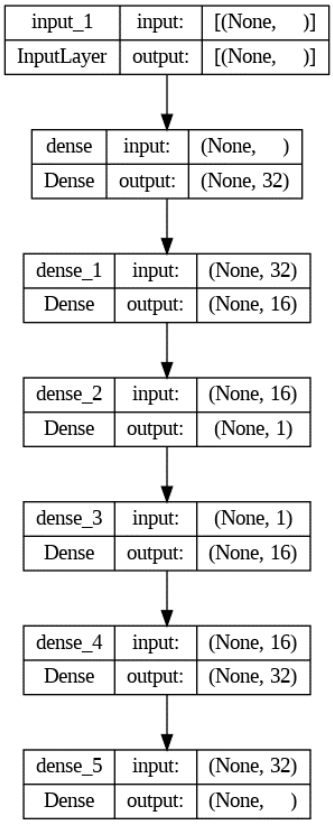
a) **Codificador,** que consta de una primera capa de entrada con el número de ítems, una capa densa con 32 neuronas, una capa densa con 16 neuronas y una salida de una neurona, la cual representa el trazo latente medido por la prueba.

b) **Decodificador,** que tiene una capa de entrada de una neurona, una capa densa de 16 neuronas, una capa densa de 32 neuronas y una capa de salida de igual número de neuronas que de ítems.

La Figura 4 muestra la arquitectura del AE candidato para implementar a nivel de forma y juntura de formas.

Figura 4

AE aplicado para formas y junturas



3.3 Método de equiparación entre aplicaciones

Como se señaló en la sección 2.1[1], siempre existe una pareja de formas entre aplicaciones consecutivas que comparte aproximadamente tres cuartas partes de los ítems. Teniendo en cuenta esto, el algoritmo propuesto para equiparar las calificaciones dentro de cada aplicación puede aplicarse de manera más simplificada entre aplicaciones. Consistiría en aplicar un AE únicamente a la pareja de formas que contenga el número máximo de ítems compartidos. Así, se utilizaría el Método SL para equiparar la aplicación 2 a la aplicación 1 (θ_2^*) mediante las funciones g(.) y f(.), obtenidas a partir de:

$$\theta_2^* = g(f(\theta_2 , \theta_{12}), \theta_1),$$

donde:

- θ_2 es la habilidad de los evaluados con la forma de la segunda aplicación después de la equiparación dentro de la aplicación,
- θ_{12} es la habilidad de los evaluados de la juntura a partir del AE,
- θ_1 es la habilidad de los evaluados con la forma de la primera aplicación después de la equiparación dentro de la aplicación,
- $f(.)$ es la función que equipara las habilidades de la forma de la segunda aplicación con la de la juntura, y
- $g(.)$ es la función que equipara la habilidad de la juntura con la de la forma de la primera aplicación.

[1] Consultar Anexo 1

4. Resultados

En este capítulo se presentarán los resultados de la implementación del algoritmo propuesto para la calificación de cada una de las pruebas evaluadas en el examen Saber 11 para las aplicaciones 2018-1, 2018-2, 2019-1 y 2019-2. En la primera subsección, se mostrarán los resultados de la metodología aplicada para cada aplicación, garantizando comparabilidad mediante el uso del método propuesto entre cada forma y, en la segunda subsección, se mostrarán los resultados empleando la metodología entre aplicaciones, garantizando así una comparabilidad temporal entre todos los evaluados a través del tiempo.

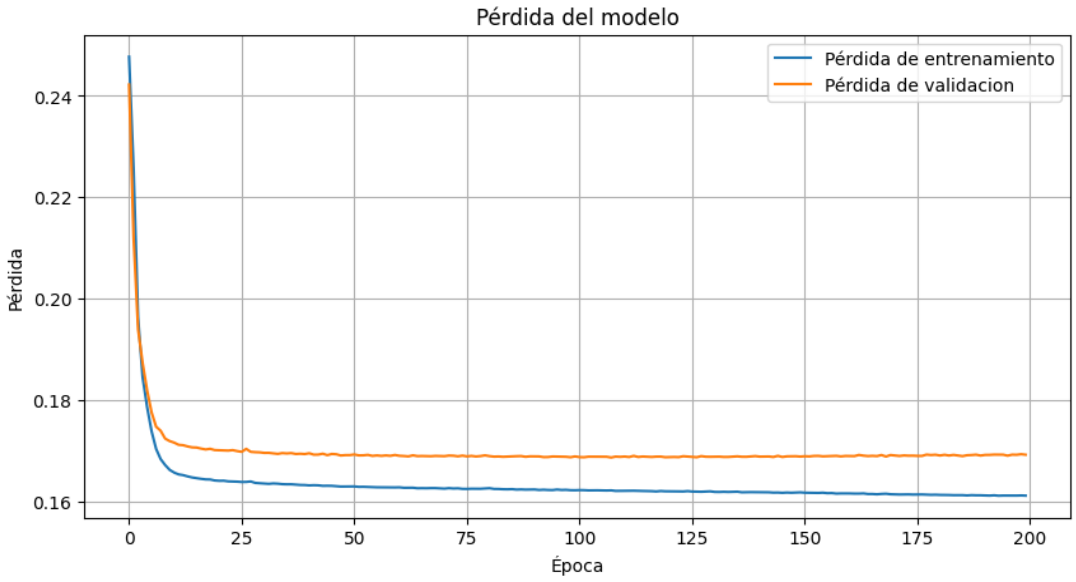
4.1 Asignación de puntajes en la aplicación

Como se mencionó en la sección anterior, la red neuronal definida para la estimación de la habilidad consta de 2 capas densas de neuronas en la parte codificadora y 2 más en la decodificadora. Además, en todos los casos se usaron 20 épocas de entrenamiento, puesto que con un número mayor no se observó disminución en la función de pérdida. Por otro lado, el optimizador escogido fue *Adam*, la función de pérdida fue el *error cuadrático medio* y el lote de ejemplos varió dependiendo del caso, en 64 o 32. Adicional a lo anterior, para poder estimar la habilidad en una prueba de una aplicación, se debe realizar el ajuste de un AE por cada forma y cada pareja de formas, se definió una regla en el código que, en caso de que en la última época la pérdida no fuera inferior a 0.2, se vuelve a ejecutar el entrenamiento, pero disminuyendo el lote de ejemplos a 32 o 16.

Por medio de Python (Rossum et al, 2009) y el paquete TensorFlow (Developers, 2022), se ejecutaron cada uno de los AE necesarios por prueba, los cuales consisten en uno por cada forma aplicada (14) y uno por cada pareja de formas (13). Es decir, por cada prueba se ajustan 27 redes neuronales. La *Gráfica 2* muestra la evaluación de la función de pérdida a través de 200 épocas para la Forma A de Matemáticas en 2018-1. Este comportamiento es muy similar para cada una de las formas de cada prueba en cada aplicación.

Gráfica 2

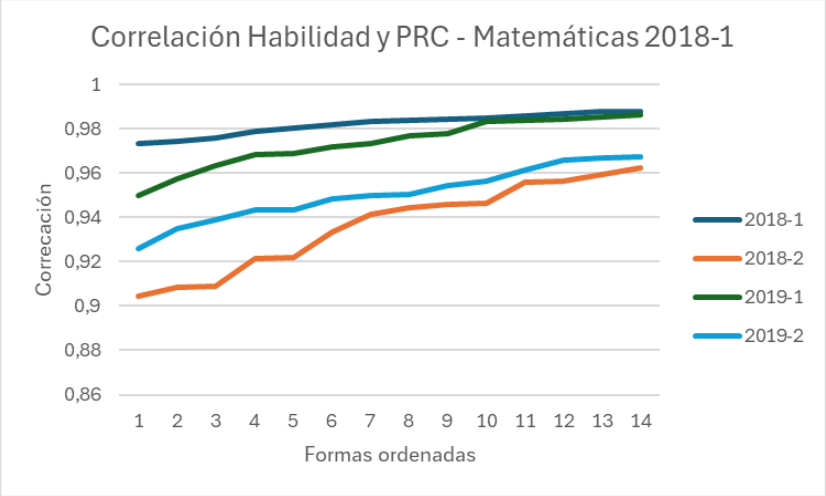
Función de pérdida a través de 200 épocas para la Forma A de Matemáticas en 2018-1



Una vez ejecutado esto, se procede a comparar la estimación de la habilidad proporcionada por el AE, la cual resultó altamente correlacionada con el porcentaje de respuestas correctas para todas las formas, pruebas y aplicaciones. La *Gráfica 3* muestra estas correlaciones por forma, después de ser ordenadas de menor a mayor, para la prueba de matemáticas. El anexo 3 presenta la correlación entre el porcentaje de respuestas correctas y la habilidad estimada para todas las pruebas evaluadas en el examen.

Gráfica 3

Correlaciones para la prueba de matemáticas.



En la prueba de Matemáticas, se puede notar que las aplicaciones de calendario A, las cuales son mucho más grandes en términos de número de evaluados que las de calendario B, tienden a tener menor correlación con el porcentaje de respuestas correctas. De hecho, el promedio de la correlación por aplicaciones para esta prueba es de 0.98 para calendario B de ambas aplicaciones y 0.94 y 0.95 para las aplicaciones de calendario A de 2018 y 2019, respectivamente.

4.2 Equiparación entre aplicaciones

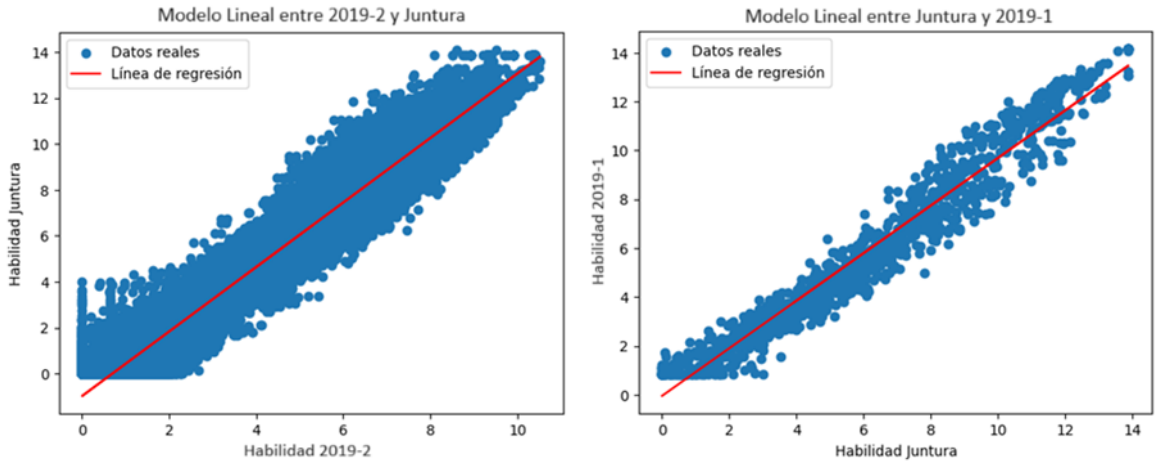
Una vez seleccionadas las parejas de formas entre aplicaciones continuas, se aplicó la metodología descrita en la sección 6.3 para equiparar todas las aplicaciones seleccionando como línea base la primera aplicación (2018-1). Además, teniendo en cuenta que el Calendario A está compuesto por muchos más evaluados que los de Calendario B y, considerando que el entrenamiento de las redes neuronales es muy susceptible a los ejemplos que recibe, para entrenar el AE de la juntura se seleccionó de forma aleatoria un subconjunto de evaluados de la forma del calendario A del mismo tamaño que la forma del calendario B.

Por otro lado, el uso de un modelo lineal como función de equiparación es razonable en todos los casos. La *Gráfica 4* muestra el enlace entre la aplicación 2019-1 (forma C) y la aplicación 2019-2 (forma A) de Matemáticas, mediante la metodología de SL. Este tipo de gráficos pueden ilustrar la relación entre las aplicaciones y la juntura. En dado caso en que no pueda asumirse una relación lineal, debe considerarse otro tipo de relación funcional.

Como paso final, una vez se obtienen todas las calificaciones en escala de la aplicación base, se procedió a estandarizar la serie restando y dividiendo por la media y desviación estándar de la aplicación base, teniendo así una escala en la que se pueden apreciar cambios a través de cada una de las aplicaciones. La *Gráfica 5* muestra la distribución de las calificaciones de matemáticas a lo largo de las 4 aplicaciones. En esta gráfica se puede observar como las aplicaciones de calendario B toman valores mayores que los de calendario A, lo cual es consistente con los resultados publicados en cada aplicación por el ICFES. Además, puede observarse que las calificaciones son más parecidas dentro calendarios que entre calendarios, lo cual también es consistente con los resultados históricos del examen Saber 11.

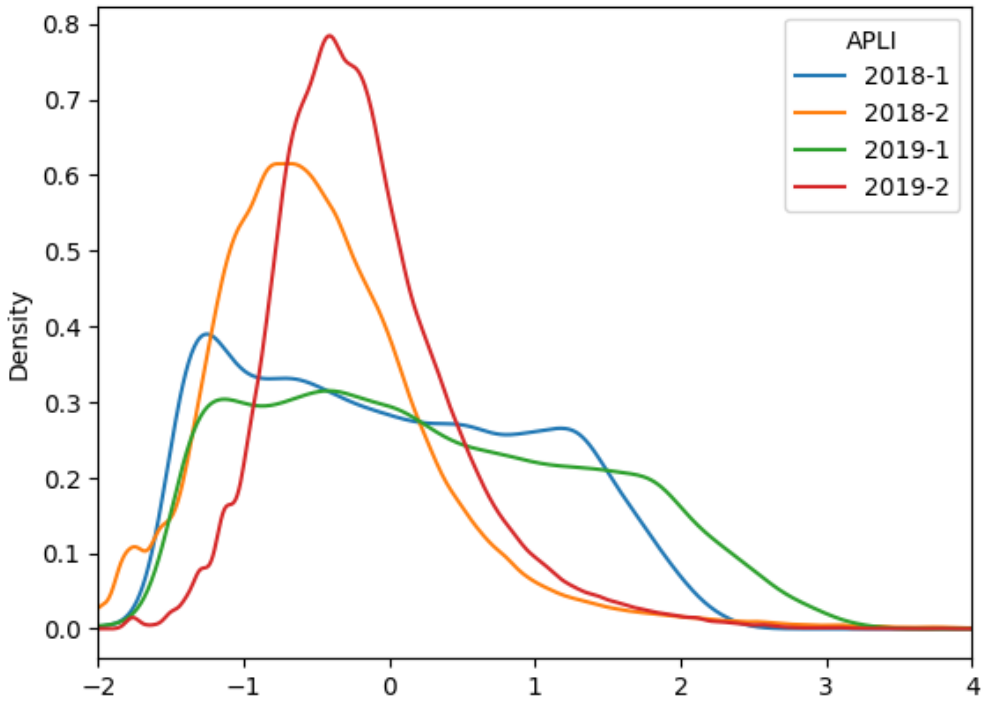
Gráfica 4

Enlace entre 2019-1 y 2019-2 para matemáticas.



Gráfica 5

Distribución de las calificaciones de Matemáticas después de estandarizar.



Las *Gráficas 6, 7 y 8* muestran la distribución de Ciencias, Sociales y Lectura, respectivamente. Además, la *Tabla 7* presenta la media y desviación estándar por aplicación de cada una de las pruebas después del proceso de estandarización y después de multiplicar por 10 y sumar

50, dejando así en términos de puntaje como son presentados por el ICFES, como se especificó en la sección 2.5.

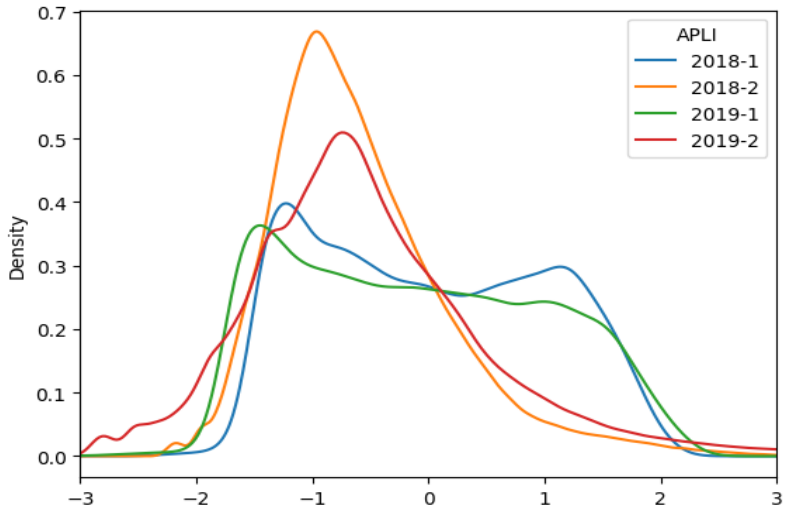
Tabla 7

Media y desviación del puntaje de las pruebas del Saber 11.

PRUEBA	APLICACIÓN	CALENDARIO	MEDIA	DES. EST.
Matemáticas	2018-1	B	50,00	10,00
	2018-2	A	45,61	7,82
	2019-1	B	52,71	11.,40
	2019-2	A	48,35	6.27
Ciencias Naturales	2018-1	B	50,00	10,00
	2018-2	A	44,13	7,63
	2019-1	B	48,93	10,89
	2019-2	A	45,01	10,90
Sociales y Ciudadanas	2018-1	B	50,00	10,00
	2018-2	A	44,48	11,98
	2019-1	B	52,10	14,13
	2019-2	A	44,29	10,58
Lectura Crítica	2018-1	B	50,00	10,00
	2018-2	A	45,05	6,74
	2019-1	B	49,70	9,65
	2019-2	A	45,04	13.62

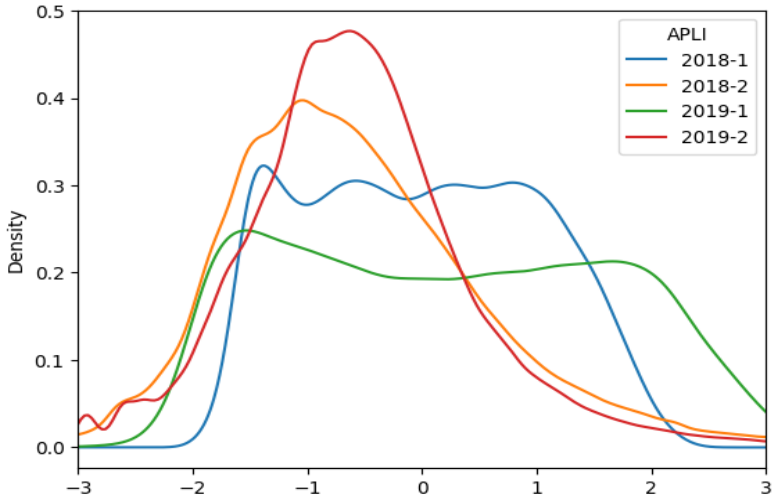
Gráfica 6

Distribución de las calificaciones de Ciencias Naturales después de estandarizar.



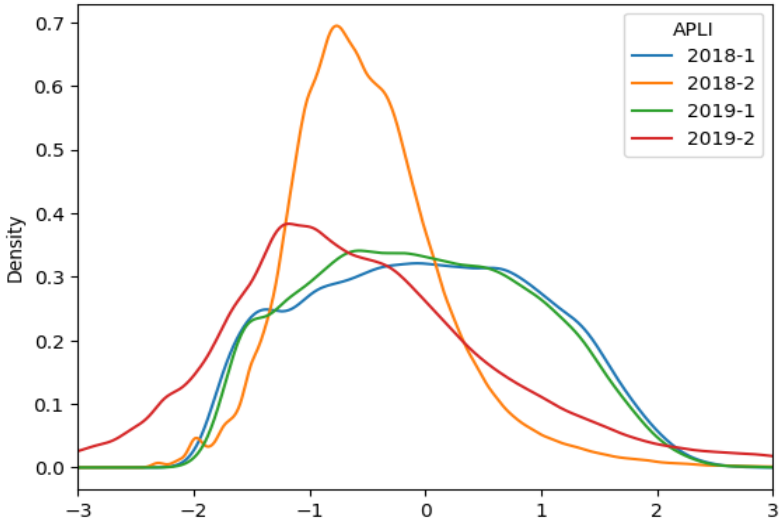
Gráfica 7

Distribución de las calificaciones de Sociales y Ciudadanas después de estandarizar.



Gráfica 8

Distribución de las calificaciones de Lectura Crítica después de estandarizar.



4.3 Evaluación y comparación del método de calificación

Para evaluar el ajuste de las calificaciones generadas mediante la metodología propuesta, se procedió a realizar predicciones para cada uno de los evaluados en el conjunto de prueba.

Posteriormente, se transformaron estos valores utilizando la condición de que si la predicción es mayor a 0.5, se asigna el valor 1; en caso contrario, se asigna el valor 0. Esto se debe a que la capa final de los AE proporciona valores reales entre 0 y 1, ya que la última capa tenía como función de activación la función sigmoide (ver sección 7.1). Para contar con valores de referencia y así poder comparar los porcentajes de aciertos de las predicciones, se ejecutó un modelo 2PL para cada una de las aplicaciones por prueba, utilizando el mismo método de equiparación usado en la metodología propuesta, es decir, Stocking-Lord.

La *Tabla 8* presenta los porcentajes de aciertos de las predicciones, comparando los valores reales del conjunto de prueba con las predicciones obtenidas mediante los AE y el modelo de TRI ajustado, para todas las pruebas y aplicaciones. Se destacan en negrilla aquellos casos donde la proporción de aciertos es mayor.

Tabla 8

Porcentaje de aciertos de predicción por los AE y por 2PL

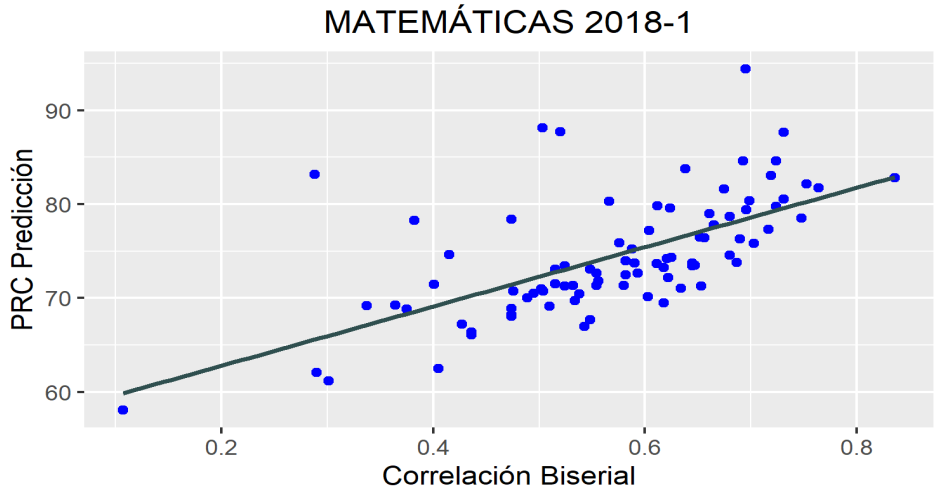
	AE (%)	TRI (%)
CN	**72,81**	72.06
2018-1	75,86	**76,15**
2018-2	**72,21**	71,45
2019-1	75,79	**76,02**
2019-2	**73,20**	72,38
LC	**72,60**	71.70
2018-1	**72,39**	72,06
2018-2	**72,12**	71,16
2019-1	75,16	**75,65**
2019-2	**72,99**	72,07
MA	**73,51**	72.62
2018-1	75,20	**75,32**
2018-2	**74,22**	73,20
2019-1	**76,98**	76,93
2019-2	**72,61**	71,79
SC	**72,13**	71.44
2018-1	74,49	**74,79**
2018-2	**72,37**	71,69
2019-1	75,60	**75,95**
2019-2	**71,69**	70,91
Total general	**72,76**	71,95

Es evidente que para todos los casos de todas las pruebas en el calendario A, los aciertos de predicción mediante la metodología propuesta son siempre superiores a los resultados del modelo 2PL. Del mismo modo, en todos los casos, excepto en la aplicación 2018-1 de Lectura Crítica, el modelo de TRI tiene un rendimiento superior. En cuanto a los totales generales, dado que la población evaluada en el calendario A es considerablemente mayor, la metodología propuesta presenta un mejor desempeño que los modelos de TRI en todas las pruebas.

Por otro lado, en un análisis de las causas de las variaciones en la correlación entre el porcentaje de respuestas correctas y la habilidad estimada por los AE, se identificó una relación lineal entre los porcentajes de aciertos de predicción a nivel de ítem y la correlación biserial. Como se mencionó, esta última mide la relación entre el ítem y el porcentaje de respuestas correctas al test. En otras palabras, si una forma está compuesta por ítems que tienen una baja correlación con el trazo latente, entonces esa forma perderá poder de predicción al ser aplicada la metodología propuesta, así mismo, si la prueba está compuesta por ítems con buenas características psicométricas, el poder predictivo de los AE será mayor. La *Gráfica 9* muestra la correlación biserial y el porcentaje de aciertos pronosticados para la prueba de matemáticas 2018-1. En el *Anexo 4* se incluyen las gráficas que muestran la relación entre la correlación biserial y la del porcentaje de aciertos de predicción a nivel de prueba y aplicación.

Gráfica 9

Relación entre la correlación biserial y el porcentaje de predicción del AE a nivel de ítem



5. Discusión y análisis de resultados

La metodología propuesta tuvo resultados interesantes y pueden contribuir en la asignación de puntajes en exámenes estandarizados. En primer lugar, se evidencia una variabilidad notoria en el porcentaje de ajuste entre distintas aplicaciones, destacándose un rendimiento superior durante el calendario B. La metodología inicialmente empleada se encontró restringida al uso de una única dimensión en el trazo latente, limitación impuesta por la naturaleza unidimensional de los modelos de calificación previamente utilizados por el ICFES. Sin embargo, se identificó un potencial para mejorar el ajuste de los resultados al considerar la inclusión de múltiples dimensiones en el trazo latente.

Por otro lado, al analizar la distribución de las calificaciones por aplicación, se puede observar que en algunos casos es razonable pensar que la población evaluada podría estar conformada por subpoblaciones con diferente distribución de habilidad (gráfica 5 y gráfica 6). En estas situaciones, cuando existe evidencia de esto, al aplicar modelos de TRI, se suelen cambiar los modelos usuales (1PL, 2PL, 3PL) por modelos multigrupo, comúnmente empleado en evaluaciones como PISA, para abordar la variabilidad poblacional. No obstante, lo destacado es que la metodología propuesta, demostró ser robusta y eficaz sin necesidad de realizar ajustes teniendo esto en cuenta.

Este resultado sugiere que la metodología basada en AE es capaz de gestionar de manera efectiva la diversidad poblacional, incluso en situaciones en las que la aplicación de la TRI aconseja modelos multigrupo (Bock, 1997). La resiliencia de la metodología ante la variabilidad poblacional plantea preguntas interesantes sobre la capacidad adaptativa y la generalización de la red neuronal utilizada. La robustez demostrada de la metodología propuesta tiene implicaciones significativas para la calificación de exámenes estandarizados, especialmente aquellos que enfrentan diversidad poblacional. Estos resultados sugieren que la implementación de AE en combinación con la TRI podría ofrecer una alternativa eficaz y adaptable, superando la necesidad de modelos específicos para poblaciones diversas.

Otro aspecto destacado es la identificación de una relación lineal positiva significativa entre el ajuste a nivel de ítem y la correlación biserial. Este hallazgo indica que a medida que la correlación entre el ítem y el trazo latente aumenta, el comportamiento de la red neuronal mejora

proporcionalmente. La calidad de construcción del ítem ejerce una influencia directa sobre el desempeño de la red, evidenciando una similitud con los principios fundamentales de la TRI. La relación positiva implica que ítems mejor construidos, en términos de su correlación con el trazo latente, contribuyen significativamente a la mejora global del rendimiento de la red.

Esta relación entre el ajuste, la correlación habilidad-respuestas correctas y la información del test a nivel de forma subraya la importancia de un análisis detallado del ajuste para guiar estrategias específicas que mejoren la capacidad de los ítems para estimar con precisión la habilidad de los evaluados. Estas implicaciones tienen un impacto directo en la calidad general de la evaluación y sugieren que la consideración de estas interrelaciones puede contribuir significativamente a la mejora continua de los procesos de calificación.

El análisis detallado revela que, al emplear AE, se obtiene una herramienta valiosa y alternativa a la TRI para evaluar la calidad de los instrumentos de medición de trazos latentes. El comportamiento conjunto de las características mencionadas proporciona una visión integral de la efectividad de los ítems para medir la habilidad de los evaluados. Esto destaca la capacidad única de las AE para capturar patrones complejos y relaciones no lineales, que a menudo no son plenamente abordadas por enfoques tradicionales como la TRI.

6. Conclusiones y trabajos futuros

6.1 Conclusiones

En este trabajo de maestría en ingeniería de sistemas y computación, los resultados obtenidos ofrecen una visión integral sobre la aplicación de AutoEncoders en la calificación de los exámenes Saber 11, abordando aspectos fundamentales tanto en términos de ajuste de modelos, gestión de la diversidad poblacional y comparabilidad de la construcción de las escalas de medición entre aplicaciones.

La identificación de múltiples poblaciones en aplicaciones de AE para la calificación de los exámenes Saber 11 es un hallazgo significativo. A pesar de que en la literatura se sugiere aplicar modelos TRI multigrupo en situaciones donde se presentan subpoblaciones con diferente distribución de habilidad, la metodología propuesta basada en redes neuronales y AutoEncoders demostró ser robusta y eficaz en presencia o no de subpoblaciones evaluadas. Esto sugiere que la metodología basada en AE es capaz de gestionar efectivamente la diversidad poblacional, incluso en contextos donde la TRI aconseja modelos específicos para diferentes grupos poblacionales.

La metodología de equiparación de Stocking-Lord resultó de vital importancia en la implementación de los AutoEncoders para cada una de las formas distribuidas durante el examen, permitiendo así comparabilidad dentro y entre aplicaciones para cada una de las pruebas. Esto fue posible debido a la particularidad en la forma de construcción del examen Saber 11, la cual hace uso del diseño en bloques incompletos balanceados.

La relación lineal positiva entre el ajuste a nivel de ítem y la correlación biserial resalta la importancia de la calidad de construcción del ítem en la mejora del rendimiento de la red neuronal. Este hallazgo, con paralelismos notables con la TRI, destaca la atención esencial a la calidad de los ítems para optimizar el desempeño de las redes neuronales artificiales en la calificación de exámenes estandarizados.

En términos de eficiencia temporal, se destaca la capacidad de los AE para converger rápidamente, permitiendo la posibilidad de disminuir el tiempo de procesamiento en la asignación de puntajes en exámenes estandarizados. Este hallazgo tiene implicaciones prácticas significativas

y destaca la adaptabilidad de los modelos de AE a la variabilidad inherente en los datos de los exámenes Saber 11.

La integración de estos resultados sugiere que la implementación de AE en combinación con la TRI podría ofrecer una alternativa eficaz y adaptable en la calificación de exámenes estandarizados. Este enfoque supera la necesidad de modelos específicos para poblaciones diversas y destaca la importancia de considerar la calidad de construcción de los ítems para optimizar la precisión de la evaluación.

6.2 Trabajos futuros

Los hallazgos de este estudio motivan ahondar e incursionar en el uso de AE en exámenes estandarizados. Desde esta perspectiva, cabe resaltar que dentro de esta motivación se sugiere la necesidad de investigaciones que exploren en detalle las condiciones bajo las cuales las AE pueden lograr una convergencia más rápida y precisa, ya sea modificando el número de capas, neuronas, función de pérdida y optimizador.

Por otro lado, la identificación de las capacidades y limitaciones de las AE en comparación con la TRI posibilitaría el diseño de estudios que evalúen la sensibilidad y especificidad de ambas metodologías en diferentes contextos de evaluación, contribuyendo así a una comprensión más profunda de sus aplicaciones respectivas.

El encuentro de la correlación abre oportunidades para explorar en mayor detalle las características específicas de los ítems que resultan más informativas para la estimación de habilidad. Se podrían diseñar estrategias específicas de construcción de ítems basadas en estos patrones identificados para mejorar aún más la precisión de la medición de habilidad en evaluaciones estandarizadas, como por ejemplo usar técnicas de lenguaje natural. Perspectivas futuras podrían explorar en mayor detalle las condiciones bajo las cuales los AE muestran un rendimiento óptimo y considerar la generalización de estos enfoques en diferentes tipos de evaluaciones.

Anexo 1. A. Conteo de ítems comunes entre formas por aplicación.

A continuación se presentan las tablas que muestran los ítems comunes entre las 14 formas por prueba de cada aplicación.

Matemáticas:

2018-1

FORMA	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
A	44	22	11	22	22	11	33	11	22	22	22	22	22	22
B	22	44	22	22	22	22	11	11	22	33	22	11	22	22
C	11	22	43	11	22	21	21	21	10	22	21	22	32	22
D	22	22	11	44	22	11	11	33	22	22	22	22	22	22
E	22	22	22	22	44	22	22	22	22	22	0	22	22	22
F	11	22	21	11	22	43	21	21	32	22	21	22	10	22
G	33	11	21	11	22	21	43	21	21	22	21	22	21	11
H	11	11	21	33	22	21	21	43	21	22	21	22	21	11
I	22	22	10	22	22	32	21	21	43	11	21	11	21	22
J	22	33	22	22	22	22	22	22	11	44	22	22	11	11
K	22	22	21	22	0	21	21	21	21	22	43	22	21	22
L	22	11	22	22	22	22	22	22	11	22	22	44	11	33
M	22	22	32	22	22	10	21	21	21	11	21	11	43	22
N	22	22	22	22	22	22	11	11	22	11	22	33	22	44

2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	22	11	22	22	22	22	22	33	11	11	22	22	22
B	22	43	32	22	11	22	21	22	21	21	10	11	21	22
C	11	32	43	11	22	22	21	22	21	21	21	22	10	22
D	22	22	11	44	22	22	22	22	11	33	11	22	22	22
E	22	11	22	22	44	11	22	22	22	22	22	22	11	33
F	22	22	22	22	11	44	22	22	11	11	22	33	22	22
G	22	21	21	22	22	22	43	0	21	21	21	22	21	22
H	22	22	22	22	22	22	0	44	22	22	22	22	22	22
I	33	21	21	11	22	11	21	22	43	21	21	22	21	11
J	11	21	21	33	22	11	21	22	21	43	21	22	21	11
K	11	10	21	11	22	22	21	22	21	21	43	22	32	22
L	22	11	22	22	22	33	22	22	22	22	22	44	11	11
M	22	21	10	22	11	22	21	22	21	21	32	11	43	22
N	22	22	22	22	33	22	22	22	11	11	22	11	22	44

2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	22	11	11	22	22	22	22	33	22	22	11	22	22
B	22	44	22	22	11	22	11	33	22	11	22	22	22	22
C	11	22	44	22	11	22	22	22	22	33	22	22	22	11
D	11	22	22	44	22	22	11	11	22	22	22	22	22	33
E	22	11	11	22	44	11	22	22	22	22	22	33	22	22
F	22	22	22	22	11	44	33	11	22	11	22	22	22	22
G	22	11	22	11	22	33	44	22	11	22	22	22	22	22
H	22	33	22	11	22	11	22	44	11	22	22	22	22	22
I	33	22	22	22	22	22	11	11	44	22	22	22	22	11
J	22	11	33	22	22	11	22	22	22	44	22	11	22	22
K	22	22	22	22	22	22	22	22	22	22	44	22	0	22
L	11	22	22	22	33	22	22	22	22	11	22	44	22	11
M	22	22	22	22	22	22	22	22	22	22	0	22	44	22
N	22	22	11	33	22	22	22	22	11	22	22	11	22	44

2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	22	11	33	11	22	22	22	22	22	22	22	22	11
B	22	44	11	22	22	22	11	11	22	22	22	22	22	33
C	11	11	44	22	22	33	22	22	22	22	22	11	22	22
D	33	22	22	44	22	22	22	22	22	22	11	11	11	22
E	11	22	22	22	44	22	22	22	22	22	11	33	11	22
F	22	22	33	22	22	44	11	11	22	22	22	22	22	11
G	22	11	22	22	22	11	44	22	22	22	33	22	11	22
H	22	11	22	22	22	11	22	44	22	22	11	22	33	22
I	22	22	22	22	22	22	22	22	44	0	22	22	22	22
J	22	22	22	22	22	22	22	22	0	44	22	22	22	22
K	22	22	22	11	11	22	33	11	22	22	44	22	22	22
L	22	22	11	11	33	22	22	22	22	22	22	44	22	11
M	22	22	22	11	11	22	11	33	22	22	22	22	44	22
N	11	33	22	22	22	11	22	22	22	22	22	11	22	44

Ciencias Naturales:

2018-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	52	26	26	26	26	26	26	26	26	26	26	26	0	26
B	26	52	26	26	13	26	13	26	26	26	39	26	26	13
C	26	26	52	26	39	26	26	26	13	26	26	13	26	13
D	26	26	26	52	13	26	26	26	13	26	26	13	26	39
E	26	13	39	13	52	26	26	26	26	13	26	26	26	26
F	26	26	26	26	26	52	13	26	13	26	13	39	26	26
G	26	13	26	26	26	13	52	13	26	39	26	26	26	26
H	26	26	26	26	26	26	13	52	39	26	13	13	26	26
I	26	26	13	13	26	13	26	39	52	26	26	26	26	26
J	26	26	26	26	13	26	39	26	26	52	13	26	26	13
K	26	39	26	26	26	13	26	13	26	13	52	26	26	26
L	26	26	13	13	26	39	26	13	26	26	26	52	26	26
M	0	26	26	26	26	26	26	26	26	26	26	26	52	26
N	26	13	13	39	26	26	26	26	26	13	26	26	26	52

2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	52	26	26	26	13	26	26	26	26	39	26	26	13	13
B	26	52	26	13	26	26	13	39	13	26	26	26	26	26
C	26	26	52	26	26	0	26	26	26	26	26	26	26	26
D	26	13	26	52	13	26	26	26	26	26	13	26	39	26
E	13	26	26	13	52	26	39	26	26	26	26	13	26	26
F	26	26	0	26	26	52	26	26	26	26	26	26	26	26
G	26	13	26	26	39	26	52	26	26	26	13	26	13	26
H	26	39	26	26	26	26	26	52	26	13	13	26	26	13
I	26	13	26	26	26	26	26	26	52	13	39	26	26	13
J	39	26	26	26	26	26	26	13	13	52	26	13	26	26
K	26	26	26	13	26	26	13	13	39	26	52	26	26	26
L	26	26	26	26	13	26	26	26	26	13	26	52	13	39
M	13	26	26	39	26	26	13	26	26	26	26	13	52	26
N	13	26	26	26	26	26	26	13	13	26	26	39	26	52

2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	52	26	39	13	13	26	26	26	26	26	13	26	26	26
B	26	52	26	26	26	26	26	26	26	26	26	26	0	26
C	39	26	52	26	26	13	13	26	26	26	26	13	26	26
D	13	26	26	52	26	26	13	26	13	26	26	26	26	39
E	13	26	26	26	52	26	13	26	39	26	26	26	26	13
F	26	26	13	26	26	52	26	39	26	13	13	26	26	26
G	26	26	13	13	13	26	52	26	26	26	39	26	26	26
H	26	26	26	26	26	39	26	52	13	26	26	13	26	13
I	26	26	26	13	39	26	26	13	52	13	26	26	26	26
J	26	26	26	26	26	13	26	26	13	52	26	39	26	13
K	13	26	26	26	26	13	39	26	26	26	52	13	26	26
L	26	26	13	26	26	26	26	13	26	39	13	52	26	26
M	26	0	26	26	26	26	26	26	26	26	26	26	52	26
N	26	26	26	39	13	26	26	13	26	13	26	26	26	52

2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	52	13	39	26	26	13	26	26	26	13	26	26	26	26
B	13	52	26	13	26	26	13	39	26	26	26	26	26	26
C	39	26	52	13	26	26	13	13	26	26	26	26	26	26
D	26	13	13	52	39	26	26	26	26	26	26	13	26	26
E	26	26	26	39	52	26	13	26	13	26	26	26	13	26
F	13	26	26	26	26	52	26	13	39	26	26	26	13	26
G	26	13	13	26	13	26	52	26	26	26	26	39	26	26
H	26	39	13	26	26	13	26	52	26	13	26	26	26	26
I	26	26	26	26	13	39	26	26	52	13	26	13	26	26
J	13	26	26	26	26	26	26	13	13	52	26	26	39	26
K	26	26	26	26	26	26	26	26	26	26	52	26	26	0
L	26	26	26	13	26	26	39	26	13	26	26	52	13	26
M	26	26	26	26	13	13	26	26	26	39	26	13	52	26
N	26	26	26	26	26	26	26	26	26	26	0	26	26	52

Lectura Crítica:

2018-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	36	18	9	18	18	27	18	18	18	9	18	9	18	18
B	18	36	18	9	18	18	18	18	18	9	18	27	9	18
C	9	18	36	18	9	18	18	27	18	18	18	18	18	9
D	18	9	18	36	27	18	9	18	18	18	18	18	18	9
E	18	18	9	27	36	9	18	18	18	18	18	18	9	18
F	27	18	18	18	9	36	18	9	18	18	18	18	18	9
G	18	18	18	9	18	18	36	18	18	27	18	9	9	18
H	18	18	27	18	18	9	18	36	18	9	18	9	18	18
I	18	18	18	18	18	18	18	18	36	18	0	18	18	18
J	9	9	18	18	18	18	27	9	18	36	18	18	18	18
K	18	18	18	18	18	18	18	18	0	18	36	18	18	18
L	9	27	18	18	18	18	9	9	18	18	18	36	18	18
M	18	9	18	18	9	18	9	18	18	18	18	18	36	27
N	18	18	9	9	18	9	18	18	18	18	18	18	27	36

2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	36	27	9	18	18	9	18	18	9	18	18	18	18	18
B	27	36	18	9	18	18	18	18	18	9	18	18	9	18
C	9	18	36	18	18	18	18	9	18	9	18	18	18	27
D	18	9	18	36	18	18	18	18	9	18	9	27	18	18
E	18	18	18	18	36	18	0	18	18	18	18	18	18	18
F	9	18	18	18	18	36	18	27	18	9	18	18	18	9
G	18	18	18	18	0	18	36	18	18	18	18	18	18	18
H	18	18	9	18	18	27	18	36	18	18	9	9	18	18
I	9	18	18	9	18	18	18	18	36	27	18	18	9	18
J	18	9	9	18	18	9	18	18	27	36	18	18	18	18
K	18	18	18	9	18	18	18	9	18	18	36	18	27	9
L	18	18	18	27	18	18	18	9	18	18	18	36	9	9
M	18	9	18	18	18	18	18	18	9	18	27	9	36	18
N	18	18	27	18	18	9	18	18	18	18	9	9	18	36

2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	36	18	18	9	18	18	18	18	27	18	9	18	18	9
B	18	36	0	18	18	18	18	18	18	18	18	18	18	18
C	18	0	36	18	18	18	18	18	18	18	18	18	18	18
D	9	18	18	36	9	18	18	27	18	18	18	18	9	18
E	18	18	18	9	36	9	18	18	9	18	18	27	18	18
F	18	18	18	18	9	36	9	18	18	9	18	18	27	18
G	18	18	18	18	18	9	36	18	18	18	27	9	18	9
H	18	18	18	27	18	18	18	36	9	18	9	18	18	9
I	27	18	18	18	9	18	18	9	36	18	18	18	9	18
J	18	18	18	18	18	9	18	18	18	36	9	9	18	27
K	9	18	18	18	18	18	27	9	18	9	36	18	18	18
L	18	18	18	18	27	18	9	18	18	9	18	36	9	18
M	18	18	18	9	18	27	18	18	9	18	18	9	36	18
N	9	18	18	18	18	18	9	9	18	27	18	18	18	36

2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	35	17	0	17	16	18	17	18	16	18	18	18	16	18
B	17	35	18	8	16	18	26	18	16	9	18	18	16	9
C	0	18	36	18	18	18	18	18	18	18	18	18	18	18
D	17	8	18	35	25	18	17	18	16	18	9	9	16	18
E	16	16	18	25	35	9	16	18	17	9	18	18	7	18
F	18	18	18	18	9	36	18	27	9	18	9	18	18	18
G	17	26	18	17	16	18	35	9	7	18	18	9	16	18
H	18	18	18	18	18	27	9	36	18	18	18	18	9	9
I	16	16	18	16	17	9	7	18	35	18	18	18	25	9
J	18	9	18	18	9	18	18	18	18	36	27	9	18	18
K	18	18	18	9	18	9	18	18	18	27	36	18	9	18
L	18	18	18	9	18	18	9	18	18	9	18	36	18	27
M	16	16	18	16	7	18	16	9	25	18	9	18	34	18
N	18	9	18	18	18	18	18	9	9	18	18	27	18	36

Sociales y Ciudadanas:

2018-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	22	22	22	22	22	22	11	33	11	22	22	22	11
B	22	44	22	22	22	22	22	22	22	22	0	22	22	22
C	22	22	44	11	11	22	11	22	22	22	22	22	33	22
D	22	22	11	44	22	22	22	22	11	11	22	33	22	22
E	22	22	11	22	44	22	22	11	22	22	22	11	22	33
F	22	22	22	22	22	44	22	11	11	33	22	22	22	11
G	22	22	11	22	22	22	44	33	22	22	22	11	22	11
H	11	22	22	22	11	11	33	44	22	22	22	22	22	22
I	33	22	22	11	22	11	22	22	44	22	22	22	11	22
J	11	22	22	11	22	33	22	22	22	44	22	22	11	22
K	22	0	22	22	22	22	22	22	22	22	44	22	22	22
L	22	22	22	33	11	22	11	22	22	22	22	44	11	22
M	22	22	33	22	22	22	22	22	11	11	22	11	44	22
N	11	22	22	22	33	11	11	22	22	22	22	22	22	44

2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	33	22	22	22	22	11	22	22	11	22	22	22	11
B	33	44	11	22	22	22	22	22	11	22	22	22	11	22
C	22	11	44	22	22	22	11	22	22	11	22	22	22	33
D	22	22	22	44	11	11	22	22	11	22	22	22	33	22
E	22	22	22	11	44	22	33	22	22	11	22	11	22	22
F	22	22	22	11	22	44	11	22	22	33	22	11	22	22
G	11	22	11	22	33	11	44	22	22	22	22	22	22	22
H	22	22	22	22	22	22	22	44	22	22	0	22	22	22
I	22	11	22	11	22	22	22	22	44	22	22	33	22	11
J	11	22	11	22	11	33	22	22	22	44	22	22	22	22
K	22	22	22	22	22	22	22	0	22	22	44	22	22	22
L	22	22	22	22	11	11	22	22	33	22	22	44	11	22
M	22	11	22	33	22	22	22	22	22	22	22	11	44	11
N	11	22	33	22	22	22	22	22	11	22	22	22	11	44

2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	22	22	22	22	22	22	22	22	22	0	22	22	22
B	22	44	11	22	22	22	22	11	22	11	22	22	33	22
C	22	11	44	11	22	33	11	22	22	22	22	22	22	22
D	22	22	11	44	11	22	22	33	11	22	22	22	22	22
E	22	22	22	11	44	11	22	22	22	22	22	11	22	33
F	22	22	33	22	11	44	22	11	11	22	22	22	22	22
G	22	22	11	22	22	22	44	11	22	33	22	22	11	22
H	22	11	22	33	22	11	11	44	22	22	22	22	22	22
I	22	22	22	11	22	11	22	22	44	22	22	33	22	11
J	22	11	22	22	22	22	33	22	22	44	22	11	22	11
K	0	22	22	22	22	22	22	22	22	22	44	22	22	22
L	22	22	22	22	11	22	22	22	33	11	22	44	11	22
M	22	33	22	22	22	22	11	22	22	22	22	11	44	11
N	22	22	22	22	33	22	22	22	11	11	22	22	11	44

2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	44	11	22	22	11	22	22	22	22	11	22	22	33	22
B	11	44	33	22	22	22	11	22	22	22	11	22	22	22
C	22	33	44	22	11	22	22	22	22	11	22	22	11	22
D	22	22	22	44	22	22	22	0	22	22	22	22	22	22
E	11	22	11	22	44	11	22	22	33	22	22	22	22	22
F	22	22	22	22	11	44	22	22	22	33	22	11	22	11
G	22	11	22	22	22	22	44	22	22	22	22	33	11	11
H	22	22	22	0	22	22	22	44	22	22	22	22	22	22
I	22	22	22	22	33	22	22	22	44	11	22	11	22	11
J	11	22	11	22	22	33	22	22	11	44	22	22	22	22
K	22	11	22	22	22	22	22	22	22	22	44	11	11	33
L	22	22	22	22	22	11	33	22	11	22	11	44	22	22
M	33	22	11	22	22	22	11	22	22	22	11	22	44	22
N	22	22	22	22	22	11	11	22	11	22	33	22	22	44

Anexo 1. B. Conteo de ítems comunes entre formas de emparejamiento de aplicaciones.

A continuación se presentan las tablas que muestran los ítems comunes entre las 14 formas de la pareja de aplicaciones que se van a equiparar. En las filas aparece la primera aplicación y en las columnas la de la segunda, en orden cronológico.

Matemáticas:

2018-1 y 2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	24	13	5	23	15	16	25	3	13	12	4	16	12	15
B	12	11	4	22	15	16	15	12	1	11	16	5	23	26
C	4	13	17	1	6	7	14	5	14	11	16	7	12	6
D	25	14	3	34	24	13	22	15	14	23	12	13	23	24
E	5	5	8	13	17	7	3	18	5	13	16	7	13	17
F	1	12	13	12	13	2	11	13	11	22	22	2	21	13
G	13	12	15	12	15	5	24	3	23	22	14	16	11	4
H	14	13	13	23	24	2	21	15	24	33	22	13	22	13
I	12	23	12	23	12	12	21	13	11	22	21	1	32	23
J	12	0	4	22	26	5	15	12	12	22	16	16	12	15
K	22	21	11	22	12	12	33	0	21	21	11	12	21	12
L	15	5	6	13	15	4	12	6	15	13	2	15	1	4
M	15	24	16	12	5	17	24	5	14	11	15	6	23	16
N	15	16	6	13	4	15	12	6	4	2	2	4	12	15

2018-2 y 2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	14	14	0	14	22	14	11	11	14	0	3	22	22	14
B	16	5	11	16	11	27	22	0	16	11	5	11	22	16
C	13	2	11	13	11	24	22	0	13	11	2	11	22	13
D	20	9	11	9	0	20	11	0	20	11	9	0	11	9
E	6	6	0	17	11	17	11	0	6	0	6	11	11	17
F	25	14	11	3	11	14	11	11	25	11	3	11	22	3
G	5	5	0	5	0	5	0	0	5	0	5	0	0	5
H	28	17	11	17	22	28	22	11	28	11	6	22	33	17
I	13	13	0	13	22	13	11	11	13	0	2	22	22	13
J	19	8	11	8	0	19	11	0	19	11	8	0	11	8
K	19	19	0	8	11	8	0	11	19	0	8	11	11	8
L	22	11	11	0	11	11	11	11	22	11	0	11	22	0
M	22	22	0	11	11	11	0	11	22	0	11	11	11	11
N	9	9	0	20	11	20	11	0	9	0	9	11	11	20

2019-1 y 2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	0	11	11	11	0	11	11	11	0	0	0	0	11
B	22	0	11	11	0	11	11	22	11	11	11	11	22	0
C	33	11	11	22	11	22	11	22	22	11	11	22	22	0
D	22	11	11	22	11	22	0	11	11	11	0	11	11	0
E	11	11	11	22	22	11	11	11	22	0	0	11	0	11
F	11	0	0	0	0	0	11	11	11	0	11	11	11	0
G	11	0	11	11	11	0	22	22	22	0	11	11	11	11
H	22	0	22	22	11	11	22	33	22	11	11	11	22	11
I	11	11	0	11	11	11	0	0	11	0	0	11	0	0
J	22	11	22	33	22	22	11	22	22	11	0	11	11	11
K	22	11	11	22	22	11	22	22	33	0	11	22	11	11
L	22	11	0	11	11	11	11	11	22	0	11	22	11	0
M	11	0	11	11	0	11	0	11	0	11	0	0	11	0
N	11	0	22	22	11	11	11	22	11	11	0	0	11	11

Ciencias Naturales:

2018-1 y 2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	26	26	39	13	13	0	13	13	13	26	26	26	13	26
B	13	13	13	13	13	13	26	13	0	13	0	26	0	26
C	13	13	13	0	0	0	0	13	13	0	13	13	0	0
D	13	26	13	0	13	13	13	26	13	0	13	26	0	13
E	13	26	26	0	13	0	0	13	13	13	26	13	13	13
F	26	13	26	13	0	0	13	13	13	13	13	26	0	13
G	0	13	13	0	13	0	0	0	0	13	13	0	13	13
H	26	26	26	13	13	13	26	26	13	13	13	39	0	26
I	13	26	26	13	26	13	26	13	0	26	13	26	13	39
J	13	0	13	13	0	0	13	0	0	13	0	13	0	13
K	0	26	13	0	26	13	13	13	0	13	13	13	13	26
L	13	13	26	13	13	0	13	0	0	26	13	13	13	26
M	0	13	0	0	13	13	13	13	0	0	0	13	0	13
N	13	39	26	0	26	13	13	26	13	13	26	26	13	26

2018-2 y 2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	19	20	19	20	0	13	13	19	0	13	13	7	6	20
B	3	8	6	11	3	0	11	3	3	11	14	8	6	11
C	11	11	14	14	3	0	22	11	3	22	25	11	14	14
D	11	20	11	7	13	13	31	24	13	18	31	7	11	7
E	20	19	23	22	3	13	13	20	3	13	16	6	10	22
F	15	28	15	15	13	26	17	28	13	4	17	2	2	15
G	24	18	24	18	0	13	16	24	0	16	16	5	11	18
H	9	8	9	8	0	0	17	9	0	17	17	8	9	8
I	20	32	20	19	13	26	26	33	13	13	26	6	7	19
J	18	20	21	23	3	13	12	18	3	12	15	7	8	23
K	14	32	17	22	16	26	20	27	16	7	23	6	4	22
L	6	18	6	5	13	13	24	19	13	11	24	5	6	5
M	7	21	10	11	16	13	28	20	16	15	31	8	10	11
N	5	18	8	8	16	13	23	18	16	10	26	5	8	8

2019-1 y 2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	13	12	25	0	12	25	13	0	13	12	13	25	0	12
B	13	0	13	0	0	13	13	0	13	0	13	13	0	0
C	26	0	13	13	13	13	26	13	13	0	13	26	0	13
D	13	0	0	26	13	13	26	13	13	13	0	13	13	26
E	26	0	13	26	13	26	39	13	26	13	13	26	13	26
F	0	12	12	13	12	25	13	0	13	25	0	12	13	25
G	0	12	12	0	12	12	0	0	0	12	0	12	0	12
H	0	0	0	13	0	13	13	0	13	13	0	0	13	13
I	26	12	25	13	25	25	26	13	13	12	13	38	0	25
J	13	0	13	13	0	26	26	0	26	13	13	13	13	13
K	13	0	0	13	13	0	13	13	0	0	0	13	0	13
L	13	12	25	13	12	38	26	0	26	25	13	25	13	25
M	13	12	12	26	25	25	26	13	13	25	0	25	13	38
N	13	12	12	13	25	12	13	13	0	12	0	25	0	25

Lectura Crítica:

2018-1 y 2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	18	9	0	9	0	0	18	9	9	18	9	9	9	9
B	18	9	9	18	9	0	9	9	0	9	0	9	9	18
C	18	18	9	9	9	0	9	0	9	9	9	18	0	9
D	18	18	18	18	9	9	18	0	9	9	18	27	9	9
E	18	9	18	27	9	9	18	9	0	9	9	18	18	18
F	9	9	0	0	0	0	9	0	9	9	9	9	0	0
G	9	0	0	9	0	0	9	9	0	9	0	0	9	9
H	27	18	9	18	9	0	18	9	9	18	9	18	9	18
I	9	9	9	9	9	0	0	0	0	0	0	9	0	9
J	0	0	9	9	0	9	9	0	0	0	9	9	9	0
K	18	9	9	18	0	9	27	9	9	18	18	18	18	9
L	9	9	18	18	9	9	9	0	0	0	9	18	9	9
M	9	9	9	9	0	9	18	0	9	9	18	18	9	0
N	9	0	9	18	0	9	18	9	0	9	9	9	18	9

2018-2 y 2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	9	9	9	0	18	0	18	0	9	9	18	9	9	9
B	9	9	2	0	11	0	11	0	9	2	11	9	2	2
C	9	9	2	0	2	9	2	0	9	11	2	0	11	11
D	0	9	9	9	9	9	18	9	0	9	18	0	18	9
E	18	27	2	9	11	18	20	9	18	11	20	9	20	11
F	9	18	0	9	9	9	18	9	9	0	18	9	9	0
G	0	0	7	0	7	0	7	0	0	7	7	0	7	7
H	9	18	7	9	16	9	25	9	9	7	25	9	16	7
I	9	18	0	9	0	18	9	9	9	9	9	0	18	9
J	9	18	7	9	7	18	16	9	9	16	16	0	25	16
K	18	18	0	0	9	9	9	0	18	9	9	9	9	9
L	0	9	2	9	2	9	11	9	0	2	11	0	11	2
M	18	18	7	0	16	9	16	0	18	16	16	9	16	16
N	9	9	9	0	9	9	9	0	9	18	9	0	18	18

2019-1 y 2019-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	0	18	18	9	9	9	9	9	18	9	0	9	9
B	9	0	0	9	9	9	0	9	0	0	0	9	0	9
C	0	9	27	18	18	9	9	18	18	18	18	9	9	9
D	9	9	18	18	27	9	9	18	9	9	18	18	0	18
E	9	0	9	18	9	18	0	18	9	9	0	9	9	9
F	0	9	18	9	9	9	0	18	18	9	9	9	9	0
G	0	0	9	9	9	0	9	0	0	9	9	0	0	9
H	9	0	18	27	18	18	9	18	9	18	9	9	9	18
I	0	9	18	9	18	0	9	9	9	9	18	9	0	9
J	9	0	9	18	18	9	9	9	0	9	9	9	0	18
K	0	9	9	0	9	0	0	9	9	0	9	9	0	0
L	9	9	18	18	18	18	0	27	18	9	9	18	9	9
M	0	0	9	9	0	9	0	9	9	9	0	0	9	0
N	9	9	9	9	18	9	0	18	9	0	9	18	0	9

Sociales y Ciudadanas:

2018-1 y 2018-2

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	11	11	11	11	0	0	0	11	0	0	0	11	0	11
B	11	11	0	11	0	11	0	0	0	11	11	0	11	0
C	22	11	11	11	11	22	0	11	11	11	11	0	22	0
D	0	0	0	11	0	0	11	11	11	11	0	11	11	0
E	22	22	11	33	0	11	11	22	11	22	11	22	22	11
F	22	22	11	22	0	11	0	11	0	11	11	11	11	11
G	11	11	11	22	0	0	11	22	11	11	0	22	11	11
H	11	0	11	11	11	11	11	22	22	11	0	11	22	0
I	22	11	22	11	11	11	0	22	11	0	0	11	11	11
J	33	22	22	22	11	22	0	22	11	11	11	11	22	11
K	22	11	22	22	11	11	11	33	22	11	0	22	22	11
L	11	0	11	0	11	11	0	11	11	0	0	0	11	0
M	11	11	0	22	0	11	11	11	11	22	11	11	22	0
N	22	11	11	22	11	22	11	22	22	22	11	11	33	0

2018-2 y 2019-1

FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	11	0	11	0	0	0	11	11	0	11	11	11	0
B	11	22	0	11	11	0	11	11	22	11	11	11	22	0
C	0	11	11	11	22	11	11	11	0	11	22	0	11	22
D	0	11	11	0	11	11	0	0	0	0	11	0	11	11
E	11	33	11	11	22	11	11	11	22	11	22	11	33	11
F	11	11	0	11	22	0	22	11	11	22	11	0	11	11
G	11	33	11	11	22	11	11	11	22	11	22	11	33	11
H	11	11	0	0	11	0	11	0	11	11	0	0	11	0
I	0	11	0	22	11	0	11	22	11	11	22	11	11	11
J	11	11	0	11	22	0	22	11	11	22	11	0	11	11
K	0	22	11	22	22	11	11	22	11	11	33	11	22	22
L	0	11	0	22	11	0	11	22	11	11	22	11	11	11
M	0	11	11	0	11	11	0	0	0	0	11	0	11	11
N	11	22	11	11	33	11	22	11	11	22	22	0	22	22

2019-1 y 2019-2

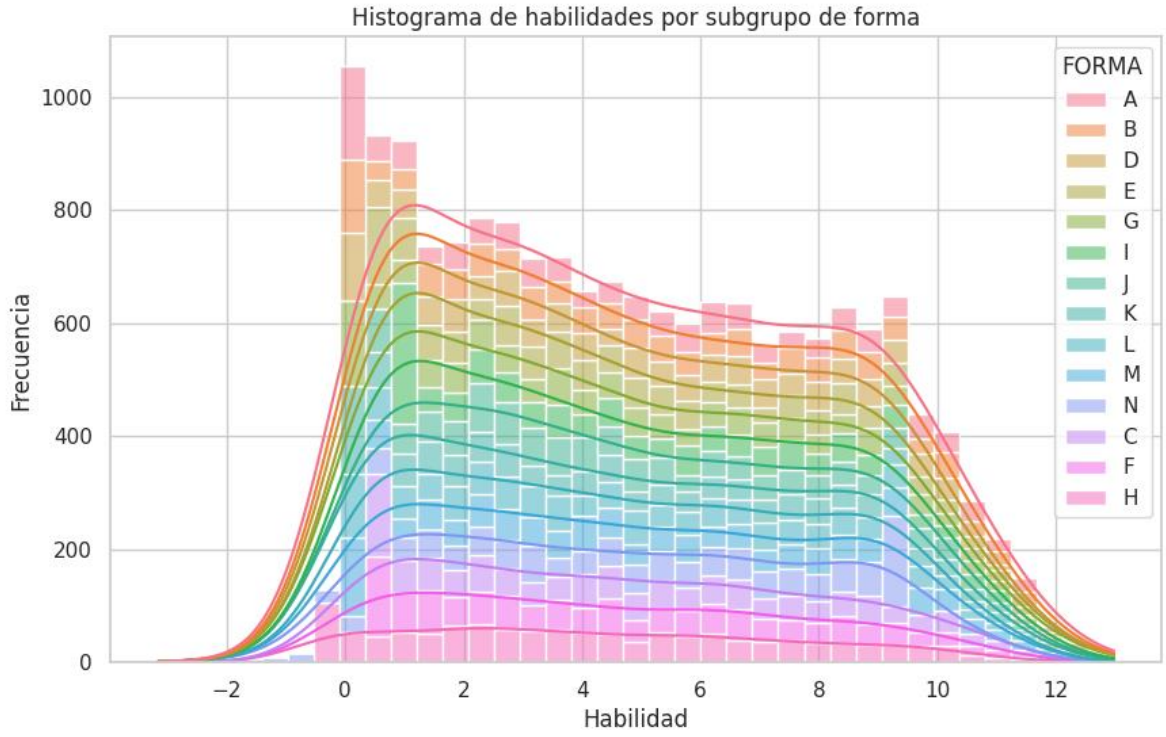
FORMAS	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	22	22	22	0	11	11	11	33	11	11	11	22	22	22
B	0	11	0	0	11	0	0	11	0	11	0	11	11	11
C	22	22	33	11	0	22	11	22	11	11	22	11	11	22
D	11	22	11	0	11	11	0	22	11	11	0	11	22	11
E	11	0	11	0	0	0	11	11	0	0	11	11	0	11
F	11	33	22	11	11	22	0	22	11	22	11	11	22	22
G	0	22	11	11	11	11	0	11	0	22	11	11	11	22
H	22	11	22	0	0	11	11	22	11	0	11	11	11	11
I	11	11	22	11	0	11	11	11	0	11	22	11	0	22
J	11	22	22	11	0	22	0	11	11	11	11	0	11	11
K	0	11	11	11	0	11	0	0	0	11	11	0	0	11
L	11	22	22	11	11	11	11	22	0	22	22	22	11	33
M	11	11	11	0	0	11	0	11	11	0	0	0	11	0
N	11	11	11	0	11	0	11	22	0	11	11	22	11	22

Anexo 2. Distribución de habilidades por prueba y aplicación.

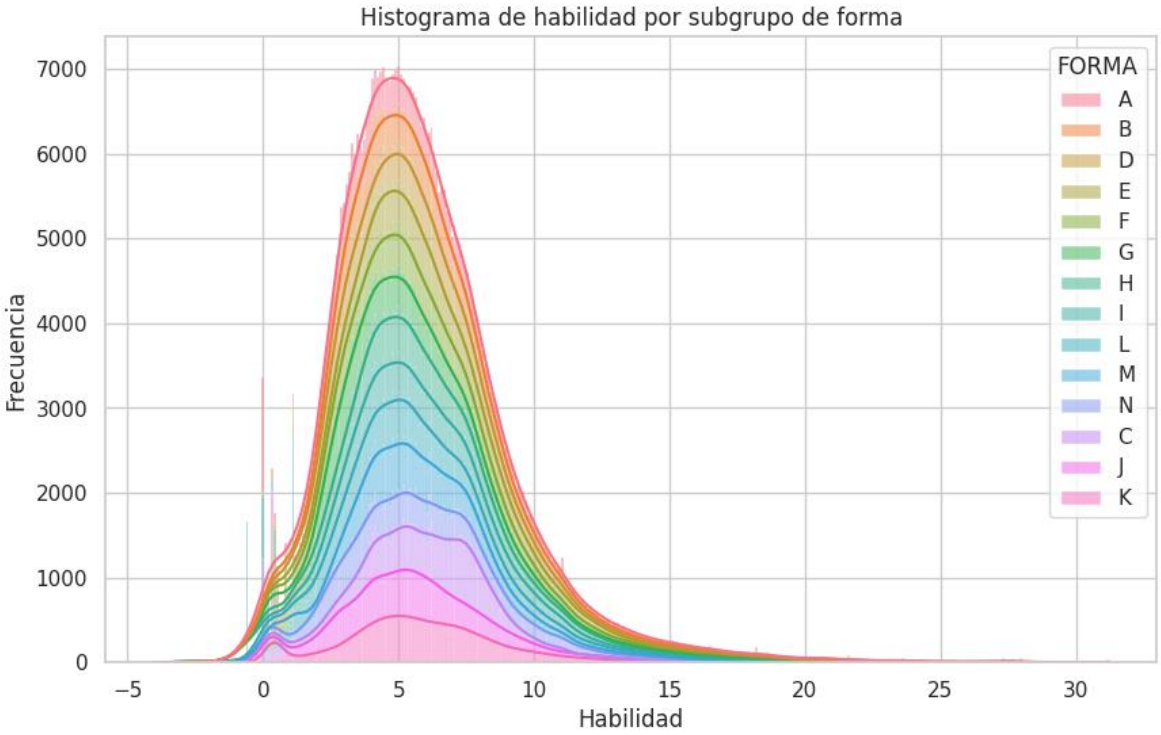
A continuación se presentan los diagramas de densidad e histogramas para todas las pruebas en todas las aplicaciones.

Matemáticas:

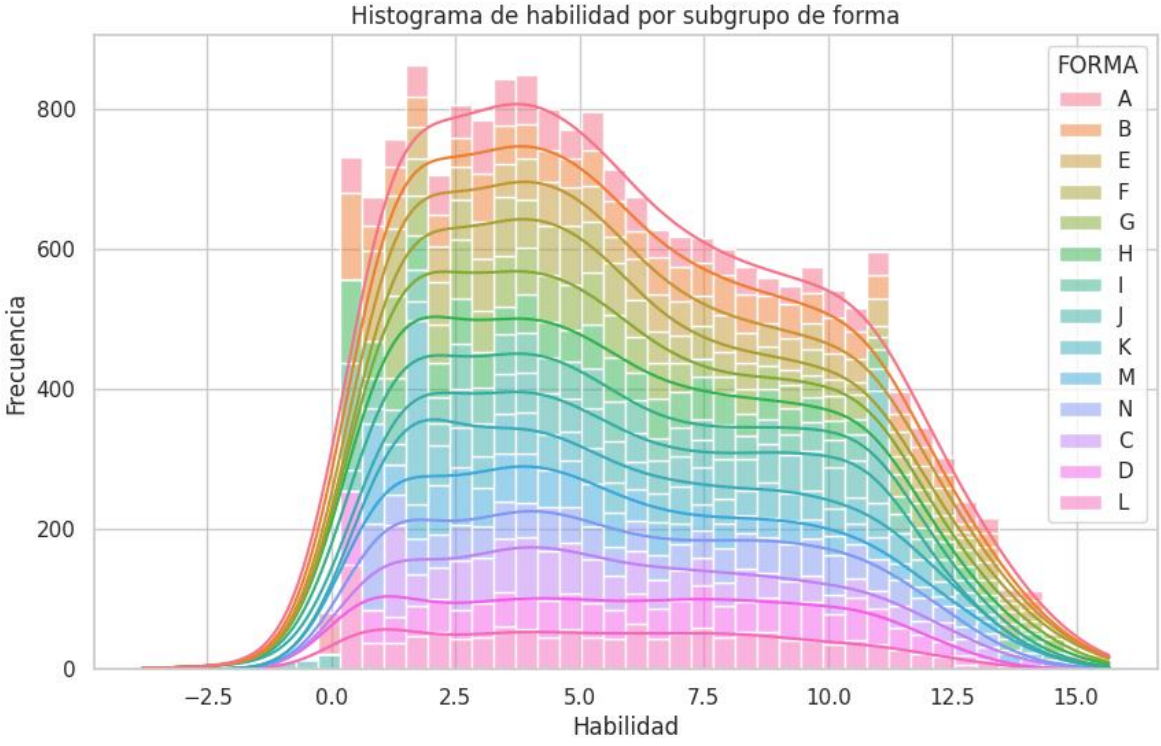
2018-1



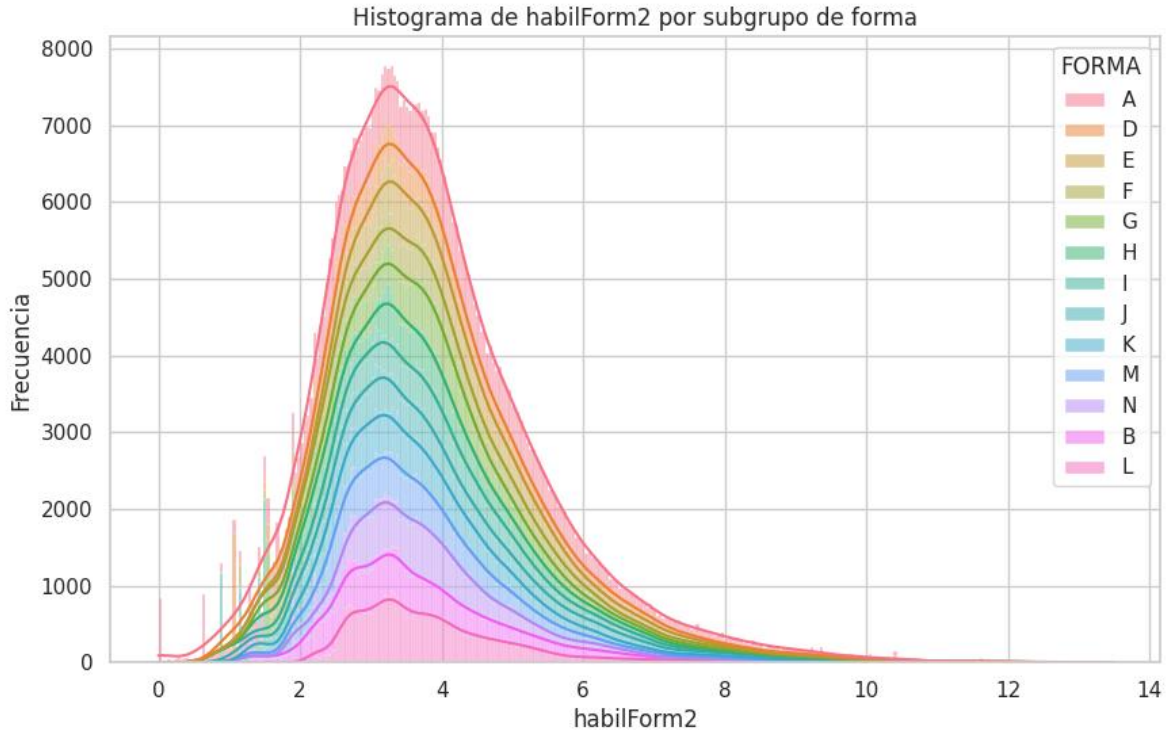
2018-2


Histograma de habilidad por subgrupo de forma
Frecuencia
Habilidad
FORMA
A
B
D
E
F
G
H
I
L
M
N
C
J
K
0
1000
2000
3000
4000
5000
6000
7000
−5
0
5
10
15
20
25
30

2019-1

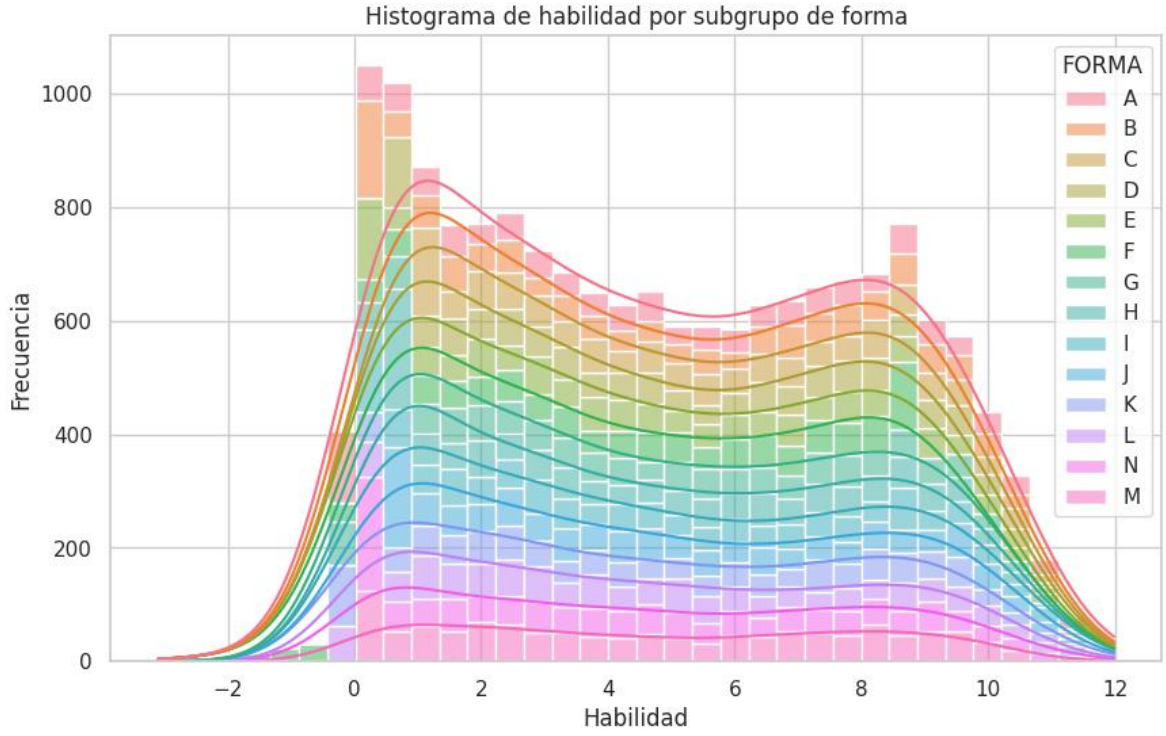

Histograma de habilidad por subgrupo de forma
Frecuencia
Habilidad
FORMA
A
B
E
F
G
H
I
J
K
M
N
C
D
L
0
200
400
600
800
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0

2019-2

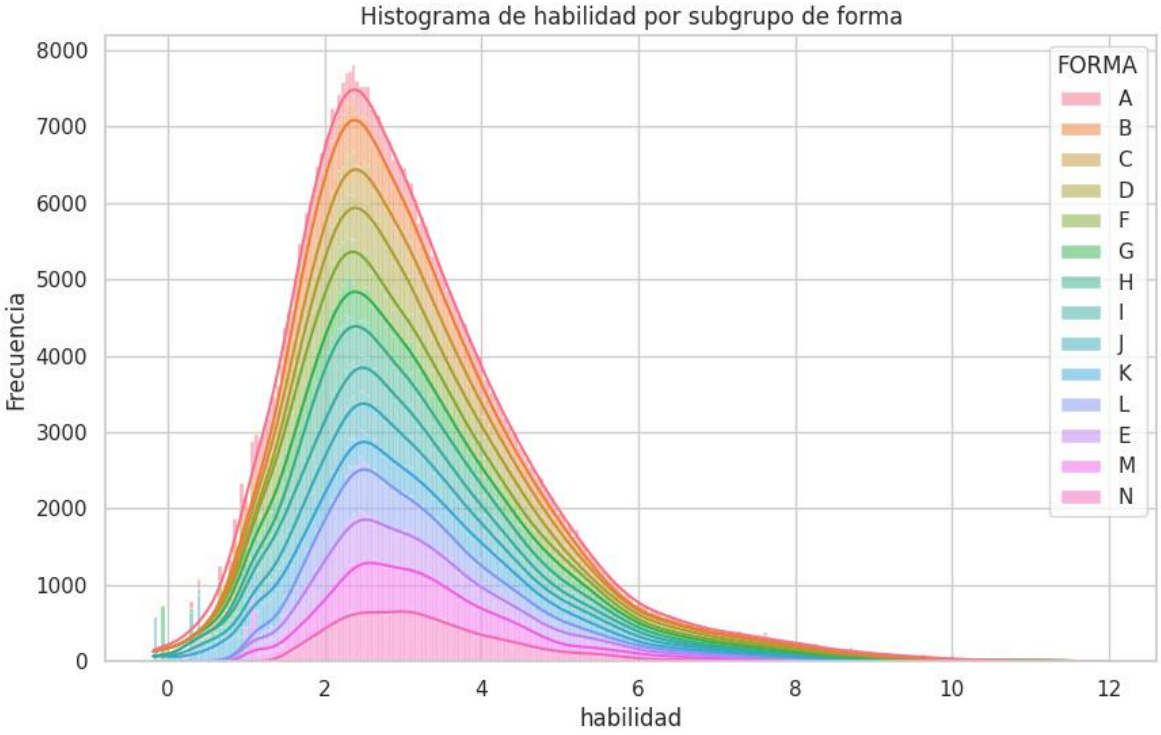


Ciencias Naturales:

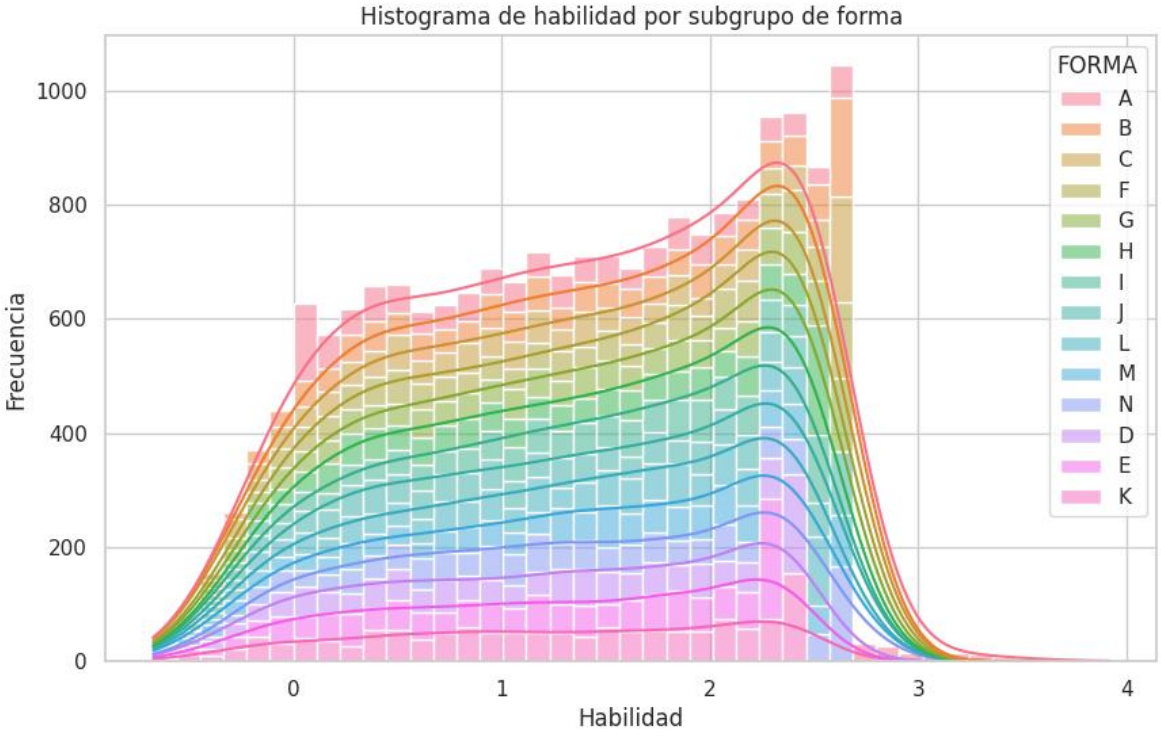
2018-1



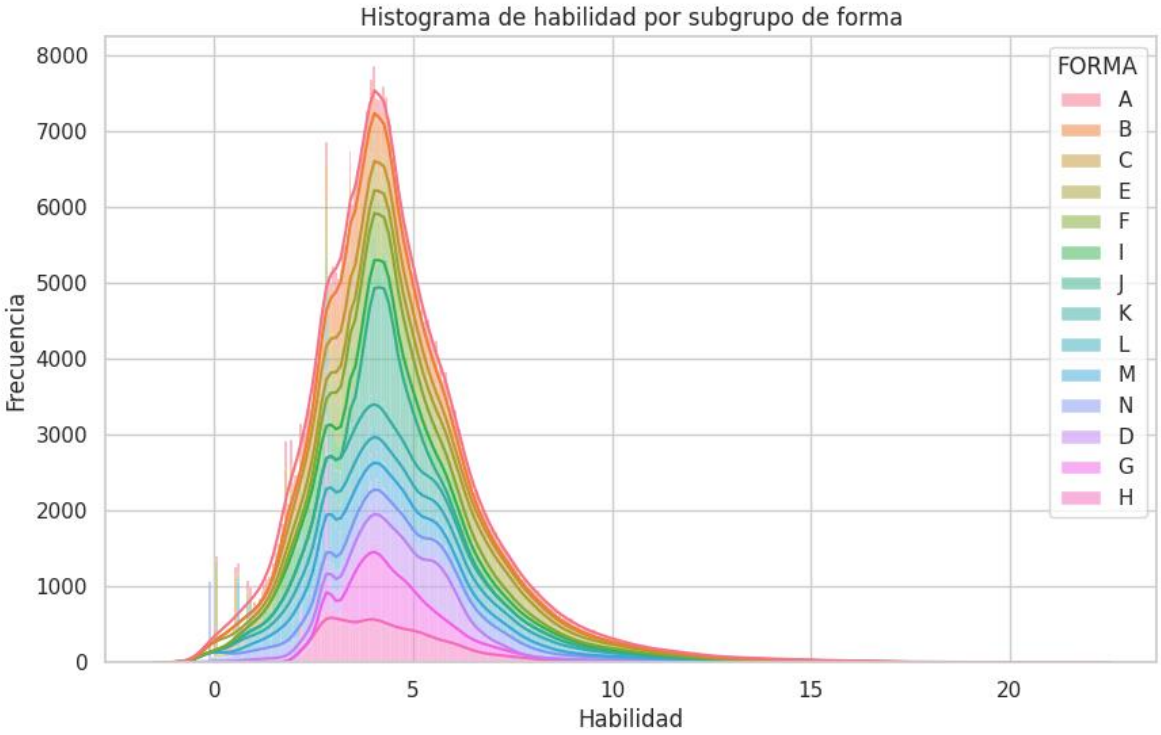
2018-2


Histograma de habilidad por subgrupo de forma
Frecuencia
habilidad
FORMA
A
B
C
D
F
G
H
I
J
K
L
E
M
N
0
1000
2000
3000
4000
5000
6000
7000
8000
0
2
4
6
8
10
12

2019-1

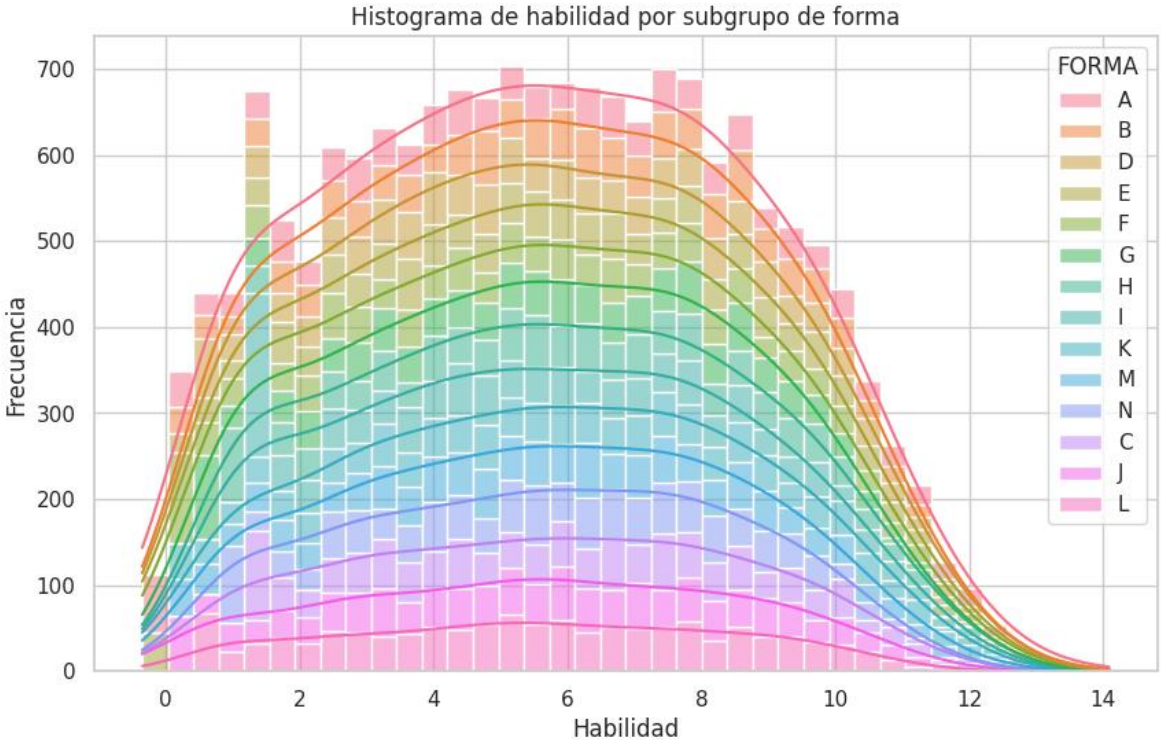

Histograma de habilidad por subgrupo de forma
Frecuencia
Habilidad
FORMA
A
B
C
F
G
H
I
J
L
M
N
D
E
K
0
200
400
600
800
1000
0
1
2
3
4

2019-2

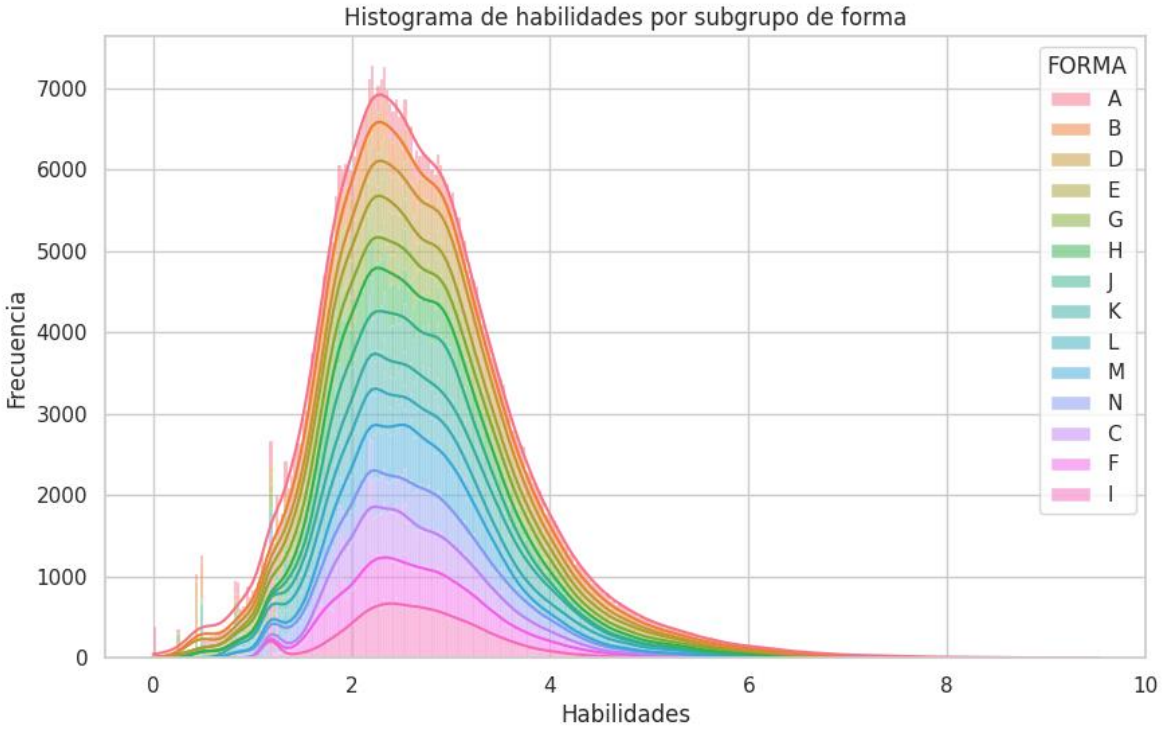


Lectura Crítica:

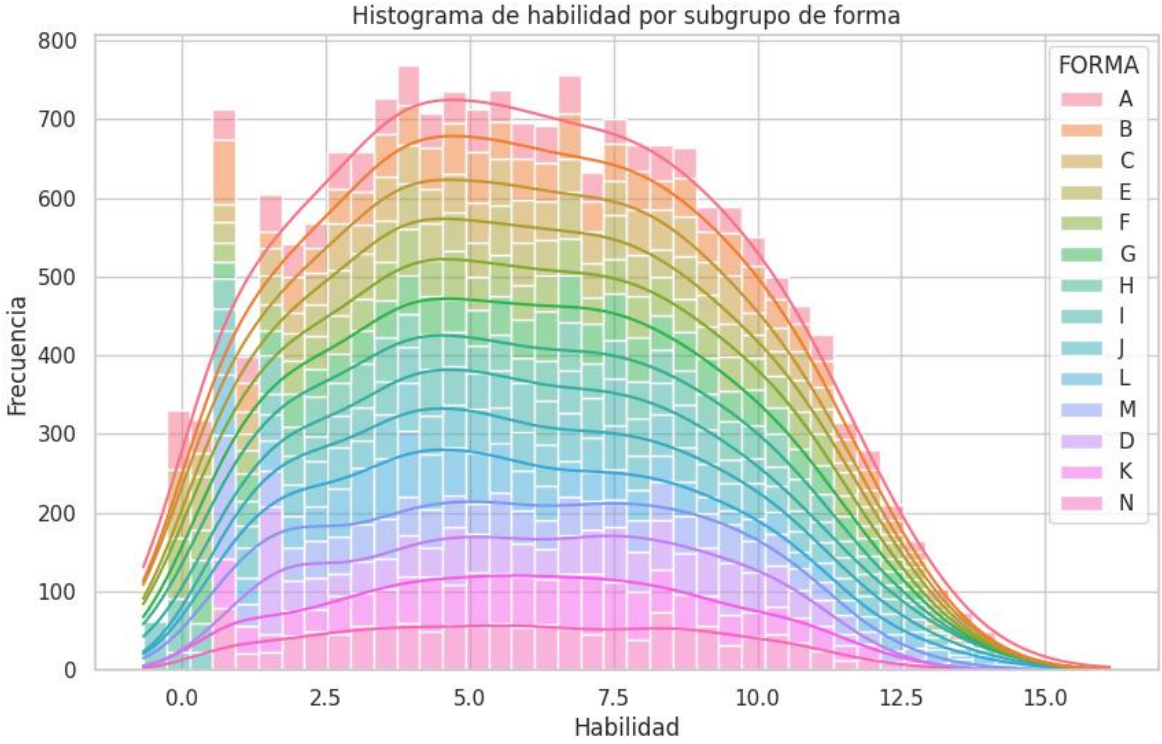
2018-1



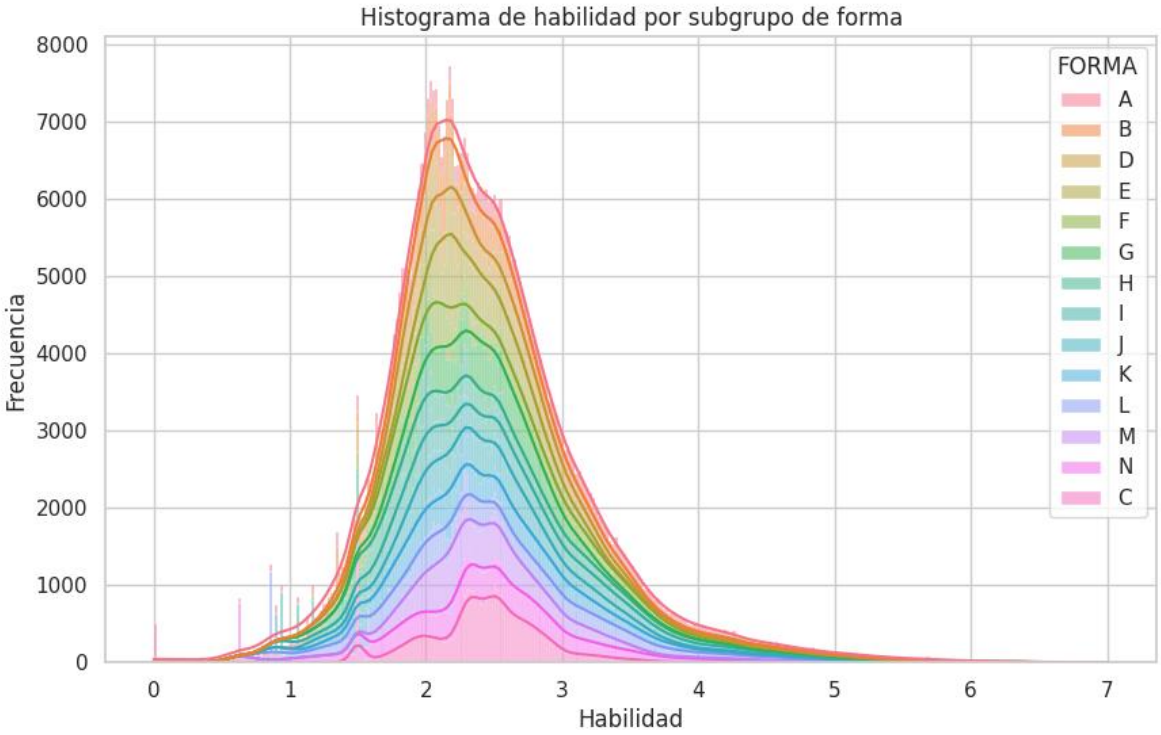
2018-2


Histograma de habilidades por subgrupo de forma
Frecuencia
Habilidades
FORMA
A
B
D
E
G
H
J
K
L
M
N
C
F
I
0
1000
2000
3000
4000
5000
6000
7000
0
2
4
6
8
10

2019-1

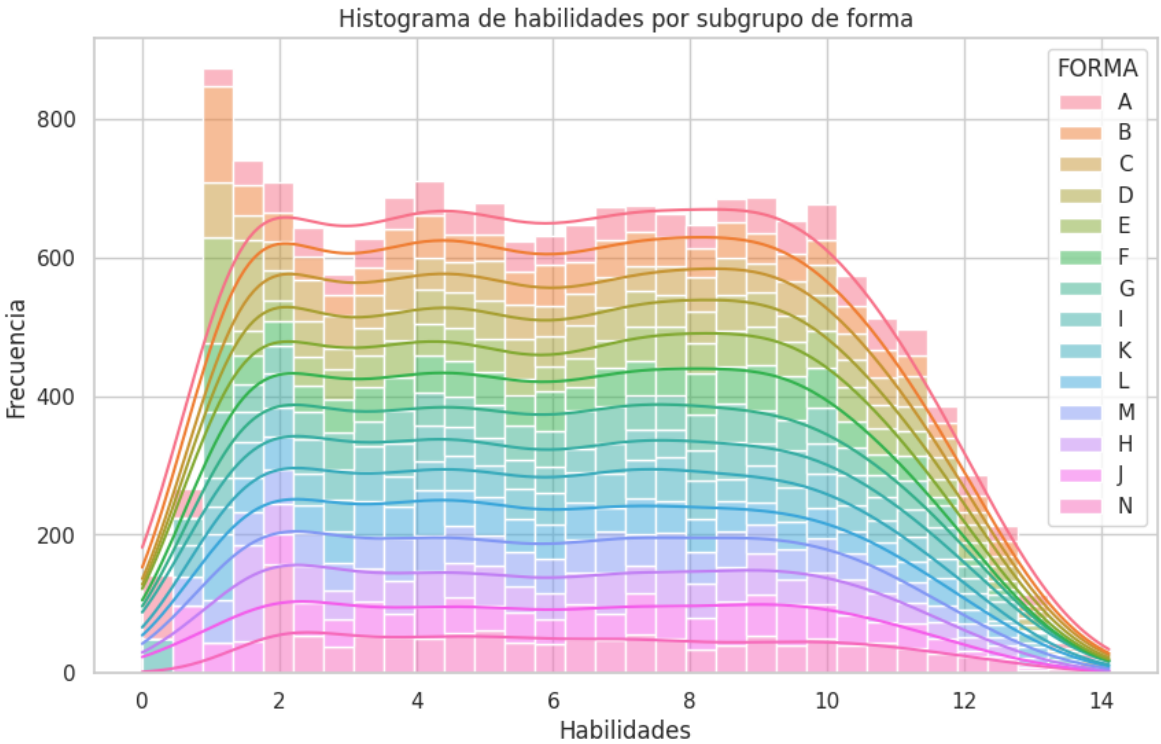

Histograma de habilidad por subgrupo de forma
Frecuencia
Habilidad
FORMA
A
B
C
E
F
G
H
I
J
L
M
D
K
N
0
100
200
300
400
500
600
700
800
0.0
2.5
5.0
7.5
10.0
12.5
15.0

2019-2

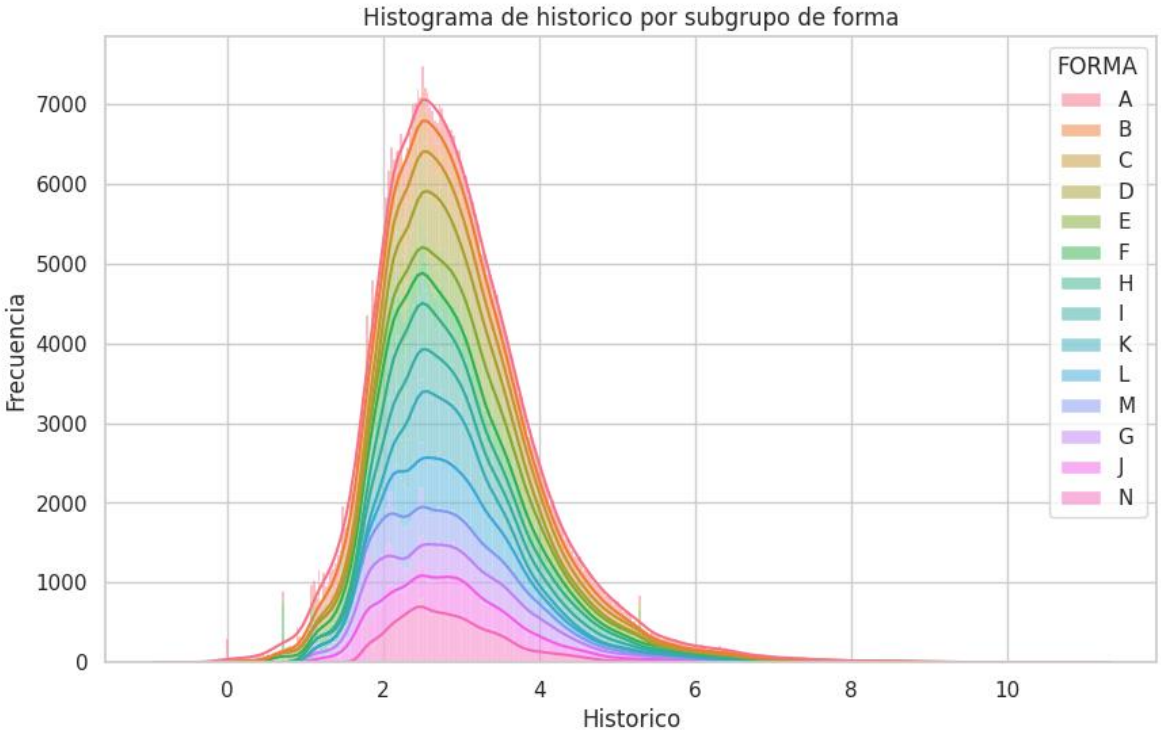


Sociales y Ciudadanas:

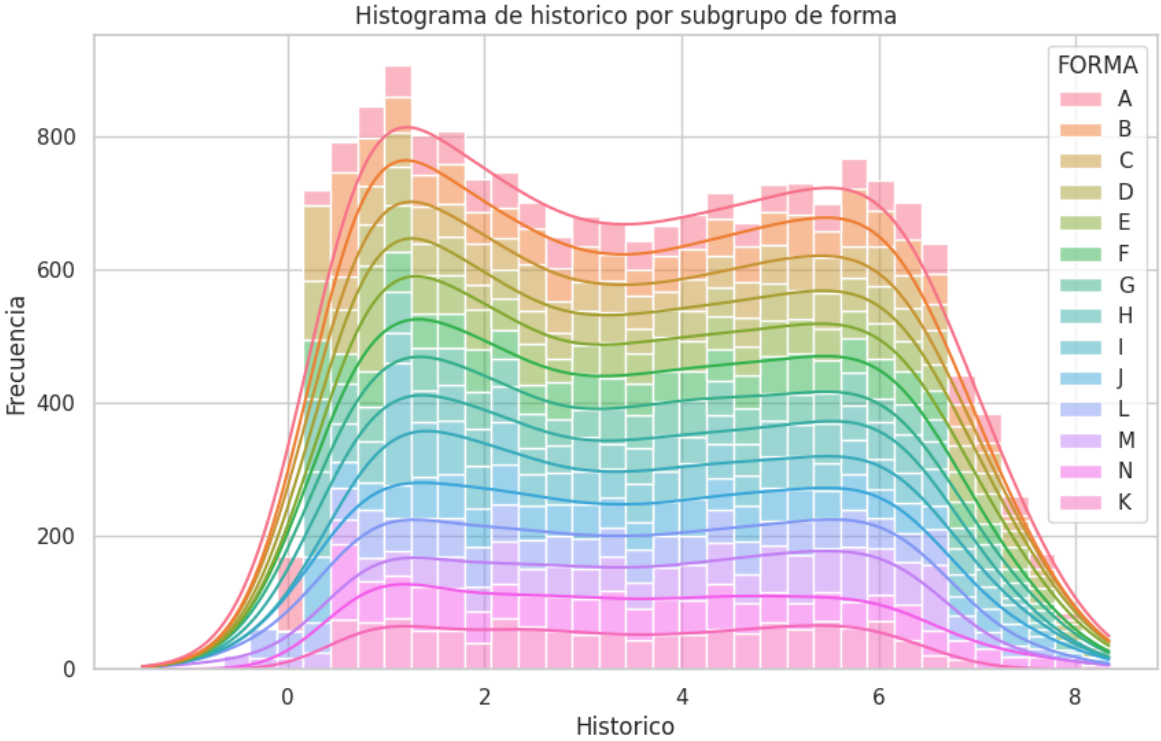
2018-1



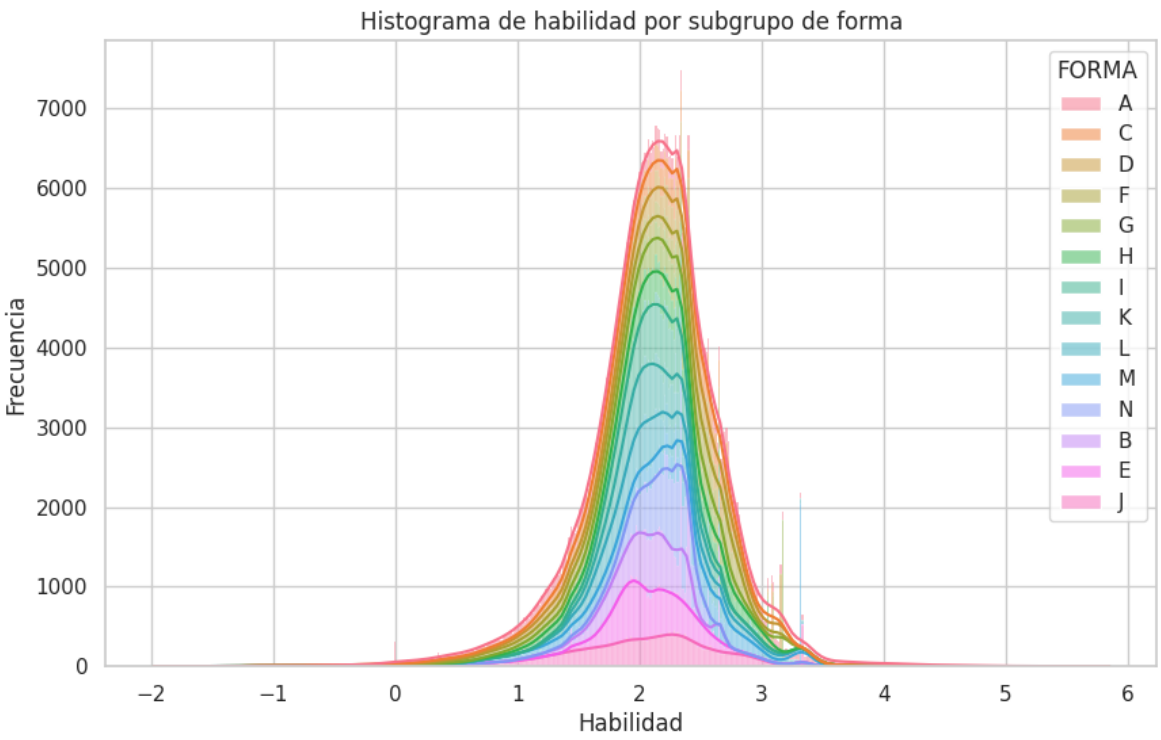
2018-2


Histograma de historico por subgrupo de forma
Frecuencia
Historico
FORMA
A
B
C
D
E
F
H
I
K
L
M
G
J
N
0
1000
2000
3000
4000
5000
6000
7000
0
2
4
6
8
10

2019-1


Histograma de historico por subgrupo de forma
Frecuencia
Historico
FORMA
A
B
C
D
E
F
G
H
I
J
L
M
N
K
0
200
400
600
800
0
2
4
6
8

2019-2


Histograma de habilidad por subgrupo de forma
FORMA
A
C
D
F
G
H
I
K
L
M
N
B
E
J
Frecuencia
0
1000
2000
3000
4000
5000
6000
7000
−2
−1
0
1
2
3
4
5
6
Habilidad

Anexo 3. Correlación entre porcentaje de respuestas correctas y habilidad generada por los AE.

Matemáticas:

APLICACIÓN	FORMA	CORRELACIÓN	APLICACIÓN	FORMA	CORRELACIÓN
2018-1	A	0.980463	2019-1	A	0.976977
	B	0.984538		B	0.983178
	C	0.984750		C	0.963264
	D	0.985836		D	0.986476
	E	0.974603		E	0.971878
	F	0.987717		F	0.949854
	G	0.978747		G	0.968485
	H	0.983864		H	0.968907
	I	0.973390		I	0.977693
	J	0.975891		J	0.984535
	K	0.981883		K	0.973218
	L	0.987086		L	0.985129
	M	0.983443		M	0.957265
	N	0.987608		N	0.983602
2018-2	A	0.908677	2019-2	A	0.961557
	B	0.908805		B	0.934883
	C	0.921340		C	0,949675
	D	0.956186		D	0.943371
	E	0.921698		E	0.967007
	F	0.946437		F	0.950531
	G	0.946091		G	0.965709
	H	0.933567		H	0.925893
	I	0.944257		I	0.943642
	J	0.962555		J	0.956506
	K	0.959536		K	0.954362
	L	0.941373		L	0.938709
	M	0.904323		M	0.948352
	N	0.955655		N	0.967131

Ciencias Naturales:

APLICACIÓN	FORMA	CORRELACIÓN	APLICACIÓN	FORMA	CORRELACIÓN
2018-1	A	0.982085	2019-1	A	0.984301
	B	0.984280		B	0.984565
	C	0.983820		C	0.983050
	D	0.984941		D	0.976869
	E	0.984904		E	0.980503
	F	0.984400		F	0.988649
	G	0.991817		G	0.979370
	H	0.973509		H	0.981080
	I	0.979273		I	0.981409
	J	0.982334		J	0.980567
	K	0.988051		K	0.973666
	L	0.977171		L	0.981353
	M	0.987847		M	0.971795
	N	0.982162		N	0.982817
2018-2	A	0.956848	2019-2	A	0.946032
	B	0.948067		B	0.954726
	C	0.948533		C	0.944709
	D	0.950336		D	0.936342
	E	0.951391		E	0.970114
	F	0.956190		F	0.954702
	G	0.958728		G	0.939683
	H	0.963862		H	0.948779
	I	0.958637		I	0.930389
	J	0.933420		J	0.955332
	K	0.970378		K	0.938033
	L	0.954461		L	0.960043
	M	0.965506		M	0.942857
	N	0.941330		N	0.972986

Lectura Crítica:

APLICACIÓN	FORMA	CORRELACIÓN	APLICACIÓN	FORMA	CORRELACIÓN
2018-1	A	0.985295	2019-1	A	0.972285
	B	0.987557		B	0.982699
	C	0.982493		C	0.980311
	D	0.985970		D	0.971913
	E	0.986950		E	0.983036
	F	0.983670		F	0.986557
	G	0.984099		G	0.982264
	H	0.984579		H	0.987766
	I	0.968797		I	0.972592
	J	0.987316		J	0.975027
	K	0.980034		K	0.975606
	L	0.984412		L	0.957641
	M	0.983833		M	0.986065
	N	0.982925		N	0.978253
2018-2	A	0.940722	2019-2	A	0.947168
	B	0.934813		B	0.857119
	C	0.951723		C	0.931817
	D	0.960656		D	0.925528
	E	0.940215		E	0.918845
	F	0.946042		F	0.913447
	G	0.956326		G	0.933334
	H	0.942334		H	0.903706
	I	0.952536		I	0.934656
	J	0.952467		J	0.912327
	K	0.962246		K	0.924736
	L	0.957789		L	0.939153
	M	0.963792		M	0.919798
	N	0.962092		N	0.950476

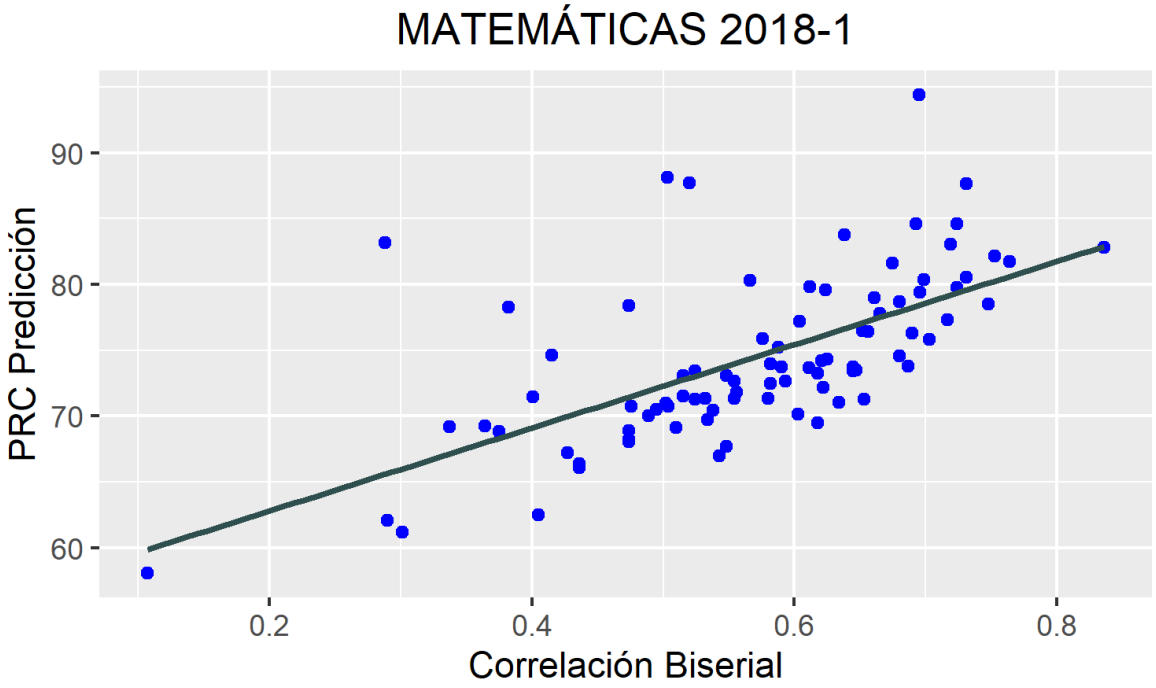
Sociales y Ciudadanas:

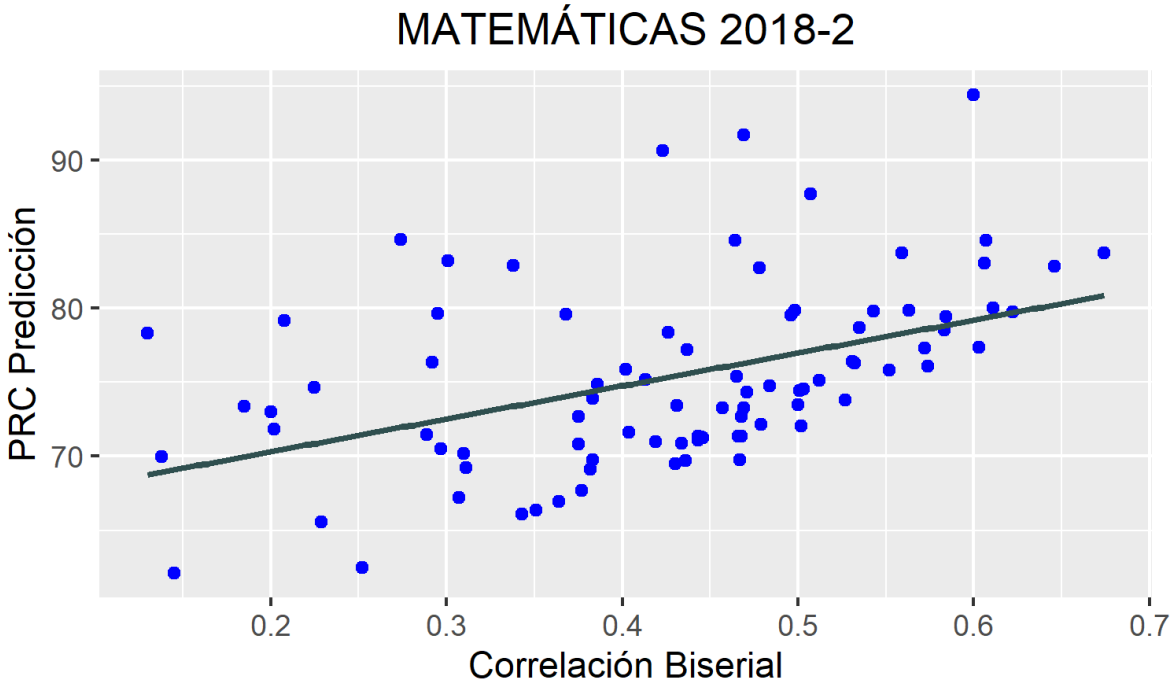
APLICACIÓN	FORMA	CORRELACIÓN	APLICACIÓN	FORMA	CORRELACIÓN
2018-1	A	0.986501	2019-1	A	0.988646
	B	0.982292		B	0.989427
	C	0.979920		C	0.989933
	D	0.981656		D	0.986863
	E	0.981105		E	0.978910
	F	0.985984		F	0.990014
	G	0.984673		G	0.985328
	H	0.983443		H	0.988238
	I	0.987867		I	0.981890
	J	0.986303		J	0.980778
	K	0.983462		K	0.990122
	L	0.984252		L	0.992540
	M	0.980403		M	0.985370
	N	0.977198		N	0.983257
2018-2	A	0.972221	2019-2	A	0.941079
	B	0.937293		B	0.946473
	C	0.930146		C	0.964311
	D	0.948160		D	0.948855
	E	0.957352		E	0.935801
	F	0.949940		F	0.965780
	G	0.963853		G	0.951213
	H	0.915970		H	0.934566
	I	0.947428		I	0.952796
	J	0.954787		J	0.956976
	K	0.856421		K	0.917444
	L	0.932874		L	0.848870
	M	0.952478		M	0.933573
	N	0.927069		N	0.957468

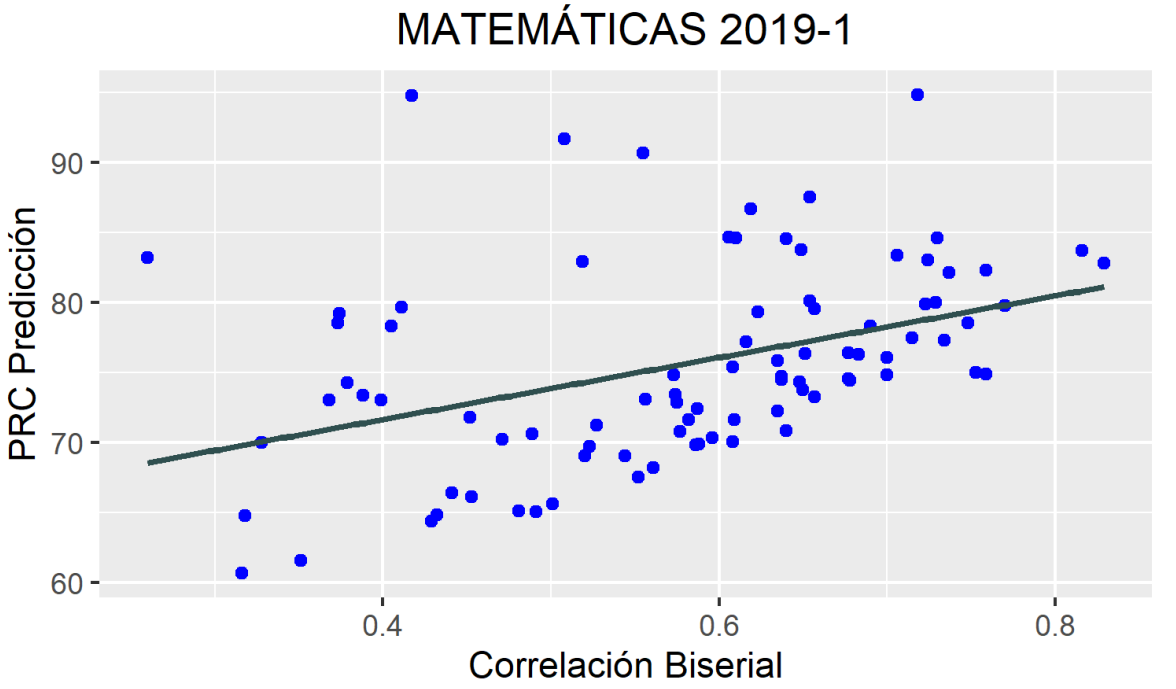
Anexo 4. Relación entre la correlación biserial y porcentaje de aciertos de predicción.

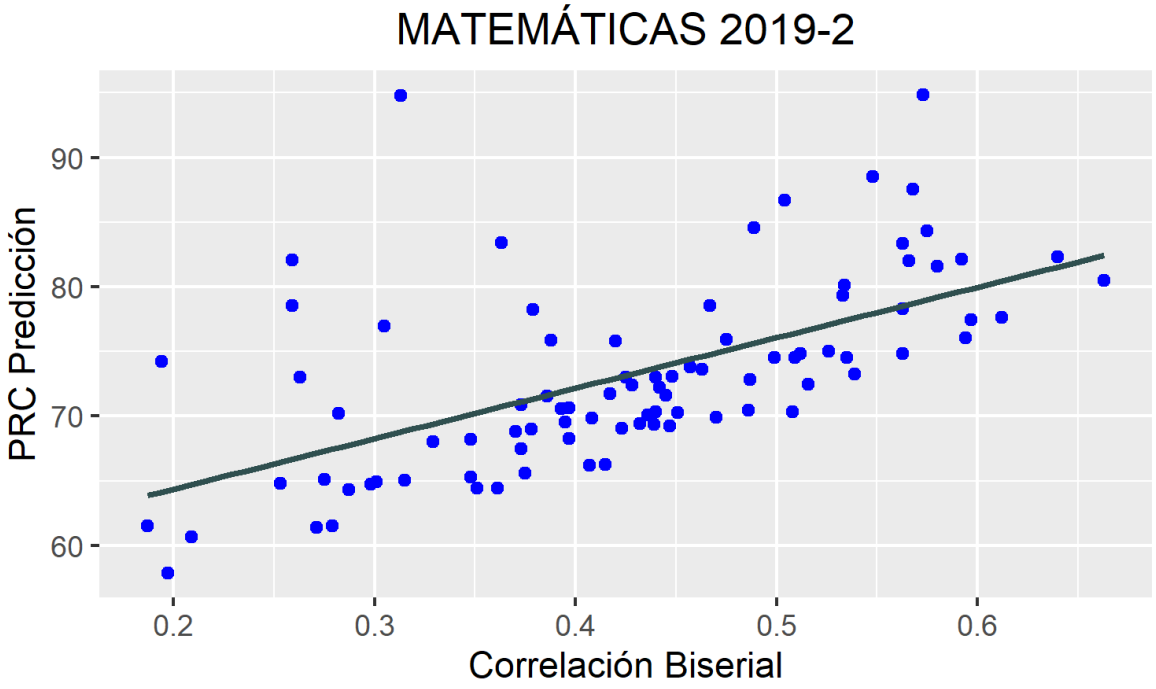
A continuación se presentan los gráficos que muestran la relación entre la correlación biserial de los ítems y el porcentaje de aciertos de predicción del AE a nivel de ítem.

Matemáticas:

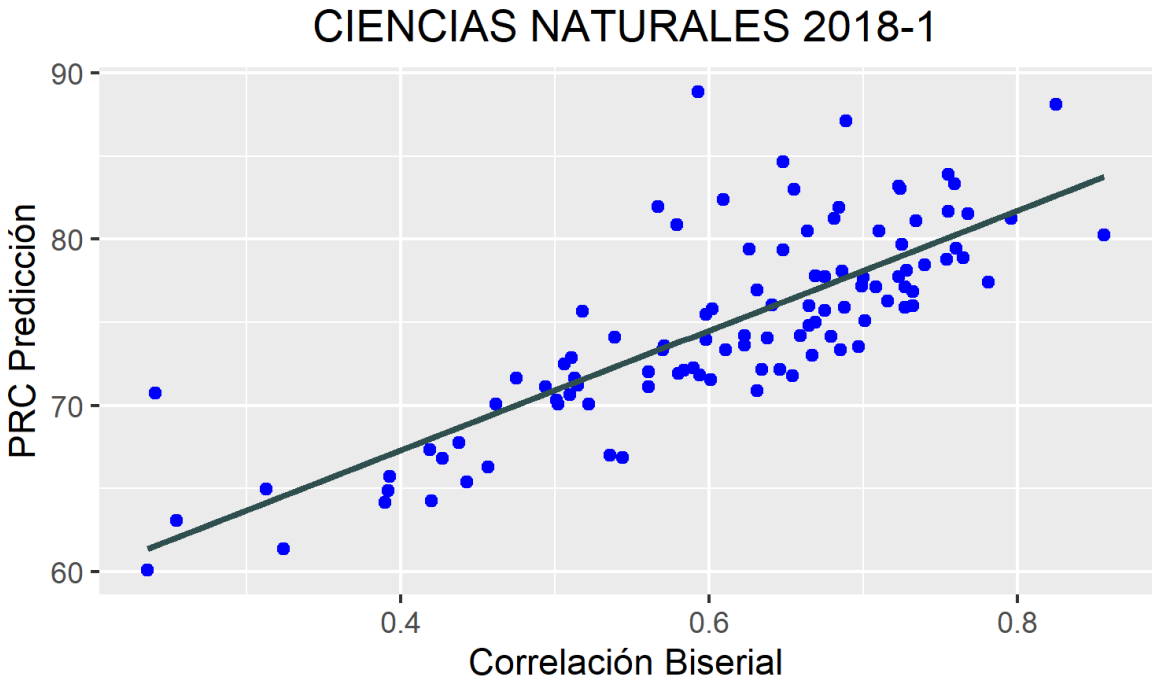


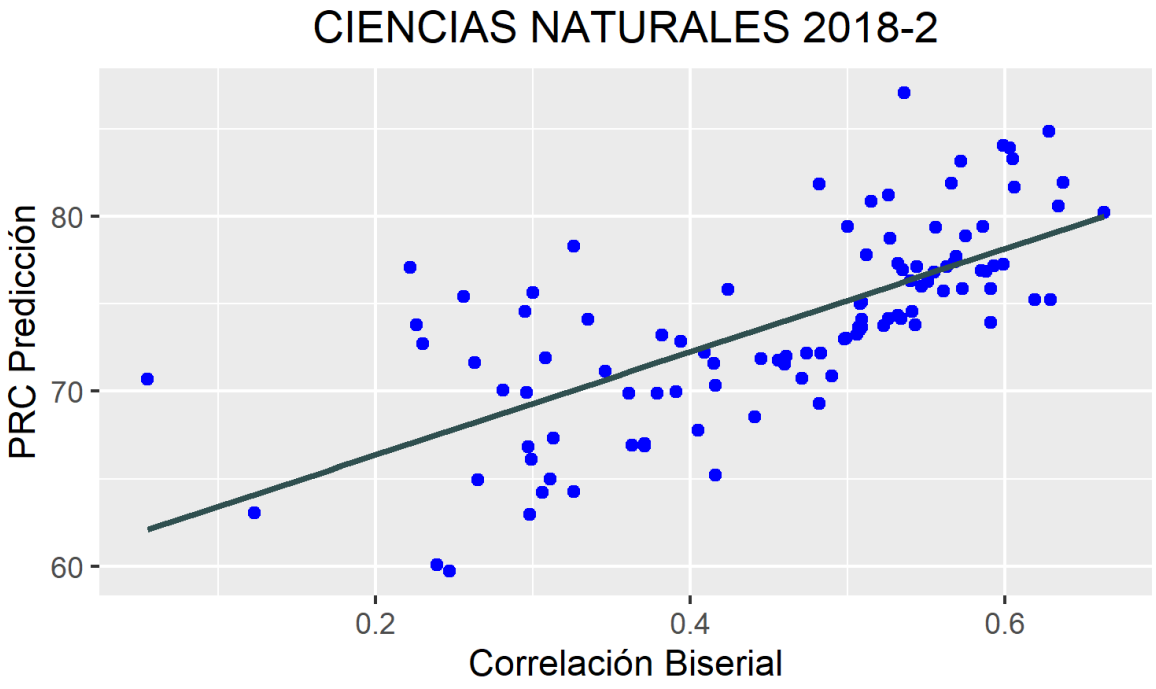

MATEMÁTICAS 2018-2
PRC Predicción
90
80
70
0.2
0.3
0.4
0.5
0.6
0.7
Correlación Biserial

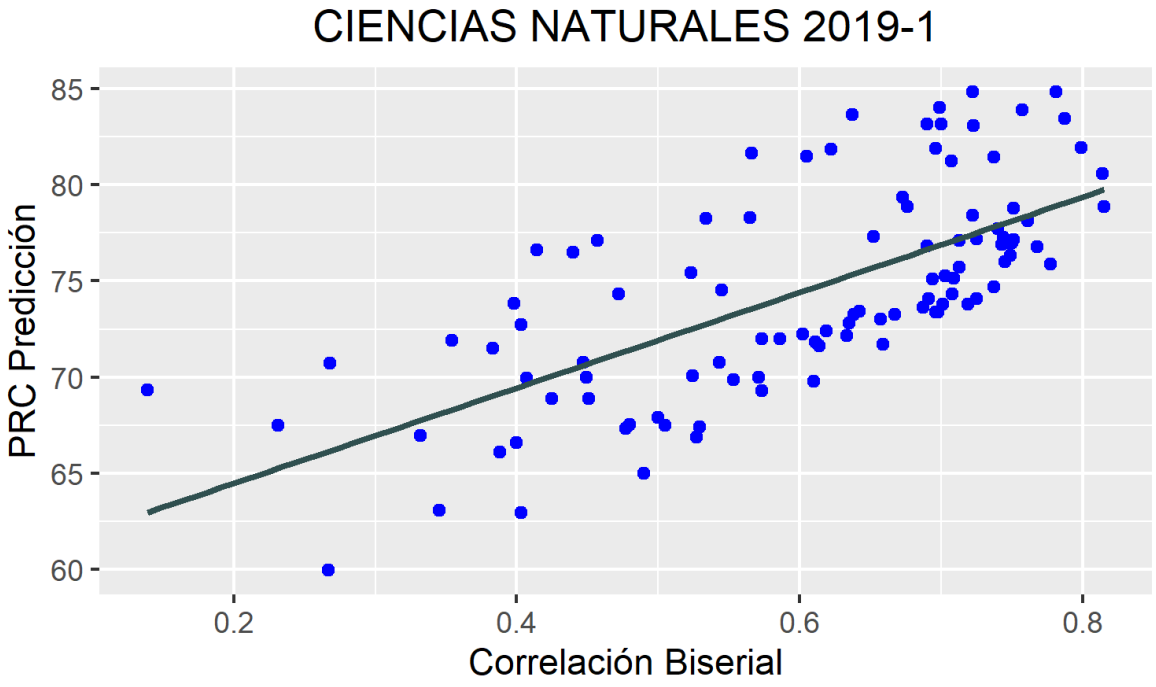

MATEMÁTICAS 2019-1
PRC Predicción
90
80
70
60
0.4
0.6
0.8
Correlación Biserial

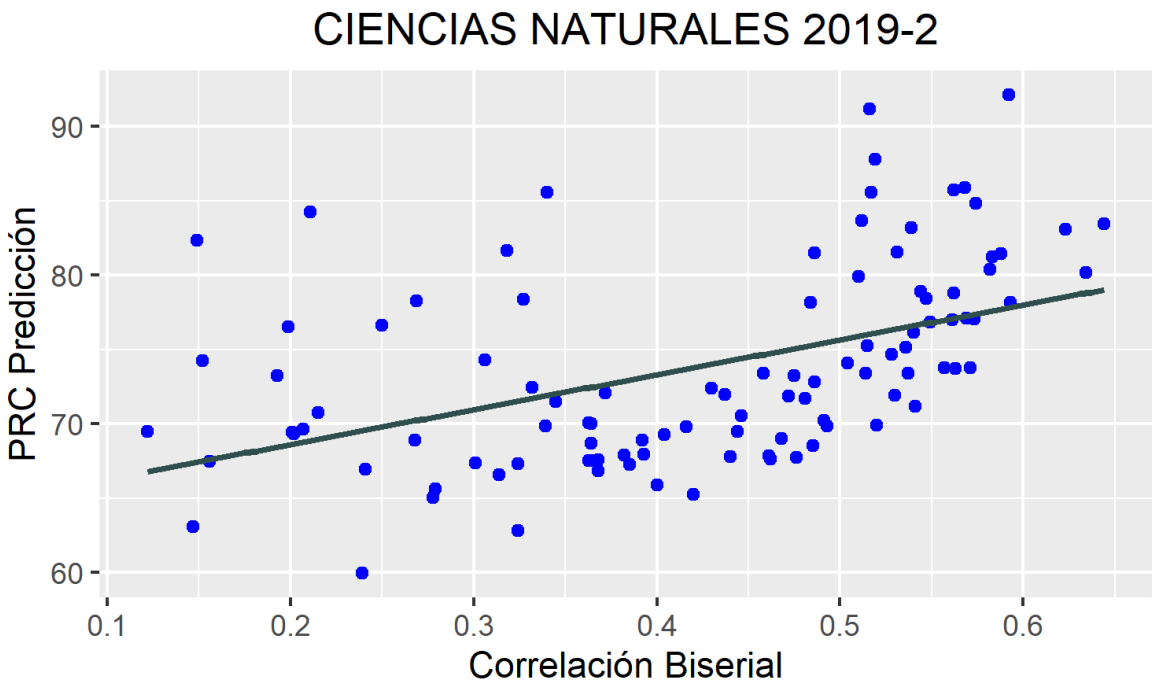


Ciencias Naturales:

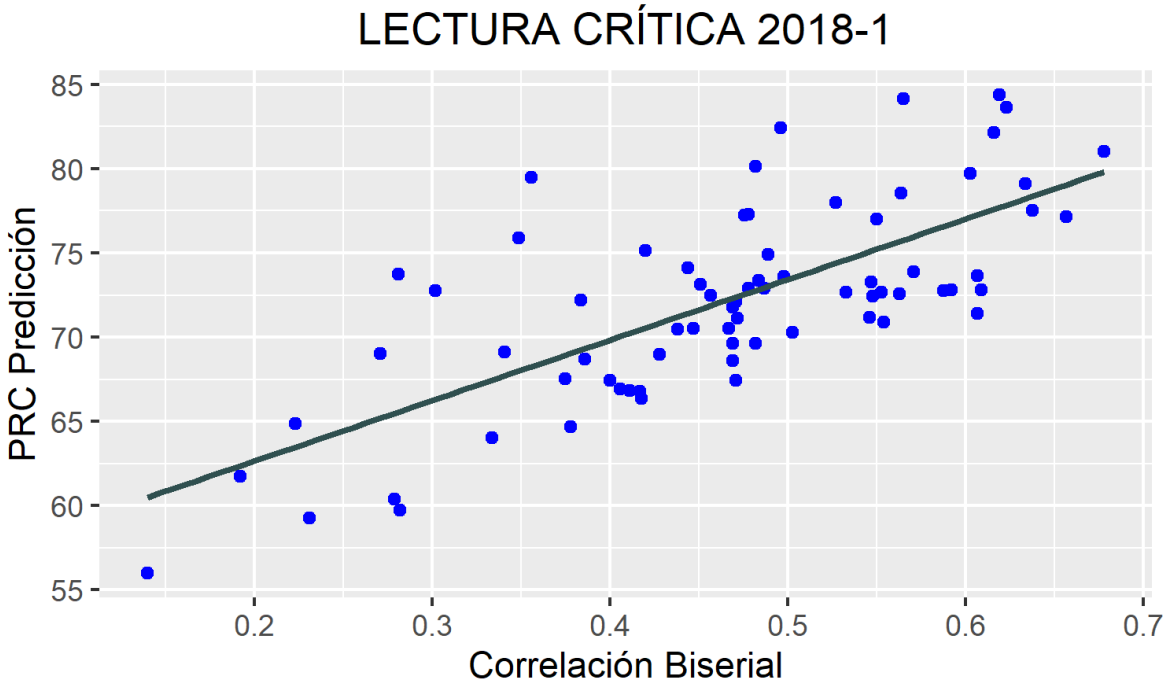


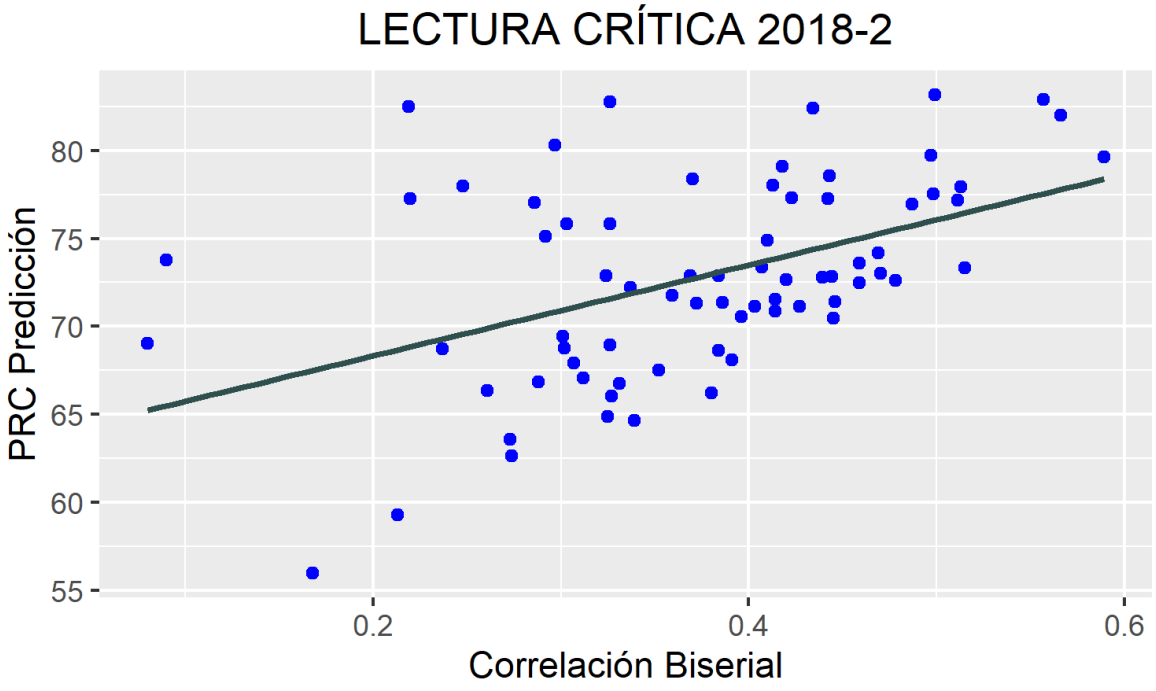

CIENCIAS NATURALES 2018-2
PRC Predicción
80
70
60
0.2
0.4
0.6
Correlación Biserial

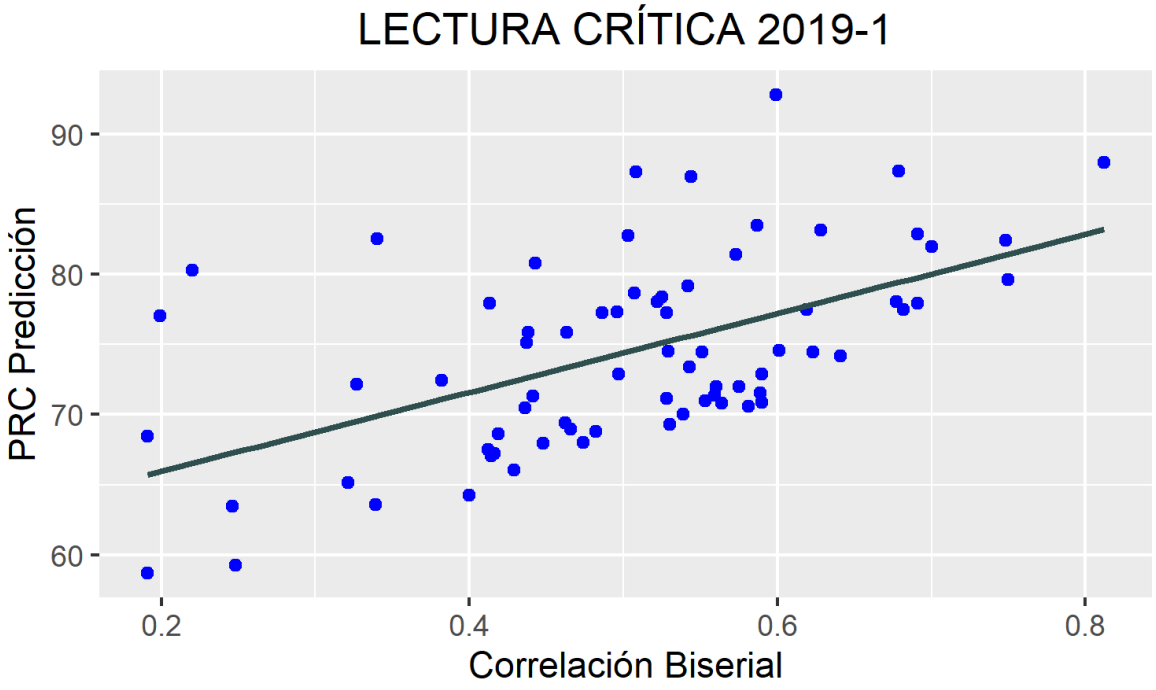

CIENCIAS NATURALES 2019-1
PRC Predicción
85
80
75
70
65
60
0.2
0.4
0.6
0.8
Correlación Biserial

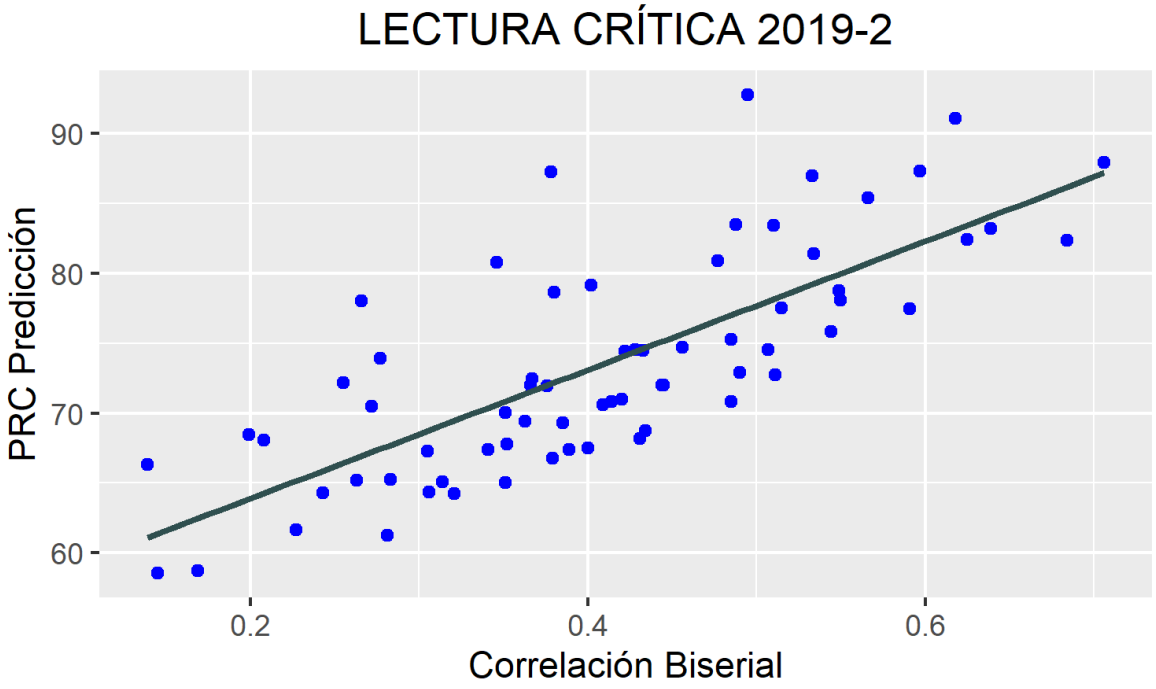


Lectura Crítica:

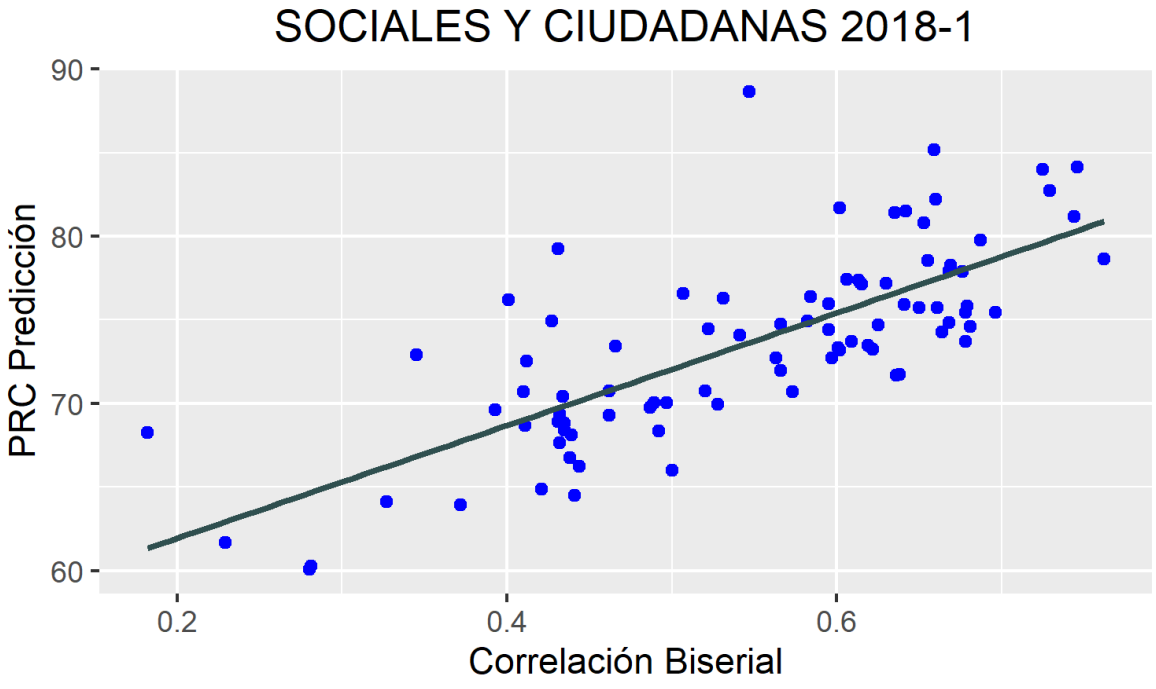


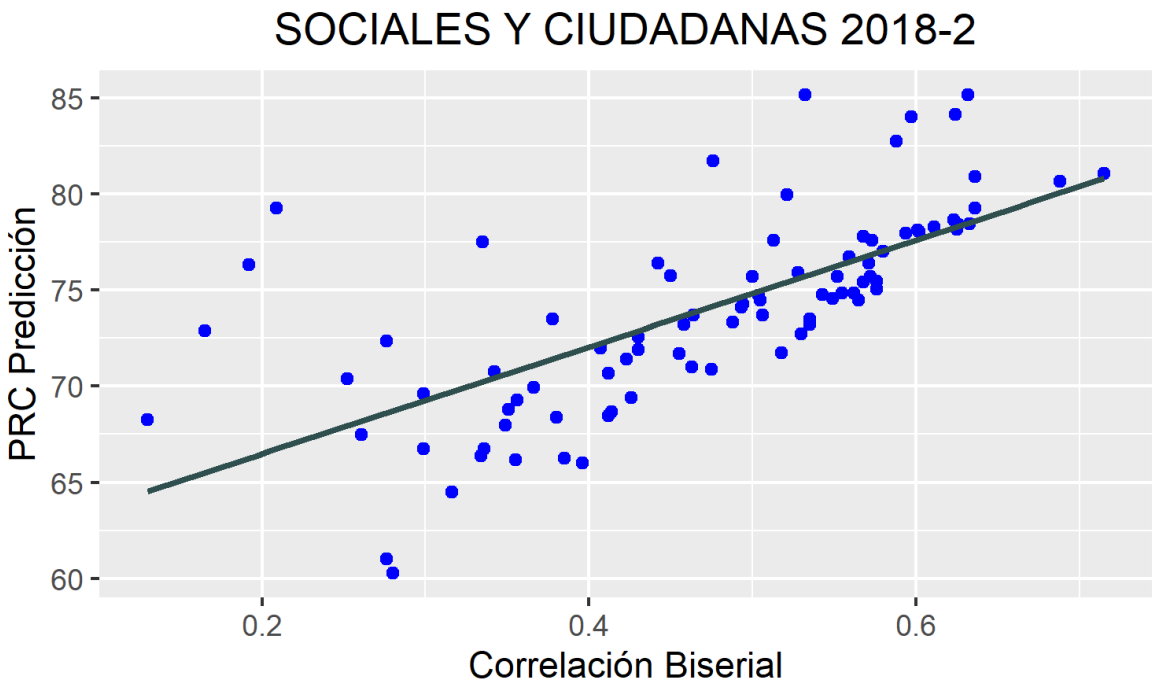

LECTURA CRÍTICA 2018-2
PRC Predicción
80
75
70
65
60
55
0.2
0.4
0.6
Correlación Biserial

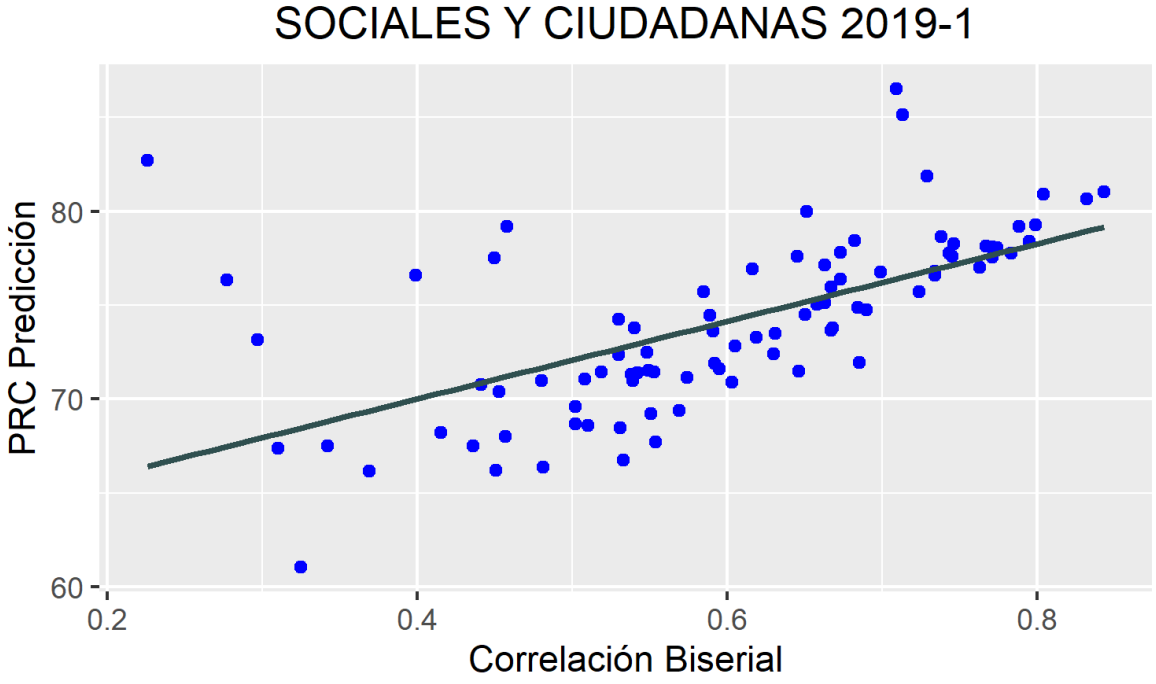

LECTURA CRÍTICA 2019-1
PRC Predicción
90
80
70
60
0.2
0.4
0.6
0.8
Correlación Biserial

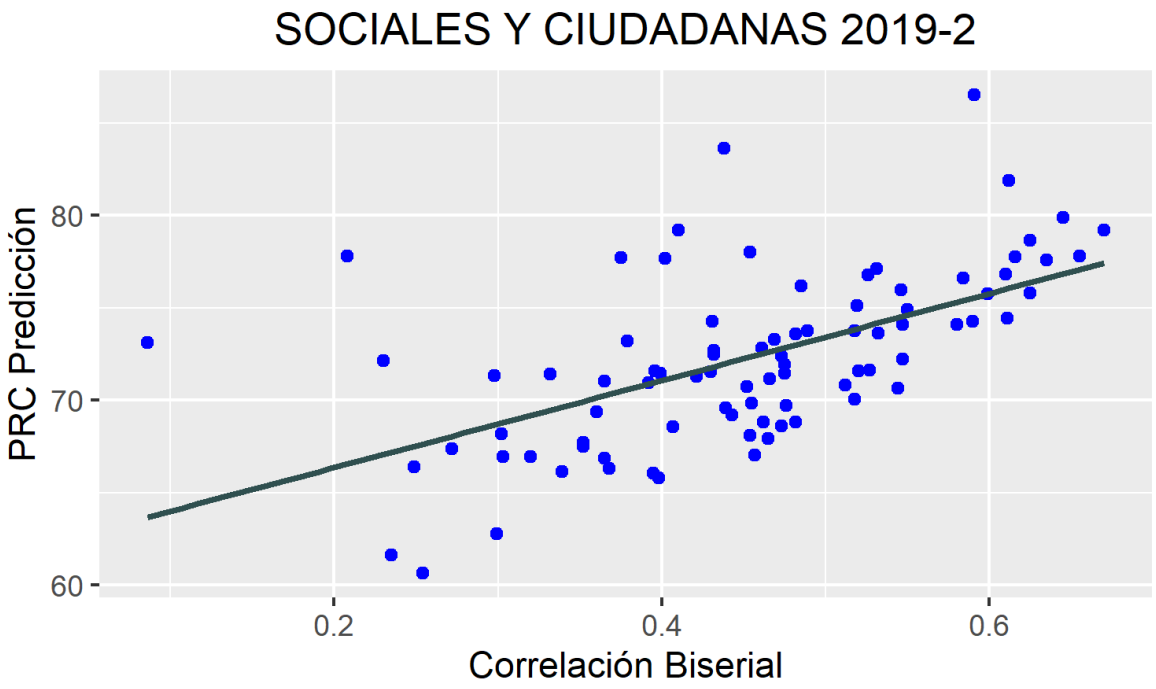


Sociales y ciudadanas:




SOCIALES Y CIUDADANAS 2018-2
PRC Predicción
85
80
75
70
65
60
0.2
0.4
0.6
Correlación Biserial


SOCIALES Y CIUDADANAS 2019-1
PRC Predicción
80
70
60
0.2
0.4
0.6
0.8
Correlación Biserial


SOCIALES Y CIUDADANAS 2019-2
PRC Predicción
80
70
60
0.2
0.4
0.6
Correlación Biserial

Bibliografía

American Educational Research Association -AERA, American Psychological Association - APA, & National Council on Measurement in Education –NCME (2018). Estándares para pruebas educativas y psicológicas. American Educational Research Association.

Amin, A. (2020), ``A Face Recognition System Based on Deep Learning (FRDLS) to Support the Entry and Supervision Procedures on Electronic Exams``. International Journal of Intelligent Computing and Information Sciences, 20(1). https://doi.org/10.21608/ijicis.2020.23149.1015

Basheer, Imad & Hajmeer, M.N.. (2001). Artificial Neural Networks: Fundamentals, Computing, Design, and Application. Journal of microbiological methods. 43. 3-31.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer New York.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A Multigroup Item Response Theory Analysis of the Psychopathy Checklist-Revised. Psychological assessment, 16(2), 155.

Bozak, A., & Aybek, E. C. (2020). Comparison of Artificial Neural Networks and Logistic Regression Analysis in PISA Science Literacy Success Prediction. International Journal of Contemporary Educational Research. https://doi.org/10.33200/ijcer.693081

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical methods, 6(9), 2812-2831.

Converse, G., Curi, M., & Oliveira, S. (2019). Autoencoders for educational assessment. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11626 LNAI. https://doi.org/10.1007/978-3-030-23207-8_8

Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. Machine Learning, 110(6). https://doi.org/10.1007/s10994-021-06005-7

Cortada de Kohan, N. (2005). Posibilidad de integración de las teorías cognitivas y la psicometría moderna. Interdisciplinaria, 22(1), 29-58.

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019). Interpretable Variational Autoencoders for Cognitive Models. Proceedings of the International Joint Conference on Neural Networks, 2019-July. https://doi.org/10.1109/IJCNN.2019.8852333

Developers, T. (2022). TensorFlow. Zenodo.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22(4), 249-262.

Dunn, T., Howlett, S. E., Stanojevic, S., Shehzad, A., Stanley, J., & Rockwood, K. (2022). Patterns of Symptom Tracking by Caregivers and Patients with Dementia and Mild Cognitive Impairment: Cross-sectional Study. Journal of Medical Internet Research, 24(1). https://doi.org/10.2196/29219

Eignor, D. R. (2006). Test Equating, Scaling, and Linking Methods and Practices.

El-Alfy, E. S. M., & Abdel-Aal, R. E. (2008). Construction and analysis of educational tests using abductive machine learning. Computers and Education, 51(1). https://doi.org/10.1016/j.compedu.2007.03.003

García-González, J. R., Sánchez-Sánchez, P. A., Orozco, M., & Obredor, S. (2019). Extracción de Conocimiento para la Predicción y Análisis de los Resultados de la Prueba de Calidad de la Educación Superior en Colombia. Formación Universitaria, 12(4). https://doi.org/10.4067/s0718-50062019000400055

Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016). Deep Learning. MIT Press. ISBN 978-0262035613

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Educational measurement: issues and practice, 12(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Sage.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. Studies in Educational Evaluation, 35(2-3), 57-63.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Hillsdale, NJ: Erlbaum.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

ICFES (2020). Resolución 268 De 2020. https://normograma.icfes.gov.co/docs/resolucion_icfes_0268_2020.htm

Icfes. (2021, Abril). Saber al detalle N 08. Retrieved from Icfes: https://www2.icfes.gov.co/documents/39286/2231027/Edicion+8+-+boletin+saber+al+detalle.pdf/0dbb437b-fded-f05d-5d2e-4426e1663e59?version=1.0&t=1647958807836#.

Jara Pinzón, D., Riascos Villegas, Á. J., & Romero, M. (2010). Detección de copia en pruebas del Estado.

Jung, J. Y., Tyack, L., & von Davier, M. (2022). Automated Scoring of Constructed-Response Items Using Artificial Neural Networks in International Large-scale Assessment. Psychological Test and Assessment Modeling, 64(4), 471-494.

Khobahi, S.; Soltanalian, M. (2019). "Model-Aware Deep Architectures for One-Bit Compressive Variational Autoencoding"

Kim, S. (2006), A Comparative Study of IRT Fixed Parameter Calibration Methods. Journal of Educational Measurement, 43: 355-381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kim, S. (2006), A Comparative Study of IRT Fixed Parameter Calibration Methods. Journal of Educational Measurement, 43: 355-381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. Applied Psychological Measurement, 18(3), 217-228.

Kim, S. H., Cohen, A. S., & Kim, H. O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. Applied Psychological Measurement, 18(3), 217-228.

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. In Foundations and Trends in Machine Learning (Vol. 12, Issue 4). https://doi.org/10.1561/2200000056

Kingma, Diederik P.; Welling, Max (2014-05-01). "Auto-Encoding Variational Bayes". arXiv:1312.6114

Kramer, Mark A. (1991). "Nonlinear principal component analysis using autoassociative neural networks" (PDF). AIChE Journal. 37 (2): 233–243. doi:10.1002/aic.690370209.

Lalor, J. P., Wu, H., & Yu, H. (2017). CIFT: Crowd-Informed Fine-Tuning to Improve Machine Learning Ability. ArXiv: Computation and Language, 6(February).

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open.

Linacre, J. M. (1994). Constructing measurement with a Many-Facet Rasch model. In Objective measurement: Theory into practice: Volume 2.

Londregan, J. (2021). Handbook of Item Response Theory, Volume 1. Measurement: Interdisciplinary Research and Perspectives, 19(1). https://doi.org/10.1080/15366367.2020.1771960

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48(2), 233-245.

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Martínez, R., Hernández, M., & Hernández, M. (2014). Psicometría. Alianza Editorial.

MinEducacion (2022) Regresan las Pruebas Saber 3°, 5°, 7° y 9°. https://www.mineducacion.gov.co/portal/salaprensa/Noticias/410085:Regresan-las-Pruebas-Saber-3-5-7-y-9-para-cerca-de-200-mil-estudiantes-de-1-300-sedes-educativas-de-todo-el-pais

Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892

Muñiz, J. (2018) Introducción a la Psicometría. Teoría Clásica y TRI.

Muñiz, José. (2010). Las Teorías de los Tests: Teoría Clásica y Teoría de Respuesta a los Ítems. Papeles del psicólogo: revista del Colegio Oficial de Psicólogos, ISSN 0214-7823, Vol. 31, Nº. 1, 2010 (Ejemplar dedicado a: Metodología al servicio del psicólogo), pags. 57-66. 31.

Mutch, C., & Tisak, J. (2005). Measurement error and the correlation between positive and negative affect: Spearman (1904, 1907) revisited. Psychological reports, 96(1), 43-46.

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3(1). https://doi.org/10.1016/0022-2496(66)90002-2

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Ostini, Remo; Nering, Michael L. (2005). Polytomous Item Response Theory Models. Quantitative Applications in the Social Sciences. Vol. 144. SAGE. ISBN 978-0-7619-3068-6.

Phelps, R. P. (2011). Standards for educational & psychological testing. New Orleans, LA: American Psychological Association.

PISA 2019, Released Field Trial and Main Survey New Reading Items. https://www.oecd.org/pisa/test/PISA2018_Released_REA_Items_12112019.pdf

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

Rasch, G. (1960). ON GENERAL LAWS AND THE MEANING OF MEASUREMENT IN. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 20-July 30, 1960 (Vol. 4, p. 321). Univ of California Press.

Rios, J. A., & Soland, J. (2021). Parameter estimation accuracy of the Effort-Moderated Item Response Theory Model under multiple assumption violations. Educational and Psychological Measurement, 81(3), 569-594.

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory.

Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. Applied Psychological Measurement, 18(3), 229-244.

Stevens, R. (2006). Machine learning assessment systems for modeling patterns of student learning. In Games and Simulations in Online Learning: Research and Development Frameworks. https://doi.org/10.4018/978-1-59904-304-3.ch017

Stone, JV (2013), "Bayes' Rule: A Tutorial Introduction to Bayesian Analysis"

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1441-1450).

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & Wainer, H. (Eds.), Test Scoring (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thurstone, L. L., & Chave, E. (1929). J. The measurement of attitudes. Chicago, III.: University of Chicago Press.

Tran, Viet Hung (2018). "Copula Variational Bayes inference via information geometry". arXiv:1803.10998

Vakadkar, K., Purkayastha, D., & Krishnan, D. (2021). Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. SN Computer Science, 2(5), 1-9.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Virla, M. Q. (2010). Confiabilidad y coeficiente Alpha de Cronbach. Telos, 12(2), 248-252.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450.

Wolins, L., Wright, B. D., & Rasch, G. (1982). Probabilistic Models for some Intelligence and Attainment Tests. Journal of the American Statistical Association, 77(377). https://doi.org/10.2307/2287805