eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Trabajo de Fin de Grado

Grado en Ingeniería Informática

Computación

Análisis multifactorial del rendimiento escolar en las asignaturas de lengua castellana y matemáticas

Ibon Urbina Atxa

Dirección
Usue Mori (FISS, UPV/EHU)
Amaia Carrión-Castillo (BCBL)

25 de junio de 2023

Resumen

Recientes estudios afirman una correlación positiva entre las notas escolares y el éxito en la vida. En nuestra investigación, a través de técnicas de aprendizaje automático, hemos tratado de predecir las notas en las asignaturas de lengua castellana y matemáticas, para luego, identificar qué tipo de variables han sido las que más importancia han tenido en dichas predicciones.

La base de datos empleada ha sido proporcionada por BCBL. Es un conjunto de datos etiquetado que contiene cerca de 2000 estudiantes de edad entre 6-16 años y 67 variables de distintos ámbitos: territoriales, socioeconómicas, cognitivas, escolares, hábitos de lectura, etc. En cuanto al aprendizaje automático, primero, hemos tratado de predecir ambas notas de manera independiente, utilizando modelos *single-output* de regresión lineal y de árboles de aprendizaje automático (*CART*). Después, hemos intentado mejorar dichos resultados predeciendo ambas variables simultáneamente con la versión *multi-output* de los mismos algoritmos.

Los análisis del estudio muestran, por un lado, que entre los algoritmos de regresión lineal y árboles de clasificación y regresión, son los primeros los que predicen tanto la nota en lengua castellana como la nota en matemáticas con mayor precisión (*RMSE* = 1.16 nota en lengua castellana y *RMSE* = 1.37 en matemáticas); siendo en ambos casos, las variables relacionadas con la cognición y los hábitos de lectura familiares (variables ambientales) e individuales las más importantes. Por otro lado, no hemos conseguido mejorar los resultados de los modelos *single-output* predeciendo ambas notas simultáneamente (en algunos casos incluso, hemos obtenido peores resultados); por lo que hemos concluído, que los modelos que mejor se han ajustado a nuestro problema han sido los de regresión lineal en formato *single-output*.

Índice de contenidos

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

Las personas cambian cuando se dan cuenta del potencial que tienen para cambiar las cosas.

Paulo Coelho (1947 - act.)

¿Cómo me va ir en la vida? Ésta probablemente haya sido una pregunta recurrente durante tu juventud. Y es que, nos preocupa nuestro futuro. Ansíamos una buena vida, un buen trabajo y buenos amigos. Sin embargo, son tantas y tan diversas las variables que influyen en nuestro yo venidero, que es difícil predecir y gestionar nuestro futuro.

En nuestra investigación nos centraremos en uno de los primeros y más importantes filtros que debemos pasar desde pequeñitos: la escuela. De agrado para algunos y un suplicio para otros, para muchos de nosotros esta época ha quedado enterrada bajo tierra. Sin embargo, como individuos, debemos ser conscientes de lo decisiva que puede llegar a ser esta etapa. Ya no solo por los problemas psicológicos que puedan surgir en la niñez y manifestar en la vida adulta (según la psiquiatra alemana Alice Miller [1]: "la experiencia nos enseña que, en la lucha contra las enfermedades psíquicas, sólo disponemos, a la larga, de una sola arma: encontrar emocionalmente la verdad de la historia única y singular de nuestra infancia"), sino por la importancia que tiene el desempeño académico en nuestro futuro. Son muchas las investigaciones que afirman una correlación positiva entre las notas y el éxito en la vida [2]. En concreto, mencionan que notas más altas implican a largo plazo mayores ingresos, felicidad y satisfacción vital. Y es que una buena nota, a parte de indicar buenas habilidades cognitivas como indicaremos después, también representa rasgos de la personalidad importantes [3]. Entre ellos, el autocontrol, la autodisciplina, la persevarancia y la diligencia. De hecho, ya se esta hablando del *sorpasso* del cociente emocional al cociente intelectual en cuanto a medida de éxito se refiere [4]; en muchos casos, cualidades como la perseverancia pueden tener más valor que el propio cociente intelectual. Todas estas nociones las empezó a popularizar el

periodista estadounidense Daniel Goleman en 1995 con su libro "Inteligencia Emocional" [4]; donde presento dicho término como una capacidad para dirigirnos con efectividad a los demás y a nosotros mismos, de reconocer y gestionar nuestras emociones, de tomar consciencia de uno mismo, etc.

Dentro de todo lo que abarca el periodo escolar, focalizaremos nuestro estudio en el rendimiento académico. A través de las notas podemos medir cómo de bien o cómo de mal va cada alumno en cada asignatura. Dichas notas, sobretodo en la escuela, suelen ponderarse a través de exámenes, trabajos y comportamiento en el aula. A través de diferentes asignaturas y métodos de calificación, se pretenden evaluar indirectamente múltiples habilidades y conocimientos: tal y como la nota en educación física mide aptitudes físicas como la fuerza, la velocidad y la flexibilidad, la nota en matemáticas mide aptitudes mentales como el razonamiento, la lógica y la resolución de problemas.

En consonancia con "La Teoría de las Inteligencias Múltiples" de Howard Gardner [5], podemos intuir que en cada asignatura predominan una o varias de las 8 inteligencias propuestas: inteligencia lingüística, inteligencia lógico-matemática, inteligencia espacial, inteligencia musical, inteligencia corporal y cinestésica, inteligencia intrapersonal, inteligencia interpersonal e inteligencia naturalista. Sin embargo, Gardner subraya que tradicionalmente han sido las inteligencias lingüística y lógico-matemática las más valoradas tanto en la sociedad como en el entorno académico [6]. Argumenta, que la mayoría de los exámenes están sesgados y contaminados para recurrir principalmente a las habilidades lingüísticas y lógicas [5]. Por ejemplo, en un examen donde se pretende medir la inteligencia musical, puede existir una pregunta enrevesada y sintácticamente compleja que ponga en aprietos a alguien con buenas habilidades musicales pero malas habilidades lingüísticas. Por lo tanto, Howard concluye que en estas pruebas tan sesgadas siempre sobresaldrán los individuos con una buena inteligencia lingüística y lógico-matemática. De hecho, debido a esta sobreestimación, han acuñado a esta mezcla de inteligencias (lingüística y lógico-matemática) como *la inteligencia erudita* [6].

Por la marcada crítica social y atención que se les presta a estas dos inteligencias en la teoría de Gardner, y porque realmente eran éstas las únicas calificaciones que teníamos, se ha prestado atención a las dos asignaturas en las que, en principio, predominan estas dos inteligencias: lengua castellana y matemáticas, con la dominancia de la inteligencia lingüística e inteligencia lógico-matemática respectivamente. El objetivo principal es dar a luz los factores que influyen en el desempeño de estas dos materias. Contamos con un conjunto de datos etiquetado que dispone de cerca de dos mil estudiantes. Cada uno de ellos es descrito por atributos escolares, socioeconómicos, territoriales, cognitivos, etc; además de disponer sus notas en lengua castellana y matemáticas. Con estos datos y con técnicas de aprendizaje autómatico, intentaremos identificar patrones y comportamientos en el conjunto de datos que sean capaces de predecir tanto la nota en lengua castellana como en matemáticas. Probaremos diferentes algoritmos y seleccionaremos los que mejor se ajusten a nuestro conjunto de datos. Además, analizaremos individualmente cómo ha influido y qué importancia ha tenido cada atributo en la predicción. De hecho, nos interesa saber si existe algún ámbito en concreto que destaque sobre los demás. Es decir, ¿son las variables que describen aspectos cognitivos las más importantes?, ¿las que tienen que ver con el entorno socioeconómico? o ¿tienen todas la misma relevencia?

Desde un punto de vista general, dividiremos el estudio en tres fases:

1. En la primera fase, que corresponde al capítulo 3, acondicionaremos nuestro conjunto de datos inicial para el entrenamiento de los modelos de aprendizaje automático. Este apartado incluye tanto el análisis exploratorio preliminar como el preprocesamiento de datos necesario.

2. En la segunda fase, en el capítulo 4, el objetivo es buscar los mejores algoritmos para predecir de manera independiente la nota en lengua castellana y matemáticas; es decir, entrenando un modelo de aprendizaje automático para cada variable objetivo. Además, analizaremos la importancia de cada variable en dichas predicciones.

3. En la tercera fase, capítulo 5, trataremos de mejorar los resultados de las predicciones anteriores capturando las relaciones entre las dos asignaturas, entrenando diferentes modelos *multi-output* que predicen ambas notas simultáneamente.

Una vez finalizado el estudio y analizadas ambas aproximaciones, se establecerán en el capítulo 6, mejoras y futuros pasos para poder seguir investigando este campo.

CAPÍTULO 2

Objetivos y planificación del proyecto

La aventura es simplemente una mala planificación.

Roald Amundsen (1872 - 1928)

2.1. Objetivos del proyecto

2.1.1. Objetivo principal

El objetivo principal de esta investigación es descubrir qué factores (cuáles son y a qué ámbito pertenecen) son los más importantes a la hora de predecir la nota en lengua castellana y matemáticas.

2.1.2. Subobjetivos

Antes de determinar cuáles son los factores más importantes, primero debemos averiguar cuál es el modelo de aprendizaje automático que mejores resultados obtiene. Para ello, hemos dividido el estudio en dos fases:

1. Análisis de modelos *single-output*, modelos que predicen una única variable de salida.
2. Análisis de modelos *multi-output*, modelos que predicen múltiples variables de salida.

En la primera fase, probaremos diferentes algoritmos de aprendizaje automático y trataremos de ver cuál de ellos es el que mejor se ajusta a nuestro problema. Siempre lo haremos entrenando un modelo por cada asignatura. En esta fase, el objetivo es, por un lado buscar relaciones entre los atributos y la nota en lengua castellana y por otro lado, buscar las relaciones entre los

atributos y la nota en matemáticas. Analizaremos hasta qué punto somos capaces de hacer las predicciones y qué variables son las que mas importancia han tenido en dichas predicciones.

En la segunda fase, dejaremos de tratar ambas asignaturas de manera independiente y trataremos de predecirlas de manera simultánea con un único modelo. Al igual que antes, probaremos diferentes algoritmos de aprendizaje automático para ver cuál es el que mejor se ajusta a nuestros datos. En este caso, el objetivo es buscar relaciones entre los atributos y ambas asignaturas, además de capturar también las relaciones entre las asignaturas. Asimismo, evaluaremos si predeciéndolas de manera simultánea conseguimos mejorar los resultados de los modelos *single-output*.

2.2. Planificación del proyecto

2.2.1. Gestión del alcance

2.2.1.1. Estructura de Desglose del Trabajo

Para facilitar la realización del trabajo, hemos dividido el proyecto en diferentes niveles de trabajo que lo fragmentan en componentes más manejables y controlables. A continuación, enumeramos los distintos apartados y subapartados que a la vez están representados de manera visual y jerárquica en la figura 2.1, mediante la Estructura de Desglose de Trabajo (EDT).

1. **Planificación y seguimiento**: constituye la organización del plan de trabajo, identificación de problemas relacionados y el seguimiento del proyecto.

2. **Investigación y formación**: incluye tanto la revisión de la literatura existente sobre el tema, como la formación pertinente para poder comprender y ejecutar el plan de trabajo. Precisamente, se divide en dos apartados más pequeños:

   a) **Literatura científica**: lectura y análisis de árticulos científicos relacionados tanto con el rendimiento escolar como con las técnicas de aprendizaje automático apropiadas para tratar este tipo de problemas.

   b) **Formación en aprendizaje automático y estadística**: revisión de las técnicas aprendidas durante la carrera y el estudio de nuevos conceptos.

3. **Implementación**: conforma la preparación previa al desarrollo de los modelos de aprendizaje automático y el propio entrenamiento de dichos modelos. Se divide en dos subpaquetes:

   a) **Tratamiento del conjunto de datos**: análisis y preprocesamiento del conjunto de datos para adecuarlo a las técnicas de aprendizaje automático.

   b) **Entrenamiento de los modelos de aprendizaje automático**: creación de los modelos *single-output* y *multi-output*.

4. **Análisis y comparativa**: incluye tanto el estudio de interpretabilidad como la comparación de rendimientos entre los modelos *single-output* y *multi-output*. Se descompone en dos paquetes de trabajo más pequeños:

   a) **Feature importance y feature effects**: interpretación de los modelos. ¿Qué variables son las más importantes para los modelos? ¿Cómo influye cada variable en la predicción?

   b) **Comparativa de rendimiento entre los modelos**: comparación entre los distintos modelos creados en cada aproximación *single-output*/*multi-output* y entre las dos aproximaciones.

5. **Memoria y presentación**: lo forman la redacción de la memoria y la preparación de la presentación para el día de la defensa.

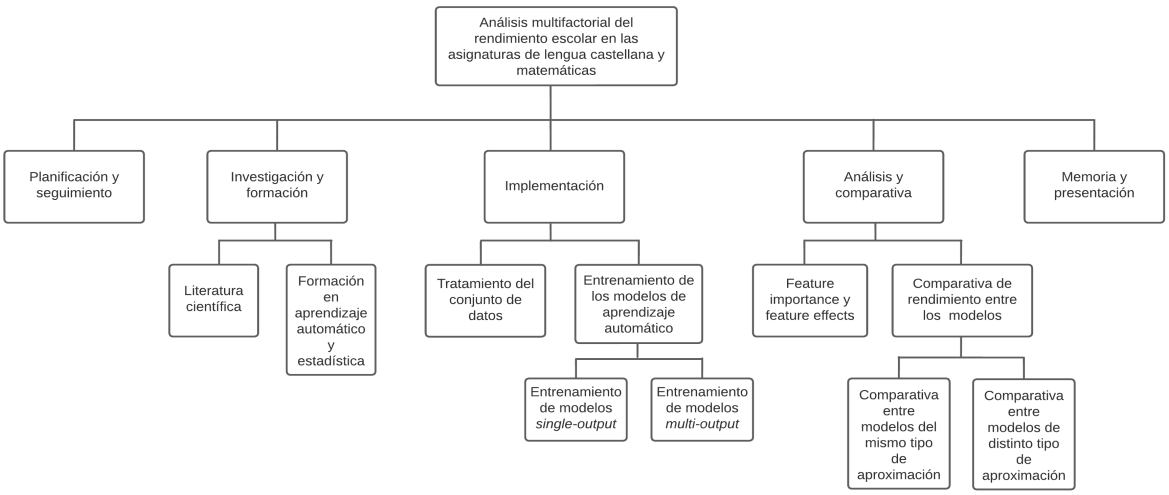

Figura 2.1: Estructura de Desglose de Trabajo (EDT) del proyecto.

2.2.1.2. Entregables

Obviando los propios entregables definidos por las tutoras una vez comenzado el proyecto para poder terminarlo a tiempo, los entregables principales del Trabajo de Fin de Grado son la memoria y la presentación de la defensa. En estas fechas:

- **Memoria**: 25/06/23.
- **Presentación de la defensa**: de 03/07/2023 a 14/07/2023.

2.2.2. Gestión del tiempo

Una vez definidos los diferentes niveles de nuestro EDT, se ha hecho una estimación inicial del tiempo que vamos a requerir para completar cada nivel. Asimismo, para ser más concretos que en la visión general de nuestro EDT, hemos añadido también diferentes tareas que debemos cumplir en cada nivel. Se puede ver un resumen de la dedicación prevista en la tabla 2.1.

1. **Planificación y seguimiento** (20 horas).
2. **Investigación y formación** (60 horas):
    a) Literatura científica (20 horas):
        1) Lectura acerca del rendimiento escolar (5 horas).
        2) Lectura acerca de procesos cognitivos y psicológicos (10 horas).
        3) Lectura acerca de modelos de aprendizaje automático aplicados en problemas similares (5 horas).
    b) Formación en aprendizaje automático y estadística (40 horas):
        1) Modelos de aprendizaje automático interpretables (20 horas).
        2) Modelos de aprendizaje automático *multi-output* (20 horas).
3. **Implementación** (110 horas):
    a) Tratamiento del conjunto de datos (70 horas):
        1) Análisis exploratorio de datos (15 horas).
        2) Preprocesamiento de datos (55 horas).
    b) Entrenamiento de los modelos de aprendizaje automático (40 horas):
        1) Entrenamiento de modelos *single-output* (20 horas).
        2) Entrenamiento de modelos *multi-output* (20 horas).
4. **Análisis y comparativa** (40 horas):
    a) Feature importance y feature effects (20 horas).
    b) Comparativa de rendimiento entre los modelos (20 horas):
        1) Comparativa entre modelos del mismo tipo de aproximación (10 horas):
            - Comparativa entre los modelos *single-output* (5 horas).
            - Comparativa entre los modelos *multi-output* (5 horas).
        2) Comparativa entre modelos de distinto tipo de aproximación (10 horas):
            - Comparativa entre el mejor modelo *single-output* y el mejor modelo *multi-output*. (10 horas)
5. **Memoria y presentación** (70 horas):
    a) Redacción de la memoria (55 horas).
    b) Preparación de la presentación (15 horas).

Tabla 2.1: Tabla que resume la previsión del tiempo que se va a emplear en cada fase.

Apartado	Horas previstas
Planificación y seguimiento	20
Investigación y formación	60
Implementación	110
Análisis y comparativa	40
Memoria y presentación	70
Total	**300**

2.2.2.1. Diagrama de Gantt

En el diagrama de Gantt 2.2 se expone el tiempo de dedicación previsto para las diferentes fases a lo largo del tiempo definido para el proyecto.

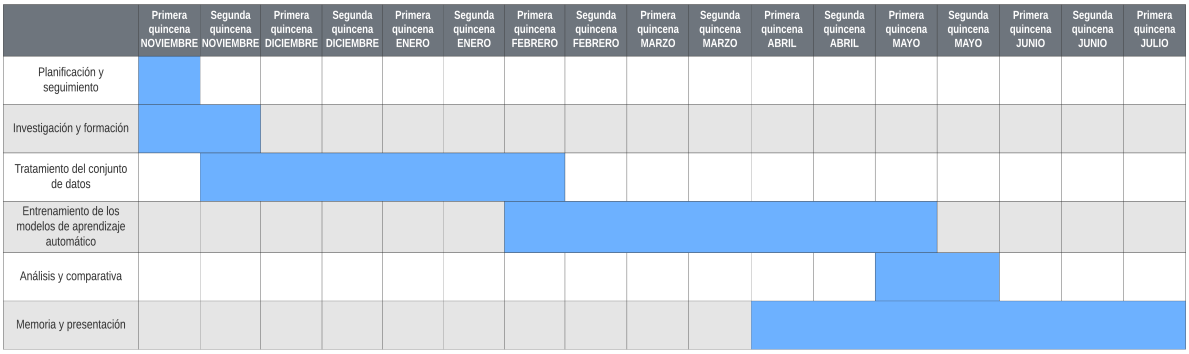

Figura 2.2: Diagrama de Gantt del proyecto.

2.2.3. Gestión de riesgos

Proyectos de este calibre están sujetos a posibles riesgos que deben identificarse y actuar en consecuencia, siguiendo un plan de actuación para reducir el impacto de los mismos.

2.2.3.1. Identificación de riesgos

1. **Planificación del tiempo desacertada**: tanto por problemas técnicos de programación como por falta de comprensión inicial de los conceptos teóricos, pueden existir desviaciones positivas (se requiere más tiempo) en el tiempo. Al igual que pueden existir desviaciones positivas, también existen las negativas (se requiere menos tiempo), pero estas, a priori, no suponen un riesgo.
2. **Pérdida de datos**: el trabajo se realizará en un ordenador personal. A pesar de ser bastante reciente, puede averiarse y perderse todo el progreso; esta pérdida puede incluir tanto los *notebooks* de programación como los ficheros de datos.

3. **Resultados no concluyentes**: para resolver el problema planteado en este estudio, hemos enfocado la investigación de una manera que en principio es lógica y funcional. Sin embargo, a medida que vayan analizándose los resultados puede que descubramos que el enfoque planteado inicialmente no sea el correcto.

2.2.3.2. Análisis cualitativo de riesgos

En la figura 2.3 análizamos tanto la probabilidad de ocurrencia del riesgo como el impacto que tendría en nuestro objetivo en relación al tiempo.

Riesgo	Probabilidad de ocurrencia	Impacto en relación al tiempo
Planificación del tiempo desacertada	Alta	Alto
Pérdida de datos	Pequeña	Alto
Resultados no concluyentes	Media	Medio

Figura 2.3: Matriz que muestra la probabilidad de ocurrencia e impacto de cada riesgo.

2.2.3.3. Planificación de la respuesta a los riesgos

En la tabla 2.2 exponemos un plan de contingencia para cada riesgo, es decir, unas respuestas que están diseñadas para ser usadas únicamente si tienen lugar determinados eventos.

2.2.4. Gestión de las comunicaciones

Dos serán las herramientas principales que emplearemos en el proceso de generación, distribución y recogida de la información entre alumno y tutoras:

- Correo electrónico, medio de comunicación escrito que nos servirá para programar las reuniones, informar acerca de imprevistos y preguntar dudas puntuales.
- *Webex*, medio de comunicación de voz y vídeo que nos permitirá llevar a cabo las reuniones programadas anteriormente para llevar el seguimiento adecuado del proyecto.
- *Google Drive*, medio de almacenamiento en la nube que se empleará tanto para guardar *backups* de seguridad como para compartir información entre alumno y tutoras.

Tabla 2.2: Tabla que muestra un plan de contingencia para cada evento de riesgo.

Riesgo	Evento	Respuesta
Planificación del tiempo desacertada.	Problemas técnicos de programación que exceden el tiempo previsto.	Acudir a foros de programación como *Stack Overflow* para ver si alguién a experimentado un problema similar. En caso de no existir la pregunta en el foro, preguntarlo nosotros mismos.
	Falta de comprensión de los conceptos teóricos y necesitar más tiempo para comprender dichos conceptos.	Intentar no perder el tiempo dándole vueltas y preguntar lo antes posibles a las tutoras por dichos conceptos.
Pérdida de datos.	Se estropea la computadora personal y tenemos que llevarla a arreglar.	Al tener varias copias de seguridad tanto en la nube (*Google Drive*) como en unidades de almacenamiento externo (disco externo *HDD*), bastará con adquirir una nueva computadora ya sea pidiéndola temporalmente a un familiar/amigo o yendo a trabajar a la sala 24h de la universidad y recuperar los datos bajándolos desde la nube o empleando el disco externo.
Resultados no concluyentes.	Los resultados obtenidos en la experimentación no nos dicen nada o no tienen sentido.	Informar cuanto antes a las tutoras de que los resultados no son concluyentes y definir nuevos enfoques para obtener resultados que sí lo sean.

2.2.5. Gestión de los recursos

Para poder completar con éxito el proyecto, necesitaremos ciertas herramientas que se exponen a continuación:

- ***Hardware***:
  - **Computadora**: medio de trabajo principal.
  - **Disco externo *HDD***: medio de almacenamiento externo para almacenar *backups* de seguridad.
- ***Software***:
  - ***RStudio***: entorno de desarrollo integrado para trabajar con el lenguaje de programación *R*.

- **Servicio**:
  - ***Gmail***: servicio de correo electrónico.
  - ***Webex***: servicio para realizar reuniones virtuales.
  - ***Overleaf***: plataforma que se empleará para la realización de la memoria.
  - **Entorno de *Google Drive***: entorno de *Google* para compartir información y creación de tablas, gráficos y documentos.
  - ***Lucidchart***: herramienta principal que se empleará en la creación de tablas y diagramas en el apartado de objetivos y planificación del proyecto.

2.2.6. Seguimiento y control

Gracias a las reuniones virtuales y mensajes de correo electrónico hemos podido hacer un seguimiento y manejar todas las situaciones que se han dado a lo largo de todo el proyecto.

2.2.6.1. Situaciones problemáticas que hemos tenido que resolver

1. **Correlaciones entre distintos tipos de variables**: medir las relaciones/correlaciones entre variables de distinto tipo (discretas, continuas, nominales y ordinales) no es tarea sencilla y en algunos casos se requiere un conocimiento avanzado en estadística. Hemos necesitado mucho más tiempo de lo previsto para poder comprender dichos conceptos, además de limitarnos a cumplir el requisito de que la medida seleccionada es apta para relacionar ambos tipos de variables (se pueden analizar también aspectos como la distribución, escala de medición, etc).

2. **Modelos de aprendizaje automático *multi-output***: el lenguaje de programación *R* no está especialmente enfocado al entrenamiento de este tipo de algoritmos. En el capítulo 5 hemos tenido que emplear un lenguaje de programación adicional (*Python*) y ajustar el entorno de desarrollo *RStudio* a dicho lenguaje. Asimismo, la documentación acerca de los modelos de regresión *multi-output* es bastante escasa, por lo que a parte de dedicarle más tiempo de lo previsto, hemos tenido que actuar con prudencia seleccionando quizás no las opciones más óptimas, pero sí las que nos aseguraban con certeza que cumplían los requisitos (capturar las relaciones entre las variables dependientes).

2.2.6.2. Comparación entre la dedicación estimada y la dedicación real

En la tabla 2.3 podemos ver una comparativa entre las horas que habíamos previsto para cada fase y las horas reales que hemos dedicado a dichas fases.

Tabla 2.3: Tabla que compara la dedicación estimada con la dedicación real.

Apartado	Horas previstas	Horas reales
Planificación y seguimiento	20	15
Investigación y formación	60	80
Implementación	110	130
Análisis y comparativa	40	30
Memoria y presentación	70	75
Total	**300**	**330**

CAPÍTULO 3

Conjunto de datos

Cuantos más datos tengamos, más posibilidades tenemos de ahogarnos en ellos.

Nassim Nicholas Taleb (1960 - act.)

3.1. Análisis exploratorio de datos

El análisis exploratorio de datos (*EDA*) [7] es un enfoque inicial empleado para analizar conjuntos de datos, con el fin de resumir sus características principales utilizando principalmente estadísticos descriptivos, gráficos estadísticos y otros métodos de visualización.

En nuestro caso, lo hemos organizado de tal manera que iniciamos el proceso con un análisis global que nos ayuda a entender el contexto general del conjunto de datos y posteriormente, nos vamos enfocando en aspectos más concretos.

3.1.1. Descripción general del conjunto de datos

En nuestro estudio hemos empleado un subconjunto de datos etiquetado de 1805 alumnos y 67 variables, obtenidos de la base de datos matriz "COEDUCA-BCBL" [8] y delimitándola al estudio de las variables *gradeLanguage* (la nota en lengua castellana) y *gradeMath* (la nota en matemáticas). A partir de ahora, cuando hablemos del conjunto de datos nos referiremos a este subconjunto.

Las instancias de este conjunto corresponden a estudiantes de edad entre 6 y 16 años pertenecientes al sistema educativo español. Concretamente, se ha delimitado el ámbito del estudio a las comunidades autónomas de Castilla y León, Andalucía y País Vasco; con un sesgo importante hacia Andalucía, como podemos ver en la figura 3.1. Debido a la extensión de cada

territorio, no se han analizado todas las escuelas pertenecientes a cada comunidad. En la tabla 3.1 podemos ver que en Castilla y León se han analizado 10 escuelas, en Andalucía 17 y en el País Vasco 3. Igualmente, a pesar de que el idioma común en todos los alumnos es el castellano, también existen estudiantes bilingües y trilingües entre los que destacan los que hablan, a parte del castellano, el euskera y/o el inglés. Asimismo, hablando de los alumnos, cabe destacar que existe una casi completa paridad en la representación de género, tal y como muestra el gráfico 3.2.

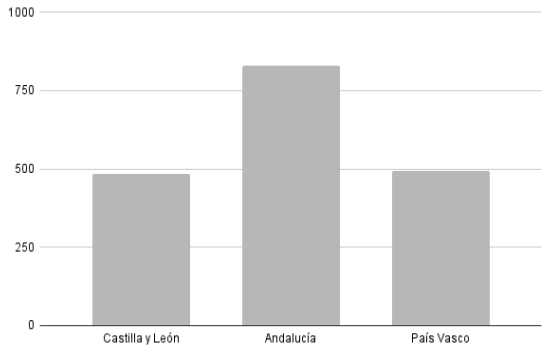

Figura 3.1: Gráfico de barras de la variable *site*.

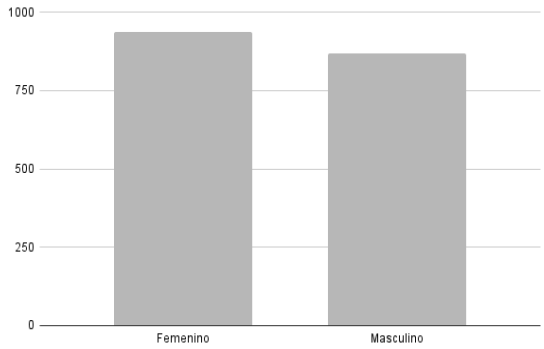

Figura 3.2: Gráfico de barras de la variable *sex*.

Tabla 3.1: Tabla que muestra todos las escuelas analizadas en cada comunidad.

Andalucía	Castilla y León	País Vasco
- Santo Tomás de Villanueva	- Juana I	- Sagrado Corazón
- San Sebastián	- La Asunción	- Urkide
- Ramón y Cajal	- Alejandría	- Amara Berri
- Tinar	- Divina Providencia	
- Los Neveros	- Padre Hoyos	
- Granada College	- Isabel de Castilla	
- Blas Infante	- La Besana	
- La Laguna	- Florida Duero	
- Ceip Alcalde León Ríos	- Pedro I	
- Ceip José Sebastián y Barandarán	- Miguel Cervantes	
- Ceip San José Obrero		
- Ceip El Pinar		
- Ceip Manuel Ciurot		
- Ceip José Payán y Garrido		
- Ceip La Paz		
- Ceip Lora Tamayo		
- Ceip Huerta del Carmen		

3.1.2. Variables del conjunto de datos

Una variable estadística es una característica de una muestra o población de datos que puede adoptar diferentes valores. En nuestro caso, son un total de 67 variables las que conforman nuestro *dataset* (ver apéndice A para una descripción detallada de cada variable) y cada una de ellas, pertenece a una categoría en concreto dependiendo del tipo de atributo que describen. Generalmente, podemos dividir las variables en dos grandes grupos: las cualitativas, que expresan una característica o cualidad observable, y las cuantitativas, que no pueden expresar una cualidad sino que son un factor o propiedad cuantificable que toma valores numéricos. Más concretamente, en el primer gran grupo, podemos diferenciar entre las cualitativas nominales, donde el orden no es importante, y las cualitativas ordinales, donde el orden sí es importante; además de que en el segundo gran grupo, están las cuantitativas discretas, que solo pueden tomar un número finito de valores entre dos valores cualesquiera, y las cuantitativas continuas, que pueden tomar valores numéricos infinitos (acudir a la tabla 3.2 para ver un ejemplo de cada tipo). El diagrama 3.3 nos muestra la distribución de dicha clasificación que corresponde a todas las variables de nuestro conjunto de datos.

Asimismo, también es posible clasificarlas en base al dominio al que pertenecen los atributos descritos. Tal y como podemos ver en la figura 3.4, los dominios examinados en nuestro estudio son siete: el cognitivo, la lectura, el socioeconómico, el escolar, el lenguaje, el territorial y el individual.

Finalmente, es necesario mencionar, que en nuestro conjunto de datos, las variables dependientes, las que queremos predecir, son *gradeLanguage* (la nota en lengua castellana) y *gradeMath* (la nota en matemáticas); mientras que las variables independientes, las que nos van a ayudar a predecir las notas, son, a priori, las 65 variables restantes.

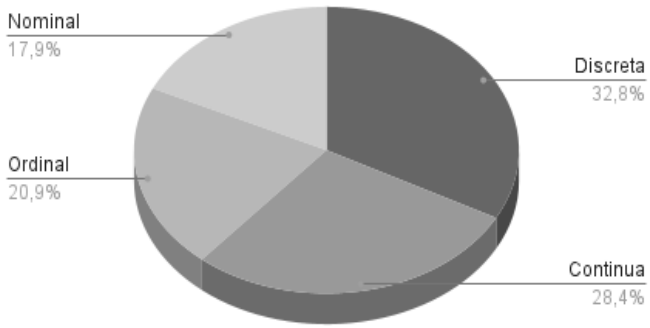

Figura 3.3: Diagrama sectorial que muestra la distribución de los tipos de variables según el atributo que describen.

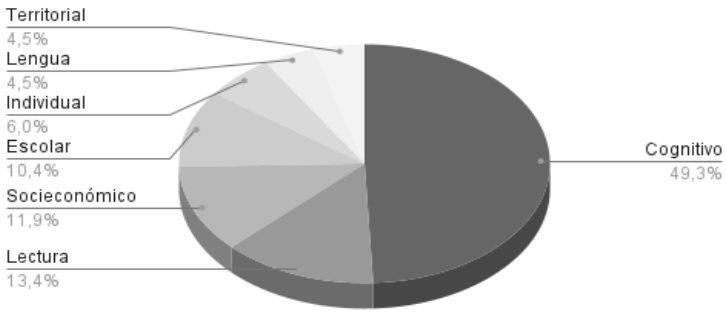

Figura 3.4: Diagrama sectorial que muestra la distribución de los diferentes dominios a los que pertenecen las variables.

Tabla 3.2: Tabla que muestra un ejemplo de cada tipo de variable dependiendo del atributo que describen.

	Tipo de variable	Ejemplo
Cualitativa	Nominal	La variable *school*, tal y como podemos ver en la tabla 3.1, representa el nombre de cada escuela. Los nombres no pueden ser ordenados, ya que a pesar de expresar sustantivos que son distintos entre sí, no tienen una jerarquía o secuencia natural.
	Ordinal	La variable *lectura.animamos.a.leer*, representa la frecuencia con la que la familia anima a su hijo/hija a leer. En este caso, la frecuencia sí que se puede cuantificar y ordenar de menor a mayor. Las categorías de la variable han sido ordenadas de esta manera: nunca $<$ a veces $<$ casi siempre $<$ siempre.
Cuantitativa	Discreta	La variable *lectura* representa, del 0 al 6, una valoración del profesor del nivel de lectura del estudiante.
	Continua	La variable *age*, representa la edad del estudiante como resultado de la diferencia entre la fecha de las pruebas del estudio y su fecha de nacimiento.

3.1.3. Técnicas empleadas en el análisis de las variables

Gracias a los métodos de visualización y a los estadísticos descriptivos, hemos podido describir, resumir las características y obtenido una valiosa información de las variables, que nos ha ayudado en el preprocesamiento posterior (ver apartado 3.2) de los datos.

Principalmente hemos empleado la visualización gráfica (hemos empleado estadísticos descriptivos en casos muy concretos en los que por ejemplo hemos querido saber la media de la nota en lengua castellana o matemáticas). En concreto, histogramas para las variables

continuas y gráficos de barras para el resto de variables. Además, cuando el gráfico no aportaba suficiente información, hemos hecho uso de las tablas de frecuencias para observar todos los valores posibles de cada variable.

En la figura 3.5 podemos ver el uso del histograma para visualizar la variable *age* y definir el rango de valores de la misma. Sin embargo, en el caso de la variable *padre.ocupacion.isco*, además de emplear un gráfico de barras (ver figura 3.6), hemos tenido que crear una tabla de frecuencias (ver figura 3.7) para poder visualizar correctamente todos sus valores.

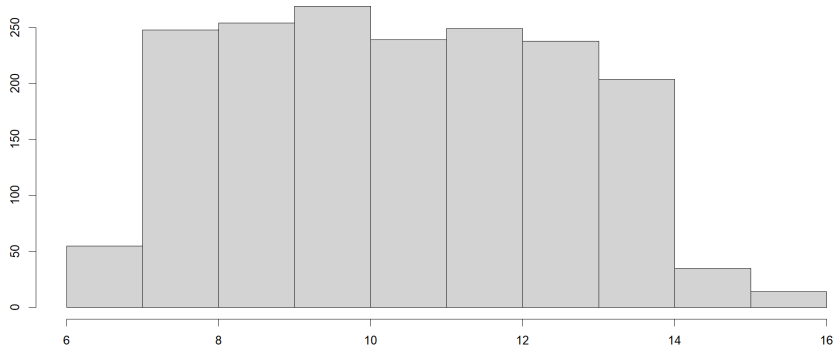

Figura 3.5: Histograma de la variable *age*. Podemos ver que la edad de los estudiantes el día de las pruebas, rondaba sobre los 6 y 16 años, siendo el grupo de estudiantes más pequeño el que corresponde a la edad de 15 y 16 años; y el más grande, el correspondiente a la edad que va desde los 7 a los 14 años.

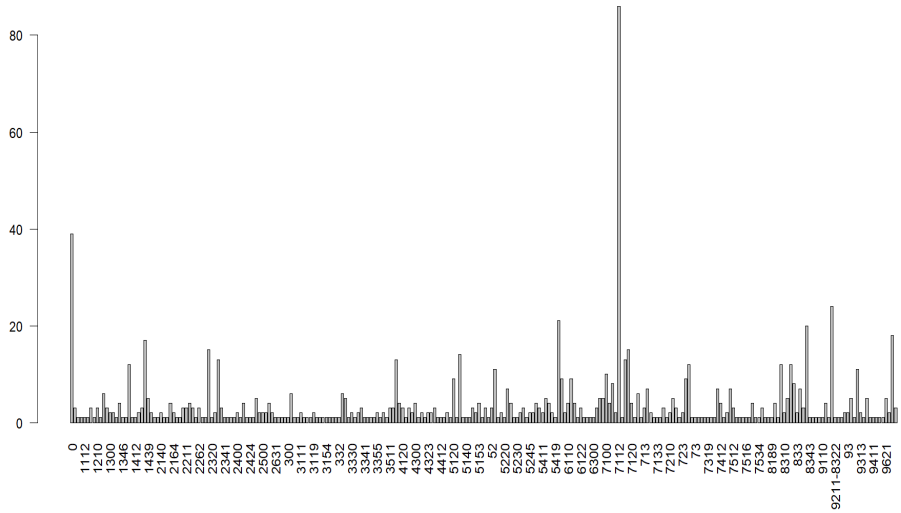

Figura 3.6: Gráfico de barras de la variable *padre.ocupacion.isco*. No tenemos acceso a todos los valores de la variable ya que no todas las barras tienen su correspondiente etiqueta en el eje X. Además, el rango de valores es demasiado grande para poder visualizarlo de manera gráfica. De todas formas, ya vemos que existe una observación con más de un código que deberemos de tratar en el preprocesamiento para que cada observación contenga un único código.

	Var1	Freq
1	0	39
2	1000	3
3	1110	1
4	11110	1
5	1112	1

Figura 3.7: Parte de la tabla de frecuencias de la variable *padre.ocupacion.isco* (la tabla original contiene 260 instancias). En ella, para cada valor tenemos su correspondiente frecuencia de aparición. Por ejemplo, el código 0 (correspondiente a las fuerzas armadas) aparece 39 veces. Asimismo, sabemos que debemos tratar en el preprocesamiento el código 11110 ya que los códigos ISCO-88 [9] tienen como mucho 4 dígitos.

Finalmente, cabe destacar que no todas las variables han sido informadadas. De hecho, un 15 % del conjunto inicial corresponde a valores *Not Available* (*NA*). A pesar de que pueda parecer que todos ellos son datos faltantes, veremos en el apartado 3.2.2 del preprocesamiento, que no es del todo así. Algunas instancias que contienen valores *NA* pueden ser informadas analizando variables del mismo dominio al que pertenecen. Asimismo, desde un principio hemos descartado los individuos que no contenían información en alguna de sus variables dependientes (*gradeLanguage* o *gradeMath*). Son cuatro individuos en total: *id = 3813*, *id = 4104*, *id = 5326* y *id = 5196*.

3.2. Preprocesamiento de datos

El objetivo del preprocesamiento de datos es facilitar su tratamiento por parte de los modelos de aprendizaje automático. Es muy común que los datos crudos obtenidos directamente del mundo real estén mal representados, duplicados, sean contradictorios, nulos, etc. Verlos de manera gráfica, como lo hemos hecho en el análisis exploratorio, nos ha ayudado a detectar este tipo de anomalías y garantizar que los datos con los que vamos a alimentar nuestro algoritmo sean correctos y consistentes.

Los tres pilares fundamentales de nuestro preprocesamiento son: la limpieza de datos, la transformación de datos y el tratamiento de datos faltantes. El primero, implica la eliminación o corrección de errores, como errores de escritura, valores atípicos y problemas similares. El segundo, implica modificar la estructura o el formato de los datos para que sean más adecuados y significativos. El tercero, implica manejar los valores faltantes para que tengan el menor impacto negativo posible.

3.2.1. Limpieza de datos

Tres han sido los casos que hemos tenido que tratar (podemos ver ejemplos en la tabla 3.3). De la siguiente manera:

1. Si existe un valor incoherente, ya sea por definición o por formato, y puede ser sustituido por otro más coherente, vamos a sustituirlo.

2. Si existe un valor incoherente, ya sea por definición o por formato, y no puede ser sustituido por otro más coherente, vamos a sustituir ese valor por *Not Available* y tratarlo como dato faltante.

3. Si existe una variable que es redundante o no aporta información, vamos a eliminarla.

Tabla 3.3: Un ejemplo de cada caso que hemos tenido que tratar en la limpieza de datos.

Caso	Solución
Valor incoherente por definición que puede ser sustituido por un valor coherente.	Un individuo tiene como valor *espana* en la variable *province* y ese valor no corresponde a ninguna provincia. Sin embargo, gracias a la variable *school* sabemos que va a la escuela *tinar* y esa escuela pertenece a *granada*; por ello, hemos sustituido *espana* por *granada*.
Valor incoherente por formato que puede ser sustituido por un valor coherente.	La variable *kbit.pc.matrices* es un percentil y por lo general estos se representan con valores discretos porque representan la posición relativa de un valor en relación con otros en un conjunto de datos. Sin embargo, existen 3 individuos con valores cuantitativos continuos: 0.1, 0.4 y 0.5. Por redondeo, los hemos sustituido de la siguiente manera: 0, 0 y 1 respectivamente.
Valor incoherente por definición que no puede ser sustituido por un valor coherente.	Varios individuos tienen el valor *a* y *nc* en la variable *lectura*. Al no ser valores posibles para esta variable y tampoco se pueden interpretar para intentar sustituirlos por otros más coherentes, hemos tratado aquellos valores como datos faltantes sustituyéndolos con el valor *Not Available*.
Variable que no aporta información.	Todos los valores de la variable *country* son iguales, corresponden a *espana*. Al ser todos iguales, no nos aportan información con la que podamos discriminar los diferentes individuos de nuestro conjunto. Además, tenemos variables territoriales más concretas, *site* y *province*, que nos ayudan a discriminar entre individuos.

3.2.2. Transformación de datos

Dos han sido los casos que hemos tenido que tratar (podemos ver ejemplos en la tabla 3.4). De la siguiente manera:

1. Si las categorías que representan la variable no son las más adecuadas y se pueden mejorar ya sea simplificándolas o especificándolas, hemos creado nuevas o eliminado alguna de ellas.

2. Si el formato de los valores de una variable no es consistente en todos los individuos, hemos especificado un formato en concreto y hemos procedido a la sustitución en los casos requeridos.

Tabla 3.4: Un ejemplo de cada caso que hemos tenido que tratar en la transformación de datos.

Caso	Solución
Se pueden crear y eliminar categorías para mejorar la comprensión de las mismas.	Existen 10 categorías para la variable *madre.ocupacion.isco*: *estudios primarios*, *graduado escolar - eso*, *bup*, *cou - bachillerato*, *fp1*, *fp2 o ciclo medio*, *fp3 o ciclo superior*, *diplomatura*, *licenciatura* y *otros*. Entre ellas, existen estudios que pertenecen al sistema educativo antiguo y se entremezclan con las que pertenecen al nuevo. Por ello, hemos recategorizado la variable, buscando equivalencias entre titulaciones (en negrita las categorías definitivas): ▪ ***primaria*** = *estudios primarios*. ▪ ***secundaria inferior*** = *graduado escolar - eso*, *bup* y *fp1*. ▪ ***secundaria superior*** = *fp2 o ciclo medio* y *cou - bachillerato*. ▪ ***formación profesional*** = *fp3 o ciclo superior*. ▪ ***universidad*** = *diplomatura* y *licenciatura*. Además, hemos eliminado la categoría *otros* ya que no aporta información socioeconómica alguna. No a todos los individuos con esta categoría les hemos asignado el valor *NA*. Si el valor de la variable *madre.diploma.mas.alto* es *otros* y el valor de la variable *madre.ocupacion.isco* no es *NA*, hemos podido hacer estas deducciones: en base a los códigos ISCO-88 [9], para acceder a determinado puesto de trabajo se requiere una cualificación mínima. Por lo tanto, teniendo el puesto de trabajo podemos deducir la cualificación teórica de dicha persona.
El formato en el que se representa cada categoría no es consistente.	Las variables objetivo *gradeLanguage* y *gradeMath* se representan en algunos individuos de manera numérica y en otros de manera ordinal; es decir, el formato no es consistente. Hemos decidido que todos los valores, tanto en *gradeLanguage* como en *gradeMath*, se van a representar en formato numérico con una nota del 0 al 10. Para ello, las observaciones que están ya en formato numérico las hemos dejado tal cual y las que estan en formato ordinal las hemos sustituido por la nota númerica más representada en cada calificación. Por ejemplo, la nota numérica que más veces aparece para lo que entendemos como suspenso es el 4; por ello, cada vez que aparece el valor *insuficiente* lo hemos sustituido por el número 4. Tanto en *gradeLanguage* como en *gradeMath* estas son las notas numéricas más representadas en cada caso: *insuficiente* = 4, *suficiente* = 5, *bien* = 6, *notable* = 7 y *sobresaliente* = 9.

3.2.3. Tratamiento de los datos faltantes

Los datos faltantes o valores faltantes, se producen cuando no se almacena ningún valor para algunas variables e instancias de la base de datos. Muchos de los modelos de aprendizaje automático no toleran valores faltantes como *input* ya que estos valores no se pueden utilizar

para comparar, categorizar y tampoco se les puede aplicar ninguna operación aritmética; de ahí radica la importancia de su tratamiento.

La mayoría de nuestros datos se han recogido gracias a encuestas y sensores, por lo que podemos dar varias explicaciones a la aparición de estos valores. En el caso de las encuestas, al encuestado se le ha podido olvidar o directamente no ha querido constestar alguna de las preguntas. En lo que a los sensores se refiere, debido a una mala configuración, calibración o preparación del dispositivo puede ser que no se haya capturado ningún resultado. Además, en ambos casos, una vez obtenidos los datos, éstos han tenido que ser cargados al fichero o base de datos correspondiente. En esta última etapa, aunque no sea lo habitual, también pueden perderse cierta cantidad datos.

En nuestro caso, llama la atención la cantidad de datos faltantes que existe: después de la limpieza y transformación de datos, un 7.3 % de los datos corresponde a valores *NA*. Los ratios de datos faltantes inferiores al 1 % suelen considerarse triviales, entre el 1 % y el 5 % manejables, entre el 5 % y el 15 % se requieren métodos sofisticados para tratarlos y más del 15 % puede perjudicar los resultados del modelo [10]. Nuestro porcentaje es suficientemente significativo como para tratarlo con métodos sofisticados; podemos verlos en el apartado 3.2.3.2.

3.2.3.1. Eliminación de las variables con alto ratio de *missingness*

En nuestro conjunto de datos existen variables que contienen desde apenas un 1 % de valores faltantes hasta las que contienen casi un 60 % de valores faltantes, tal y como podemos observar en el gráfico 3.8. Por mantener cierta cantidad de veracidad en cada variable y no tener que manipular demasiado cada variable respecto a su estructura original (intentar no imputar grandes cantidades de valores), hemos creado un umbral máximo de *missingness*, es decir, de porcentaje máximo de valores faltantes que una variable puede contener.

No se puede establecer un umbral máximo de *missingness* que sirva como regla de oro y funcione para todos los conjuntos de datos. En nuestro caso, gracias al gráfico de puntos de la figura 3.9, en el que el eje horizontal representa cada variable y el eje vertical el porcentaje de *missingness*, hemos establecido el umbral máximo en 20 %. Todas las variables que están por encima de ese umbral, han sido eliminadas; nueve en total: *madre.ocupacion.isco*, *padre.ocupacion.isco*, *madre.ocupacion.hsbc*, *padre.ocupacion.hsbc*, *t1_rt_total*, *t16.ac_spanfinal*, *t17.ac_threetask*, *t20_indice.congruencia* y *t20_indice.congruenciaerrores*.

Tras este proceso, nos han quedado un total de 53 variables predictivas; es decir, sin contar las variables *gradeLanguage* y *gradeMath*. Entre ellas, tal y como podemos ver en el gráfico 3.8, 8 están totalmente informadas y 45 contienen al menos un valor *NA*; dando lugar a un total de 7230 valores sin informar, un 7.3 % del conjunto total de datos.

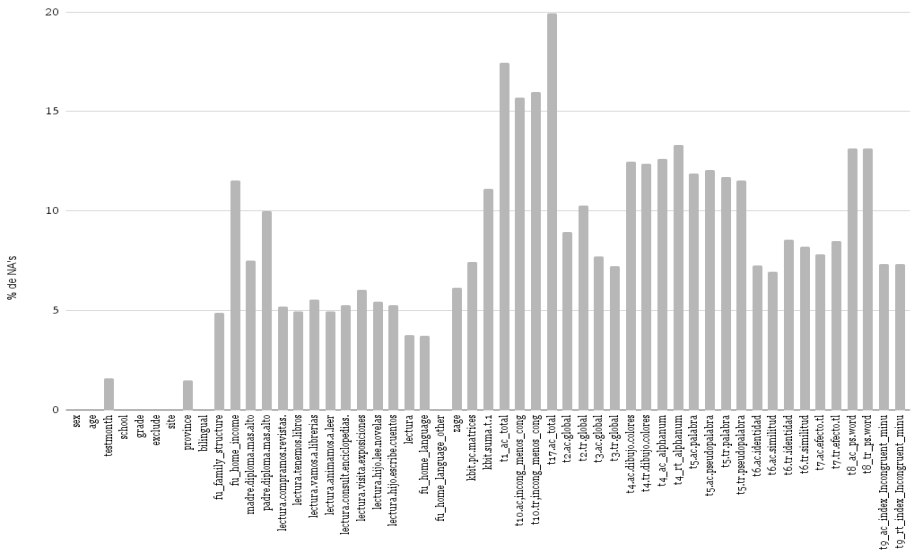

Figura 3.8: Gráfico de barras que muestra el porcentaje de datos faltantes de cada variable.

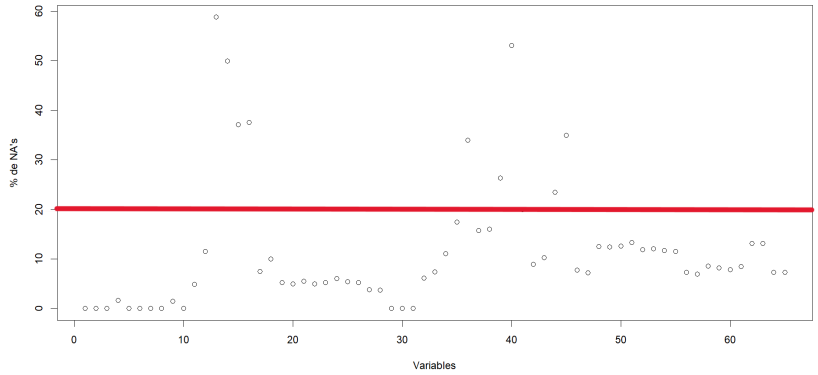

Figura 3.9: Gráfico de puntos que muestra el porcentaje de datos faltantes de cada variable.

3.2.3.2. Imputación de valores faltantes

Dependiendo de la manera en la que haya sido producida la aleatoriedad de los valores faltantes, podemos diferenciar 3 tipos de datos faltantes [10]:

1. **Datos faltantes completamente aleatorios (*MCAR*)**, cuando la probabilidad de tener un valor faltante en una instancia de una variable no depende del propio valor en esa variable ni del valor en otras variables del conjunto de datos.

2. **Datos faltantes aleatorios (*MAR*)**, cuando la probabilidad de tener un valor faltante en una instancia de una variable no depende del propio valor en esa variable pero sí del valor en otras variables del conjunto de datos.

3. **Datos faltantes no aleatorios (*MNAR*)**, cuando la probabilidad de tener un valor faltante en una instancia de una variable depende del propio valor en esa variable.

Tal y como hemos explicado en la sección 3.1.1, nuestro conjunto de datos es producto de otro más grande. Por lo tanto, analizando únicamente nuestro *dataset* es muy difícil determinar de manera fehaciente, qué tipo de aleatoriedad se ha dado en la formación de datos faltantes del conjunto de datos original. Tal y como se hace habitualmente, hemos asumido que el tipo de aleatoriedad es *MAR*. Es una asunción común que permite emplear técnicas estadísticas como la imputación para manejar los datos faltantes, que se refiere al proceso de rellenar valores faltantes, dentro de unos valores posibles, basándose en la información que hay disponible en el conjunto de datos. De hecho, es una técnica, que en nuestro caso, resulta más adecuada que otras técnicas empleadas en el tratamiento de datos faltantes como puede ser la técnica de ignorar y descartar [11], que consiste generalmente en eliminar las filas que contienen valores faltantes en alguna de sus variables. La razón es que si descartaramos todos o la mayoría de valores faltantes, perderíamos una gran cantidad de información que está disponible en nuestro *dataset* y que podría ser relevante para los modelos de aprendizaje automático.

Dentro de la técnica de imputación, existen muchísimas opciones [12]: desde métodos simples como imputación de la media, hasta otros más robustos que se basan en las relaciones entre atributos. Ya que para nosotros la imputación es una herramienta para llegar a nuestro objetivo final y no el objetivo en sí, hemos empleado los métodos que utiliza por defecto la librería *mice* [13] del lenguaje *R*; que son: *predictive mean matching (pmm)*, *logistic regression imputation (logreg)*, *polytomous regression imputation (polyreg)* y *proportional odds (polr)*. Sin embargo, a la lista por defecto hemos tenido que añadir el método *classification and regression trees (cart)*, ya que muchos de los métodos anteriores emplean regresiones lineales y en los casos en los que se produce multicolinearidad, resulta difícil estimar los coeficientes de la regresión y puede generar resultados poco fiables siendo la propia función *mice* la que lanza un error. En la tabla 3.5, resumimos qué métodos de imputación hemos utilizado según el caso.

Tabla 3.5: Métodos de la librería *mice* empleados en el proceso de imputación.

Metodo de imputación	Caso de uso
Pmm	Variables cuantitativas.
Logreg	Variables binarias.
Polyreg	Variables nominales con más de dos categorías.
Polr	Variables ordinales con más de dos categorías.
Cart	Existe multicolinearidad entre las variables y no se obtienen resultados fiables.

Para ahorrar coste computacional y porque creemos que es la manera más lógica para realizar la imputación, únicamente hemos empleado las variables que pueden aportar información real

a cada imputación. Para hacer esta selección de variables, hemos dividido la imputación en dos fases:

1. Medir las relaciones entre variables.

2. Imputar las variables con datos faltantes con la ayuda de las variables con las que comparten más relación.

La primera fase ha consistido, primero, en obtener una medida que cuantifique el grado de relación entre las variables a imputar y todas las demás variables del conjunto de datos. Por defecto, hemos empleado el coeficiente de correlación; sin embargo, no siempre es posible calcularlo. Es una medida adecuada cuando se quiere analizar la relación lineal entre dos variables continuas; pero, por poner un caso, no podemos medir este coeficiente entre dos variables cualitativas nominales. Por ello, hemos tenido que buscar diferentes medidas de relación (expuestas en la tabla 3.6) que satisfacen los requisitos de los diferentes tipos de variables. Es conveniente mencionar que muchas de las funciones que hemos empleado para medir la relación entre una variable A y otra variable B, no son conmutativas ($f(A, B) \neq f(B, A)$) y que es importante establecer correctamente el orden de las variables.

Tabla 3.6: Matriz que muestra la medida que hemos empleado para calcular la relación entre distintos tipos de variables.

		Variable independiente			
		Nominal	**Dicotómica**	**Ordinal**	**Numérica**
Variable dependiente	**Nominal**	V de Cramer	V de Cramer	V de Cramer	Eta cuadrado
	Dicotómica	V de Cramer	V de Cramer	Biserial puntual	Biserial puntual
	Ordinal	V de Cramer	Biserial puntual	Spearman	Spearman
	Numérica	Eta cuadrado	Biserial puntual	Spearman	Pearson

La segunda fase nos ha servido para imputar las variables con algún dato faltante. Como ya hemos mencionado anteriormente, nuestro objetivo es imputarlas únicamente con las variables que puedan aportarles información. Para determinar a partir de qué umbral es una variable apta para ayudar en el proceso, hemos dividido las relaciones en tres niveles: relación grande, relación media y relación pequeña. Ya que las palabras grande, media y pequeña son palabras ambiguas, han sido los valores mostrados en la tabla 3.7 los que hemos tomado como referencia.

Tabla 3.7: Umbrales de referencia para cada medida en base a la magnitud de la relación.

Medida	Relación pequeña	Relación media	Relación grande
V de Cramer	[0, 0.11)	[0.11, 0.16)	[0.16, 1]
Epsilon al cuadrado	[0, 0.13)	[0.13, 0.26)	[0.26, 1]
Eta al cuadrado	[0, 0.06)	[0.06, 0.14)	[0.14, 1]
Biserial puntual	[0, 0.3) y (-0.3, 0]	[0.3, 0.5) y (-0.5, -0.3]	[0.5, 1] y [-1, -0.5]
A de Vargha-Delaney	[0.56, 0.64) y (0.34, 0.44]	[0.64, 0.71) y (0.29, 0.34]	[0.71, 1] y [0, 0.29]
Spearman	[0, 0.3) y (-0.3, 0]	[0.3, 0.5) y (-0.5, -0.3]	[0.5, 1] y [-1, -0.5]
Pearson	[0, 0.3) y (-0.3, 0]	[0.3, 0.5) y (-0.5, -0.3]	[0.5, 1] y [-1, -0.5]

Una vez hemos asignado a cada variable imputable una lista con las variables con las que comparte una relación grande, media y pequeña (para poder ver la lista acudir al apéndice B), el proceso de imputación ha sido el siguiente:

1. Tratar de imputar cada una de las 45 variables que contienen algún dato faltante con las variables que conforman su correspondiente grupo de relación grande. No vamos a poder imputar las 45 variables, ya que son sólo 29 variables las que comparten con al menos alguna otra variable una relación grande. Por lo tanto, 16 variables van a quedar sin imputar.

2. Tratar de imputar esas 16 variables (ver 3.8) con las variables que conforman el grupo relación media. De esas 16 variables, sólo 6 comparten una relación media con al menos una variable y por ello, son las únicas que vamos a poder imputar.

3. Son 10 las variables (ver 3.8) que quedan por imputar y todas ellas comparten una relación pequeña con al menos una variable por lo que podremos imputarlas todas.

Tabla 3.8: Tabla que muestra las variables que han quedado sin imputar tras la imputación con las variables del grupo relación grande y relación media.

Variables que quedan por imputar tras procesar la imputación con el grupo relación grande	Variables que quedan por imputar tras procesar la imputación con el grupo relación media
- lectura	- t3.ac.global
- t10.tr.incong_menos_cong	- t4.ac.dibujo.colores
- t17.ac_total	- t6.ac.identidad
- t3.ac.global	- t6.ac.similitud
- t3.tr.global	- t6.tr.identidad
- t4.ac.dibujo.colores	- t6.tr.similitud
- t4_ac_alphanum	- t7.ac.efecto.tl
- t6.ac.identidad	- t7.tr.efecto.tl
- t6.ac.similitud	- t9_ac_index_Incongruent_minus_Congruent
- t6.tr.identidad	- t9_rt_index_Incongruent_minus_Congruent
- t6.tr.similitud	
- t7.ac.efecto.tl	
- t7.tr.efecto.tl	
- t8_ac_ps.word	
- t9_ac_index_Incongruent_minus_Congruent	
- t9_rt_index_Incongruent_minus_Congruent	

Tras finalizar los tres pasos del proceso de imputación, hemos conseguido que todos los atributos del conjunto de datos estén totalmente informados.

3.3. Conjunto de datos post preprocesamiento

Para concluir, haremos una serie de comparativas entre el conjunto de datos original sin procesar y el conjunto de datos final post preprocesado con el que vamos a trabajar a partir de ahora.

3.3.1. Comparativa respecto a la cantidad de individuos y variables

Tal y como podemos ver en la figura 3.10, el *dataset* con el que vamos a trabajar finalmente, contiene cuatro individuos menos que el *dataset* original. Estos cuatro individuos no contenían

información en alguna de sus variables dependientes (*gradeLanguage* o *gradeMath*) y han sido eliminados (para más información ver apartado 3.1.3).

En cuanto a la cantidad de variables del *dataset*, el gráfico de la figura 3.11 muestra que nuestro conjunto de datos final contiene doce variables menos que el conjunto de datos inicial. Las variables eliminadas son: *madre.ocupacion.isco*, *padre.ocupacion.isco*, *madre.ocupacion.hsbc*, *padre.ocupacion.hsbc*, *t1_rt_total*, *t16.ac_spanfinal*, *t17.ac_threetask*, *t20_indice.congruencia*, *t20_indice.congruenciaerrores*, *id*, *grupo* y *country*. La eliminación de estas variables es producto de la limpieza de datos (ver apartado 3.2.1) y la eliminación de las variables con alto ratio de *missingness* (ver apartado 3.2.3.1).

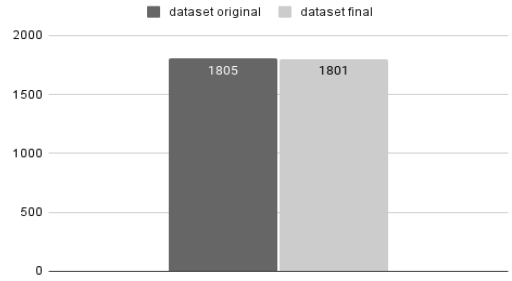

Figura 3.10: Gráfico de barras que hace una comparación respecto a la cantidad de individuos en el *dataset* original y en el *dataset* final.

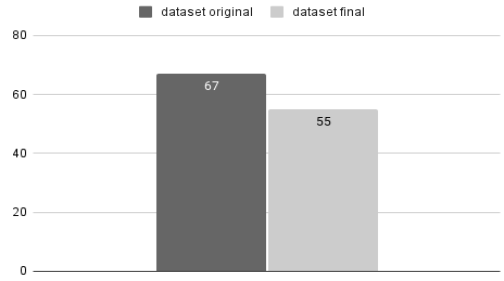

Figura 3.11: Gráfico de barras que hace una comparación respecto a la cantidad de variables en el *dataset* original y en el *dataset* final.

3.3.2. Comparativa respecto a la distribución de los diferentes dominios a los que pertenecen las variables

Analizando los diagramas 3.12 y 3.13, podemos observar que la distribución de las variables que pertenecen a diversos dominios es muy similar en ambos *datasets*. Sin embargo, podemos mencionar, por un lado, que las variables del ámbito escolar ganan una posición en el conjunto de datos final en cuanto a la proporción que les corresponde (en detrimento de las variables del ámbito socioeconómico). Por otro lado, que las variables del ámbito del lenguaje igualan en proporción a las variables del ámbito individual en el conjunto de datos final.

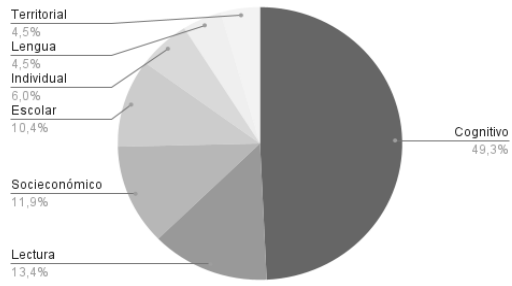

Figura 3.12: Diagrama sectorial que muestra la distribución de los diferentes dominios a los que pertenecen las variables del *dataset* original.

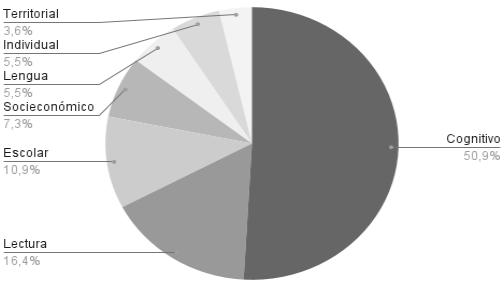

Figura 3.13: Diagrama sectorial que muestra la distribución de los diferentes dominios a los que pertenecen las variables del *dataset* final.

3.3.3. Comparativa respecto a la distribución de los tipos de variables según el atributo que describen

Si observamos los diagramas 3.14 y 3.15, podemos ver que más de la mitad de las variables que tenemos son cuantitativas. Si comparamos ambos *datasets*, la diferencia más reseñable es que en el conjunto de datos final, dentro de las variables cuantitativas, tienen más protagonismo las variables continuas que las variables discretas; al contrario de lo que ocurre en el conjunto de datos inicial.

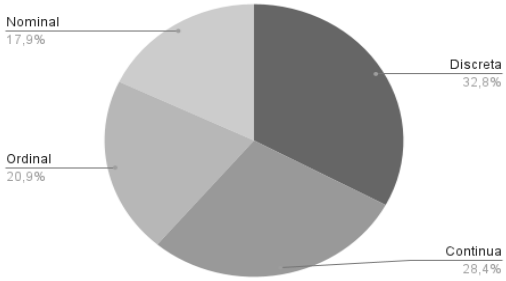

Figura 3.14: Diagrama sectorial que muestra la distribución de los diferentes tipos de variables del *dataset* original.

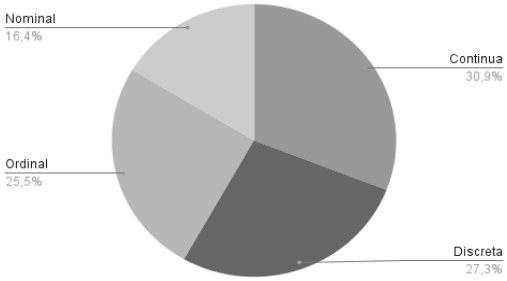

Figura 3.15: Diagrama sectorial que muestra la distribución de los diferentes tipos de variables del *dataset* final.

3.3.4. Comparativa respecto a la cantidad de valores faltantes

Lo que vemos en el gráfico de la figura 3.16 es que al inicio, nuestro *dataset* contenía 18275 valores faltantes y que gracias al preprocesamiento de datos, en especial a la imputación (ver apartado 3.2.3.2), hemos obtenido un *dataset* final totalmente informado.

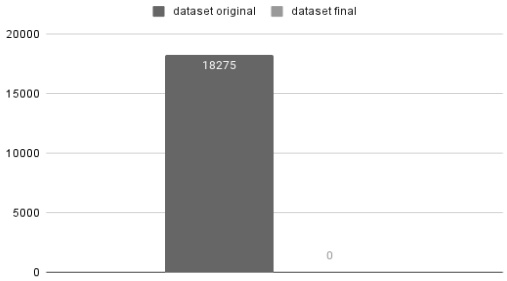

Figura 3.16: Gráfico de barras que hace una comparación respecto a la cantidad de valores faltantes en el *dataset* original y en el *dataset* final.

CAPÍTULO 4

Predicción de las variables *gradeLanguage* y *gradeMath* de manera independiente

Independiente siempre, aislado nunca.

Emilio Visconti (1829 - 1914)

4.1. El aprendizaje automático

El aprendizaje automático (*machine learning* en inglés) es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos estadísticos que permiten a las computadoras aprender, tomar decisiones y hacer predicciones de forma autónoma; es decir, sin necesidad explícita de ser programadas [14] [15].

Por lo general, el modelo de aprendizaje automático es un programa que tras un previo entrenamiento ha aprendido a mapear los datos de entrada con sus respectivos datos de salida. Los datos de entrada (también llamados variables independientes o variables predictoras) son las variables o atributos que se proporcionan al modelo para realizar una predicción. Los datos de salida (también llamados variables dependientes o etiquetas), sin embargo, son las respuestas o resultados esperados que el modelo debe ser capaz de predecir. Un ejemplo típico que suele emplearse para entender bien estos conceptos, es el del modelo que trata de aprender a diferenciar entre un perro y un gato [16]. El modelo recibe miles de imágenes (datos de entrada) que tienen su correspondiente etiqueta (datos de salida) para saber si son imágenes de perro o de gato. En el proceso de entrenamiento el modelo poco a poco va identificando diferentes patrones o características que pertenecen únicamente a perros o a gatos y gracias a ellos comienza

a aprender qué es lo que diferencia a los perros de los gatos (y viceversa). Finalmente, tras completar el entrenamiento, el modelo habrá aprendido a etiquetar datos de entrada (fotos de perros y gatos) que jamás ha visto.

Hasta ahora hemos hablado constantemente de que el modelo de aprendizaje automático aprende. ¿Pero a qué nos referimos cuando hablamos de aprender? En el libro [17] así se define: "se dice que un programa de una computadora aprende a partir de la experiencia E con respecto a una clase de tareas T y una medida de rendimiento P, si su rendimiento en las tareas T, medido por P, mejora con la experiencia E". Nos enfocaremos en los términos tarea T y experiencia E.

El aprendizaje automático nos permite abordar tareas que son demasiado difíciles de resolver con programas fijos escritos y diseñados por seres humanos. El proceso de aprendizaje en sí no es la tarea. El aprendizaje es el medio para adquirir la capacidad de realizar la tarea [18]. En nuestro ejemplo de perros y gatos, nuestro modelo está tratando de resolver una tarea de clasificación; es decir, su objetivo es aprender a asignar una etiqueta o categoría (perro o gato) a cada instancia del conjunto de datos. Sin embargo, los algoritmos de aprendizaje automático pueden resolver muchos tipos de tareas más; por ejemplo, si lo que queremos predecir es un valor numérico, estamos ante un problema de regresión; si lo que queremos es agrupar instancias similares en conjuntos o *clusters*, estamos ante un problema de *clustering*, etc. Podemos ver ejemplos de diferentes tareas en la tabla 4.1.

Tabla 4.1: Ejemplos de tareas que un algoritmo de aprendizaje automático puede aprender.

Tarea	Ejemplo
Clasificación	Además del ejemplo anterior de perros y gatos, podría ser clasificar los correos electrónicos en dos categorías: *spam* y *no spam*.
Regresión	Estimar el precio, en euros, que debería costar una casa.
Clustering	Segmentación de mercado: dividir un mercado objetivo en grupos más pequeños y homogéneos de consumidores o clientes que comparten características y necesidades similares.

Asimismo, dependiendo del tipo de experiencia E, podemos decir que hay tres formas de aprender [18]:

- **Aprendizaje supervisado:** el modelo recibe un conjunto de datos de entrenamiento que consta de ejemplos etiquetados donde cada ejemplo tiene una entrada y una salida esperada. El modelo aprende a partir de los ejemplos etiquetados para poder predecir la salida correcta para nuevos datos de entrada. Nuestro ejemplo de perros y gatos, pertenece a este tipo de aprendizaje.
- **Aprendizaje no supervisado:** el modelo recibe un conjunto de datos de entrenamiento con muchos atributos y sin etiquetas. Al no especificar cuál es la salida correcta, el modelo aprende patrones o relaciones subyacentes entre las instancias del conjunto de datos.

- **Aprendizaje por refuerzo:** el modelo aprende a través de la interacción con el entorno. Como respuesta de sus acciones, el modelo va recibiendo un *feedback* positivo o negativo, por lo que va aprendiendo qué es lo que debe hacer y lo que no. El modelo aprende a tomar las acciones que maximicen las recompensas.

Dicho todo esto, dependiendo de la tarea a la que nos vamos a enfrentar y cómo vamos a abordar el aprendizaje, deberemos elegir un algoritmo de *machine learning* u otro. Se pueden ver algunos ejemplos típicos en la tabla 4.2.

Tabla 4.2: Algunos ejemplos de algoritmos de *machine learning* que podemos emplear en base a la tarea y el aprendizaje elegido. Hemos supuesto que en la tarea de clasificación y de regresión disponemos de un conjunto de datos etiquetado.

Tarea	Tipo de aprendizaje	Algoritmos de *machine learning*
Clasificación	Supervisado	Árboles de decisión
Regresión	Supervisado	Regresión lineal y árboles de decisión
Clustering	No supervisado	*K-means clustering*
Navegación autónoma	Por refuerzo	Agente de aprendizaje por refuerzo alimentado de imágenes preprocesadas con redes neuronales convolucionales [19]

4.2. Nuestro estudio y el aprendizaje automático

Tal y como hemos mencionado en los capítulos 1 y 2, el primer objetivo de nuestra investigación es tratar de predecir *gradeLanguage* y *gradeMath* de manera independiente; es decir, creando un modelo para cada una de las variables dependientes. Queremos que, al proporcionar los atributos de un individuo a cada uno de los modelos, nos dé una predicción de la nota, en formato numérico, que va a sacar el individuo en dichas asignaturas. Al tratarse de predecir un valor continuo utilizando un conjunto de datos etiquetado, estamos frente a un problema de regresión y de aprendizaje supervisado. En este caso, el conjunto de datos que utilizaremos para entrenar los modelos, está compuesto por los datos de entrada (las variables o atributos que describen a cada estudiante) y los datos de salida (*gradeLanguage* o *gradeMath*).

4.2.1. Modelos de aprendizaje automático

Debemos seleccionar un algoritmo de aprendizaje automático que trabaje con el aprendizaje supervisado y los problemas de regresión. Además, en nuestro caso particular, nos interesa saber exactamente cuáles han sido los atributos que han determinado que el modelo se decante por una respuesta u otra. Por ello, analizando el modelo entrenado debemos ser capaces de identificar la

influencia y el peso que ha tenido cada variable en la decisión final. A esto último se le llama interpretabilidad; que tal y como se explica en el artículo [20], es el grado en que un humano es capaz de entender la causa de una decisión; cuanto más grande sea la interpretabilidad de un modelo, más fácil es para el humano comprender el porqué de ciertas decisiones o predicciones.

Por ello, emplearemos métodos tradicionales de aprendizaje automático, que son aquellas técnicas y algoritmos que han sido utilizados durante mucho tiempo antes del auge de los enfoques basados en redes neuronales y aprendizaje profundo. Entre los métodos disponibles, seleccionaremos dos que son interpretables, que sirven para resolver problemas de regresión y que no impiden realizar un aprendizaje supervisado. Éstos son: la regresión lineal múltiple y los árboles de clasificación y regresión (*CART*).

Los motivos por los que no hemos empleado técnicas de aprendizaje profundo (*deep learning*) son dos: por un lado, porque éstas funcionan particularmente bien en casos en los que se dispone de miles o millones de observaciones (nuestro conjunto de datos no es tan grande), y por otro lado, porque este tipo de modelos suelen definirse comúnmente como *black boxes*; es decir, modelos que no pueden entenderse observando sus parámetros o que no revelan su mecanismo interno [21].

4.2.1.1. Regresión lineal múltiple

Los modelos de regresión lineal múltiple buscan establecer relaciones lineales entre la variable dependiente y las variables independientes. Es decir, tal y como se expresa en la figura 4.1, modelan la variable de salida $Y \in \mathbb{R}$ como una suma ponderada de los atributos $\mathrm{X} \in \mathbb{R}^p$ [21] [22].

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon \tag{4.1}$$

Figura 4.1: Regresión lineal múltiple.

Los elementos de la regresión lineal múltiple son:

- $Y \in \mathbb{R}$, la variable dependiente.
- $\beta_0 \in \mathbb{R}$, el intercepto.
- Los betas ($\beta_j \in \mathbb{R}$), donde $j \leq p$ y $p =$ Nº de atributos, representan los pesos aprendidos. A cada atributo le corresponde un peso y por lo tanto una importancia relativa en el modelo.
- Los atributos $\mathrm{X} = (X_1, \ldots, X_i, \ldots, X_p)$, donde $i \leq p$, $p =$ Nº de atributos y $X_i \in \mathbb{R}$.
- ϵ, el error que cometemos.

Son varios los métodos que pueden emplearse para estimar los coeficientes de regresión. En nuestro caso, hemos empleado el método de los mínimos cuadrados ordinarios (*OLS*). Sin entrar en demasiados detalles, el objetivo es encontrar un hiperplano (tenemos más de una variable independiente) que minimice la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (ver fórmula 4.2) [21].

$$\hat{\beta} = \underset{\beta_0,\ldots,\beta_p}{\arg\min} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2 \tag{4.2}$$

Figura 4.2: Fórmula del método de los mínimos cuadrados ordinarios.

Una vez estimados los coeficientes de regresión y el intercepto, podemos predecir la variable de salida Y a través del modelo de la figura 4.3 [22].

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p \tag{4.3}$$

Figura 4.3: Predicción de la variable de salida Y a través del modelo de regresión lineal múltiple ($\hat{Y} \in \mathbb{R}$).

4.2.1.2. Árboles de clasificación y regresión (*CART*)

Cuando los modelos de regresión lineal no tienen la precisión que deseamos, una opción es intentar con los árboles de decisión. Los modelos basados en árboles consiguen capturar las relaciones no lineales entre los atributos y los datos de salida además de las interacciones entre los atributos [21]. El objetivo de estos modelos, al igual que el de las regresiones lineales, es predecir una variable de salida en función de diversas variables de entrada; pero ahora no solo buscando relaciones lineales. Para ello, el algoritmo divide los datos múltiples veces en función de determinados valores de corte en los atributos. A través de estas divisiones (también llamadas particiones), se crean diferentes subconjuntos del *dataset* donde cada instancia pertenece a uno de ellos. Los subconjuntos finales se llaman nodos terminales u hojas y los subconjuntos intermedios, nodos intermedios o nodos de división. Una vez creado el árbol, si queremos predecir el valor de salida de una instancia, debemos recorrer el árbol y coger el valor del nodo hoja al que hemos llegado [21].

Existen muchos algoritmos para hacer crecer el árbol. Difieren en la propia estructura del árbol (por ejemplo, número de divisiones en cada nodo), en el criterio para encontrar las divisiones, en el criterio para dejar de dividir y en cómo hacer las estimaciones en los nodos hoja. En nuestro caso, emplearemos el algoritmo de clasificación y regresión *CART* [23].

Para explicar este algoritmo, vamos a considerar que queremos predecir el consumo de nuestro coche (litros de gasolina por kilómetro) en base a la cilindrada (*cyl*) y caballos de potencia del coche (*hp*) y que tras el entrenamiento del modelo, hemos obtenido el árbol de la figura 4.4. Todas las nuevas instancias que queremos predecir, con formato (*cyl*, *hp*), pasan por

este árbol, se evalúan en un nodo específico y continúan hacia la izquierda si la respuesta es *sí* o hacia la derecha si la respuesta es *no*. Una vez llegados al nodo hoja, es el valor del nodo hoja quien nos da la predicción. Vamos a imaginar que tenemos una nueva instancia con valores (8, 180). Al evaluar el primer nodo tenemos que ir hacia la izquierda ya que *cyl* = 8. Al evaluar el segundo nodo, sin embargo, tenemos que ir hacia la derecha (*hp* = 180 es menor que 192) terminando en el nodo hoja con valor = 18. Por lo tanto, la predicción para nuestra instancia (8, 180) es que el consumo del coche es de 18 litros de gasolina por kilómetro.

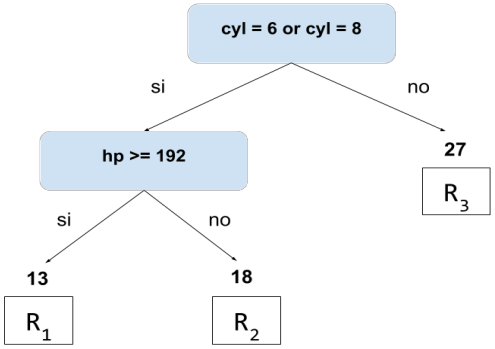

Figura 4.4: Ejemplo de árbol de clasificación y regresión.

Este ejemplo puede generalizarse diciendo [24], por un lado, que tenemos dos variables de entrada X_1 y X_2 junto a una variable de salida Y. Por otro lado, que la partición recursiva tiene como resultado tres regiones finales (R_1, R_2, R_3) donde el modelo intenta predecir Y con una constante c_m para la región R_m. Podemos ver la formulación matemática en la figura 4.5 [22] [24].

$$\hat{Y} = \sum_{m=1}^{3} c_m I(X_1, X_2) \in R_m \tag{4.4}$$

Figura 4.5: Predicción de la variable de salida Y en el algoritmo de árboles de clasificación y regresión (*CART*). $I(X_1, X_2)$ es la función identidad que devuelve 1 si los atributos X_1 y X_2 pertenecen a la región R_m; en caso contrario, devuelve 0.

¿Pero cómo obtenemos todas estas divisiones? Primero, es importante mencionar que la partición de variables se hace de manera descendente y *greedy*. Esto quiere decir que las particiones creadas anteriormente en el árbol no van a cambiar en las particiones posteriores del mismo. En el caso de las tareas de regresión, para crear estas particiones el árbol comienza analizando todo el conjunto de entrenamiento para buscar dos predictores (constantes c_1 y c_2) y los valores de división que particionan los datos en dos regiones, R_1 y R_2, de manera que se minimiza el error de la suma total de cuadrados [22] [24], tal y como se representa en la figura 4.6.

$$\text{mín}\,[\min_{c1}\sum_{i\in R_1}(y_i - c_1)^2 + \min_{c2}\sum_{i\in R_2}(y_i - c_2)^2] \tag{4.5}$$

Figura 4.6: Fórmula para minimizar el error de la suma total de cuadrados en el algoritmo *CART*.

Una vez hemos encontrado la mejor partición, dividimos el árbol en las regiones resultantes y repetimos el proceso en esas dos nuevas regiones. Continuamos así recursivamente hasta alcanzar un criterio de parada; que puede ser: el número mínimo de observaciones que debe haber en un nodo antes de la división o el número mínimo de observaciones que debe haber en un nodo hoja [21] [24].

4.2.2. Entrenamiento de los modelos

Antes de proceder a entrenar ambos modelos, hemos dividido nuestro *dataset* en dos subconjuntos: el conjunto de entrenamiento, que corresponde al 80 % de datos del *dataset* y el conjunto de prueba, que corresponde al 20 %. Esta es una técnica común para evaluar el rendimiento del modelo en datos no vistos y asegurarse de que pueda generalizar correctamente. Precisamente, para evaluar el rendimiento de los modelos hemos empleado la métrica llamada raíz del error cuadrático medio (*RMSE*) que se calcula como la raíz cuadrada de la media de los errores al cuadrado entre los valores predichos por el modelo y los valores reales, tal y como se puede ver en la figura 4.7. Al expresarse en las mismas unidades que la variable objetivo, puede interpretarse como la desviación estándar de los errores de predicción. Es decir, un valor de *RMSE* más bajo indica un mejor ajuste del modelo, ya que en promedio existe una diferencia menor entre el valor predicho y el real.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.6}$$

Figura 4.7: Fórmula de la raíz del error cuadrático medio.

En los entrenamientos de los modelos de *machine learning*, es común llevar a cabo un proceso llamado ajuste de hiperparámetros (*hyperparameter tuning*) para encontrar los valores óptimos de cada hiperparámetro. Éstos, son aquellos parámetros que pueden ajustarse antes de que comience el propio entrenamiento del modelo y que no se pueden aprender directamente a partir de los datos. En la tabla 4.3 podemos ver algunos de los hiperpárametros que pueden ajustarse en los árboles de clasificación y regresión con la librería *rpart* [25] de *R*. No hay una tabla para la regresión lineal múltiple porque no tiene hiperparámetros como tal.

Tabla 4.3: Tabla que muestra las definiciones de algunos hiperparámetros que pueden ajustarse en la librería *rpart*.

Hiperparámetro	Definición
minsplit	Mínimo número de observaciones que debe existir en un nodo para que pueda realizarse una partición.
minbucket	Mínimo número de observaciones que debe haber en un nodo hoja.
maxdepth	Profundidad máxima de cualquier nodo del árbol, teniendo el cuenta que el nodo raíz tiene profundidad 0.
cp	Parámetro de complejidad. Cualquier división potencial que no logra reducir la falta de ajuste del modelo por un factor igual o mayor que el valor de *cp*, no se intenta.

Sin embargo, nuestro objetivo principal es saber qué tipo de algoritmo (lineal, no lineal, etc) se ajusta mejor a nuestro problema y no tanto, dedicar todos nuestros esfuerzos al ajuste de hiperparámetros con el fin de encontrar el mejor modelo de cada algoritmo (una vez sabemos si es la regresión lineal o *CART* el que mejor funciona, sí que hemos intentado mejorarlo con técnicas de selección de características como se puede ver en el apartado 4.2.3). Por ello, para acelerar el proceso de entrenamiento de cada modelo, hemos empleado la librería *caret* [26] y su función *train()*, que entre otras, ofrece estas posibilidades:

- Entrenar un modelo con sus respectivos datos de entrada y salida.
- Mediante técnicas de remuestreo evaluar los efectos de los ajustes de hiperparámetros en el rendimiento del modelo.
- Seleccionar el modelo óptimo en base a los ajustes de hiperparámetros.

Para el entrenamiento de los modelos de regresión lineal no hemos empleado ninguna técnica de remuestreo, ya que no hay hiperparámetros que optimizar. Sin embargo, en los modelos de árboles de clasificación y regresión, hemos empleado la validación cruzada de k-pliegues (*k-fold cross-validation)*, con $k = 10$, para optimizar el hiperparámetro *cp*, que es el hiperparámetro que se centra en optimizar la función *train()*. Es importante matizar, que la función *train()* de *caret* entrena los modelos de árboles de clasificación y regresión empleando la librería *rpart*. No obstante, como ya hemos dicho, de todos los hiperparámetros que permite ajustar esta última librería, *caret* se centra únicamente en el parámetro de complejidad y emplea para los demás hiperparámetros los valores por defecto que define *rpart*.

En la figura 4.8, podemos ver cómo combina la función *train()* la validación cruzada 10-pliegues y el ajuste del hiperparámetro *cp* en nuestros modelos de árboles de clasificación y regresión.

```
Definir el conjunto de valores posibles para el hiperparámetro cp
for cada valor posible del hiperparámetro cp do {
     for cada iteración en validación cruzada 10-pliegues do {
          Entrenar el modelo con los datos de entrenamiento (k-1 pliegues)
          Predecir los valores del conjunto test (1 pliegue)
     }
     Calcular el rendimiento promedio sobre el conjunto test
}
Determinar el valor óptimo para el hiperparámetro cp
Entrenar el modelo final con todo el dataset de entrenamiento empleando el valor óptimo de cp
```

Figura 4.8: Pseudocódigo de la optimización del hiperparámetro *cp* a través de la validación cruzada. Los valores posibles del hiperparámetro *cp* son generados automáticamente por *caret* basándose en ciertos heurísticos y reglas predefinidas. Para profundizar en cómo lo hace, podemos acudir al código publicado en [26].

4.2.3. Comparación de modelos

La tabla 4.4 muestra los resultados obtenidos a través de los procedimientos explicados anteriormente.

Tabla 4.4: Resultados de los modelos empleados para predecir individualmente *gradeLanguage* y *gradeMath*.

Modelo	**RMSE**	
	gradeLanguage	**gradeMath**
Regresión lineal múltiple	1.158708	1.366459
Árboles de clasificación y regresión (*CART*)	1.374892	1.683497

Para nuestro caso en concreto, a pesar de que el algoritmo *CART* tiene la capacidad de capturar relaciones más complejas que la regresión lineal múltiple, que asume una relación lineal entre la variable dependiente y las variables independientes, ha sido la regresión lineal múltiple la que ha obtenido mejores resultados a la hora de predecir ambas notas.

Una vez sabido esto, hemos tratado de mejorar el modelo de regresión lineal a través de métodos de selección de características. Dentro de todas las posibilidades que ofrecen éstos [26], hemos elegido un método tipo *wrapper* llamado eliminación recursiva de características (*RFE*). Los métodos de tipo *wrapper* evalúan múltiples modelos (del mismo tipo) utilizando procedimientos que añaden y/o eliminan predictores para encontrar la combinación óptima que maximiza el rendimiento del modelo [26]. En el caso concreto de la eliminación recursiva de características, hemos empleado la selección hacía atras. Consiste en lo siguiente: primero, el algoritmo ajusta el modelo con todos los predictores y se clasifica (se hace un ránking) cada predictor según su importancia para el modelo. Definamos ahora, S, como una secuencia de números ordenados que son valores candidatos para la cantidad de predictores a retener en el modelo final ($S_1 > S_2 > S_3 \ldots$). En cada iteración de la selección de características, se retienen los top S_i predictores, se vuelve a ajustar el modelo y se calcula el rendimiento. Se determina el valor de S_i que mejor rendimiento obtiene y se utilizan los top S_i predictores para entrenar el modelo final [26].

En nuestro caso, al algoritmo previamente descrito, le hemos añadido una capa externa de remuestreo (la validación cruzada con $k = 10$) para obtener estimaciones de rendimiento que incorporen la variación debida a la selección de características. De no hacerlo, puede ocurrir lo que se denomina en el artículo [27] como *selection bias*; es decir, que una variable predictora se relacione aleatoriamente con los resultados de los datos de entrenamiento y que el algoritmo de *RFE* dé a esta variable una buena puntuación en el ránking de importancia. Sin embargo, al comprobar con otro conjunto de datos, se descubre que en realidad esa variable no era tan importante como se creía en un principio. Se puede ver el algoritmo de *RFE* empleado en la figura 4.9 [26].

```
for cada iteración en validación cruzada 10-pliegues do {
     Dividir el dataset en datos de entrenamiento y test;
     Entrenar el modelo empleando todos los predictores;
     Predecir las instancias correspondientes al grupo test;
     Clasificar las variables en base a su importancia;
     for cada tamaño de subconjunto de predictoras S_i, i = 1...length(S) do {
          Seleccionar las S_i variables predictoras más importantes;
          Entrenar el modelo empleando S_i variables predictoras;
          Predecir las instancias correspondientes al grupo test;
     }
}
Evaluar el rendimiento sobre cada tamaño de subconjunto S_i empleando los grupos test;
En base a los diferentes rendimientos, elegir el mejor tamaño de subconjunto S_i;
Estimar la lista de los predictores finales que van a emplearse en el modelo final;
```

Figura 4.9: Pseudocódigo del algoritmo de selección de características *RFE* que hemos empleado.

Como vemos en la figura 4.9, en cada iteración de la validación cruzada se hace un ránking de las variables más importantes para el modelo. Esto implica que para cada conjunto de datos de entrenamiento empleado en cada iteración, el ránking de variables puede ser diferente. Ésto puede complicar la estimación de la lista final de predictores que el algoritmo hace en la última línea de código. Así lo resuelve *caret*:

1. Selecciona el número de predictoras que el modelo final debe tener. Calcula, en promedio, con qué número de predictoras se han obtenido los mejores resultados. Imaginemos que el algoritmo determina que en promedio se obtienen los mejores resultados con 20 predictoras.

2. En cada iteración de la validación cruzada, el algoritmo ha clasificado cada variable en base a los coeficientes de regresión (hay que tener en cuenta que en este proceso únicamente hemos empleado la regresión lineal). Un coeficiente de regresión alto (valor absoluto) supone un puesto alto en el ránking, mientras que un coeficiente bajo (valor absoluto) supone un puesto bajo. Ahora que el algoritmo sabe que tiene que seleccionar el mejor subconjunto de 20 variables predictoras, suma la puntuación (el valor absoluto del coeficiente de regresión) que ha obtenido cada variable en cada uno de los ránkings y selecciona las 20 variables que mayor puntuación suman.

Podemos ver en la tabla 4.5 que tras entrenar los nuevos modelos de regresión lineal con las variables seleccionadas en el proceso de selección de características, hemos obtenido unos resultados muy similares (un poco peores de hecho) si los comparamos con los modelos iniciales, pero con una cantidad bastante menor de variables predictoras.

Tabla 4.5: Comparación entre los modelos de regresión lineal múltiples iniciales y los optimizados mediante el proceso de eliminación de características recursiva. Nota: hay que tener en cuenta que la cantidad de variables predictoras descrita es teniendo en cuenta las variables cualitativas como *dummy*.

Modelo	gradeLanguage		gradeMath	
	RMSE	Nº predictoras	RMSE	Nº predictoras
Regresión lineal múltiple inicial	1.158708	59	1.366459	59
RFE + regresión lineal multiple	1.165011	38	1.368221	31

4.3. Discusión y análisis de los resultados

Han sido los modelos de regresión lineal los que mejor rendimiento han dado. Entre ellos, los que previamente no han sido sometidos a un proceso de selección de características han conseguido resultados ligeramente mejores. Sin embargo, utilizaremos como referencia los modelos de regresión lineal entrenados tras el proceso de selección de características, porque obtienen resultados muy similares con una cantidad significativamente menor de predictoras.

4.3.1. ¿Hasta qué punto somos capaces de predecir la nota en *gradeLanguage* y *gradeMath*?

No siempre se consigue capturar mediante un modelo de aprendizaje automático la varianza total de la variable dependiente. Cuando hablamos de la varianza de la variable dependiente, nos referimos a una medida de dispersión que indica la extensión en la que los valores de esa variable están dispersos alrededor de su media. Tener una gran varianza implica que existe una gran variabilidad de los datos y que los valores observados están bastante alejados de la media. Por lo tanto, ser capaces de capturar gran parte de la varianza de la variable dependiente, significa que el modelo es capaz de explicar o predecir una cantidad significativa de la variabilidad observada en la variable dependiente. La parte de la varianza que no conseguimos capturar con el modelo de regresión lineal, se explica como las diferencias entre los valores observados de la variable dependiente y los valores predichos por el modelo. Es decir, el error que comete el modelo (conocido también como error residual).

En el caso de la regresión lineal, esta varianza que explica el modelo se calcula a través del coeficiente de determinación (R^2) [21]. Tal y como representa la fórmula de la figura 4.10, la suma de los cuadrados de los residuos (*SSE*) determina cuánta varianza queda sin explicar después de ajustar el modelo, la suma total de los cuadrados (*SST*) determina las desviaciones entre los valores observados y la media de la variable dependiente, y la división entre ambas representa la proporción de la variabilidad no explicada por el modelo en relación con la variabilidad total.

$$
\begin{aligned}
R^2 &= 1 - \frac{SSE}{SST}, \text{ donde:} \\
SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2
\end{aligned}
\tag{4.7}
$$

Figura 4.10: Fórmula del coeficiente de determinación (R^2).

En nuestro caso, a pesar de que conseguimos predecir la nota en ambas asignaturas con un error promedio menor a 1.4 puntos, conseguimos capturar en el mejor de los casos, aproximadamente el 50 % de la varianza total de la variable dependiente. En la tabla 4.6 podemos ver exactamente la varianza capturada por los modelos de *gradeLanguage* y *gradeMath*. Esta poca variabilidad explicada por el modelo puede deberse en gran medida a que en el apartado 3.2.2 del preprocesamiento de datos, se hace un cambio de formato en las variables dependientes y se pierde bastante información. Las variables ordinales se sustituyen por variables numéricas. El problema es que un valor ordinal como puede ser *sobresaliente*, se sustituye por la nota numérica más frecuente en la calificación *sobresaliente*. En este caso, todos los valores *sobresaliente* se sustituyen por el número 9, cuando no siempre se da esa relación. A veces, podrá corresponder, pero otras muchas veces la verdadera nota numérica será 9.2, 9.5, 9.75, 10, etc.

Tabla 4.6: Coeficiente de determinación de los modelos de regresión lineal tras el proceso de selección de características.

Modelo	R^2	
	gradeLanguage	**gradeMath**
RFE + regresión lineal múltiple	0.5342	0.4256

4.3.2. De las variables que hemos analizado, ¿cuáles son las que más importancia tienen en el rendimiento del modelo? ¿y cómo influyen?

4.3.2.1. Feature importance

La importancia de las características o *feature importance*, nos permite estimar, dado un modelo predictivo ya aprendido, la importancia (o relevancia relativa) de cada variable predictora respecto a las predicciones que realiza el modelo [28].

En el caso de la regresión lineal múltiple, la importancia o peso relativo de cada variable independiente se puede medir mediante el valor absoluto del estadístico t de cada variable independiente [21]. Siguendo su fórmula (ver figura 4.11), este estadístico representa el peso estimado de la variable escalado por su error estándar. Es decir, la importancia de un atributo es mayor cuando su coeficiente de regresión incrementa a la vez que cuanta más varianza tenga dicho coeficiente, menos importante va a ser la variable.

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{4.8}$$

Figura 4.11: Fórmula del estadístico *t*.

Predicción de *gradeLanguage*

Haciendo un análisis que va desde una perspectiva general hasta una más específica, podemos decir que las variables que más importancia tienen en nuestras predicciones de la variable *gradeLanguage*, son las relacionadas con la cognición y la lectura; que componen casi el 60 % del peso total de las variables predictivas.

De las variables cognitivas destacan muy por encima de las demás las asociadas con la decodificación, representando el 60 % de las variables cognitivas. Entre las variables vinculadas con la lectura, destacan las relacionadas con hábitos/esfuerzos familiares (variables ambientales) y las que se relacionan con los hábitos/esfuerzos individuales, estando levemente por encima las primeras.

Finalmente, si analizamos cada variable individualmente, sobresable de manera notable la variable *lectura* que casi tiene una importancia tres veces mayor que la segunda variable más importante en la lista, la variable *age*.

Puede verse este análisis deductivo siguendo el orden de las figuras 4.12, 4.13 y 4.14.

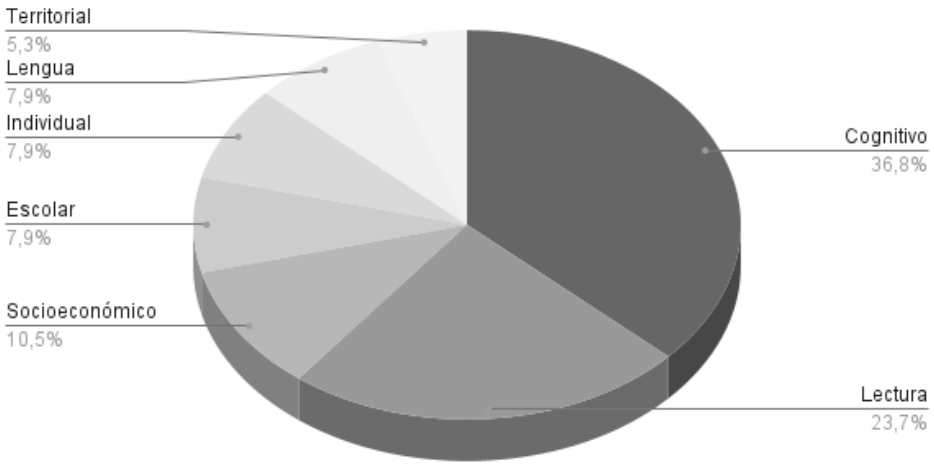

Figura 4.12: Diagrama sectorial que muestra la importancia (medida a través del estadístico *t*), a la hora de predecir la variable *gradeLanguage*, de las variables en base al ámbito al que pertenecen. Podemos ver que gran proporción del gráfico pertenece a los ámbitos de cognición y de lectura.

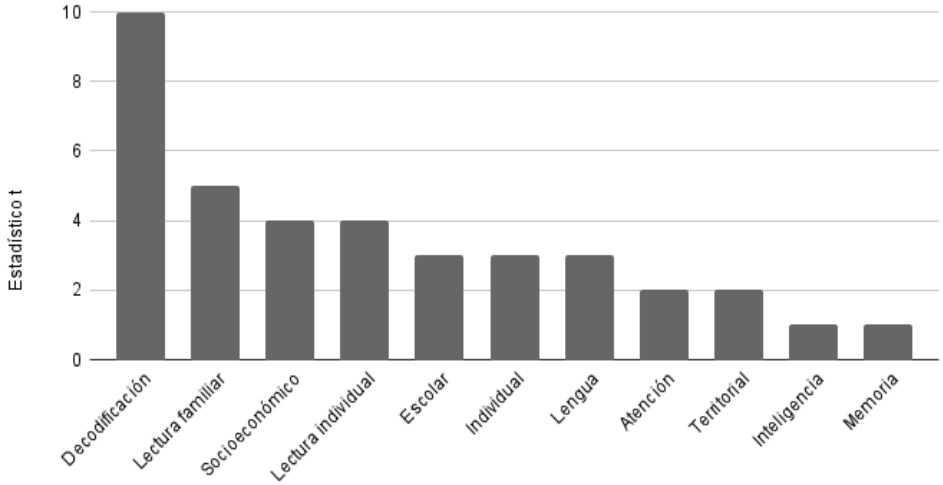

Figura 4.13: Gráfico de barras que muestra la importancia (medida a través del estadístico *t*), a la hora de predecir la variable *gradeLanguage*, de las variables en base al ámbito al que pertenecen, pero concretando un poco más cada ámbito. Destaca muy por encima de los demás el ámbito de la decodificación, que pertenece al grupo de las variables cognitivas.

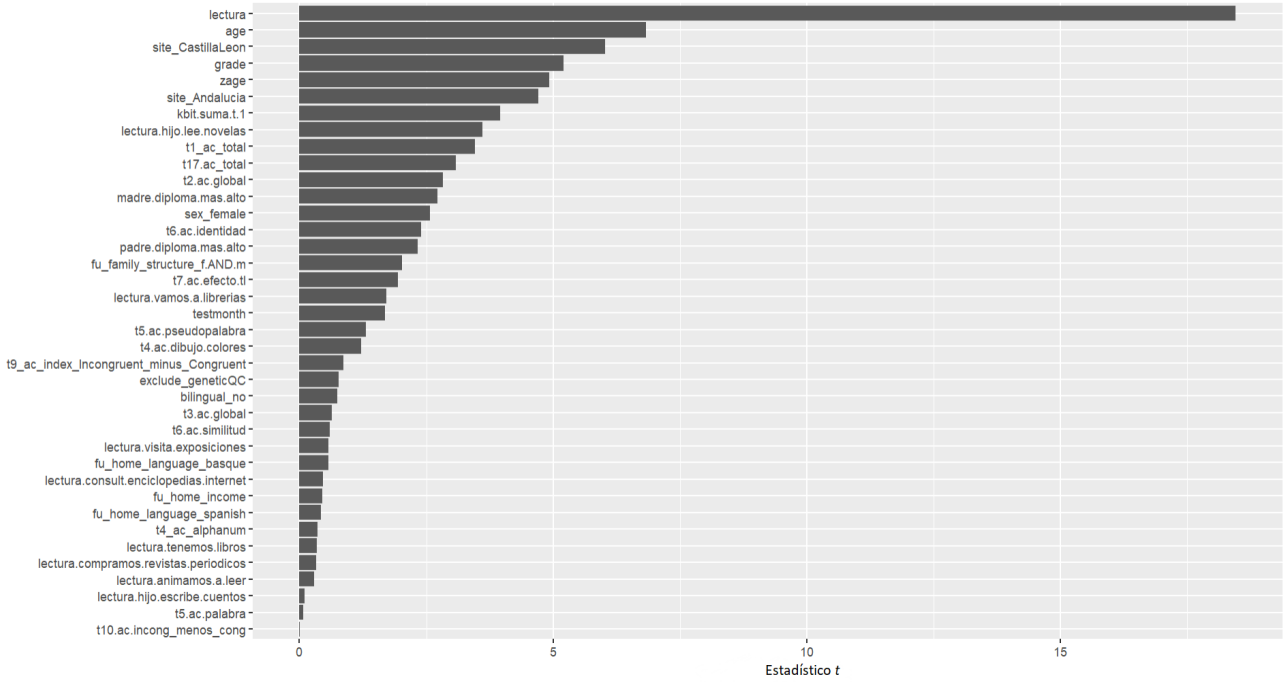

Figura 4.14: Gráfico de barras que muestra la importancia (medida a través del estadístico *t*) de cada variable a nivel individual. Es la variable *lectura*, sin lugar a dudas, la más importante para el modelo a la hora de predecir la variable *gradeLanguage*.

Predicción de *gradeMath*

Al igual que antes, explicaremos la importancia de las variables de manera deductiva siguiendo las figuras 4.15, 4.16 y 4.17.

A nivel global, tal y como ocurre con *gradeLanguage*, tanto las variables relacionadas con la lectura como las cognitivas representan más de la mitad de la importancia total. Sin embargo, al contrario de lo que ocurría antes, aquí tienen más importancia las relacionadas con la lectura que las relacionadas con la cognición.

Las variables asociadas a la decodificación y a la lectura familiar e individual siguen siendo, respectivamente, las más representativas en los ámbitos cognitivos y lectura. No obstante, tanto las relacionadas con la lectura familiar e individual como las socioeconómicas adquieren una importancia que casi está a la par de las variables asociadas a la decodificación; cosa que no ocurría con *gradeLanguage*. Además, merece mencionar que al contrario del anterior análisis, en este no aparecen las variables relacionadas con la atención que forman parte del gran grupo de las relacionadas con la cognición.

A nivel individual, la variable lectura sigue destacando muy por encima de las demás, siendo casi tres veces más importante que la segunda; que en este caso es la variable *kbit.suma.t.1*.

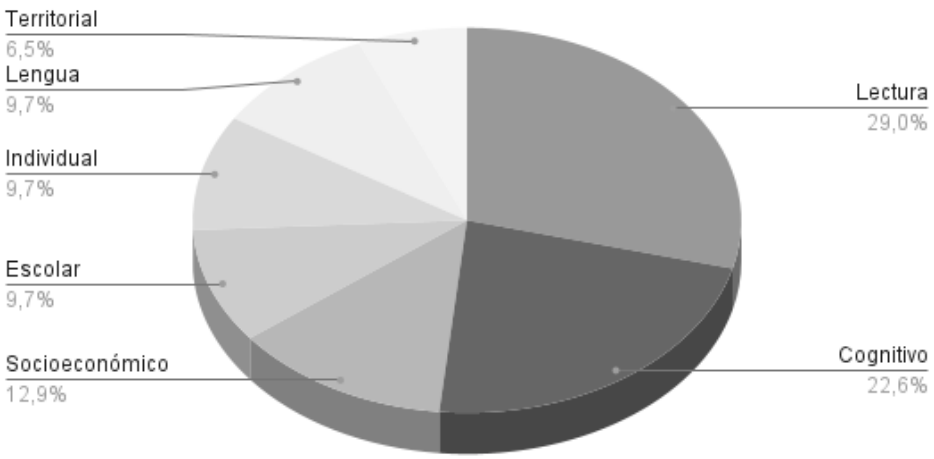

Figura 4.15: Diagrama sectorial que muestra la importancia (medida a través del estadístico *t*), a la hora de predecir la variable *gradeMath*, de las variables en base al ámbito al que pertenecen. Podemos ver que gran proporción del gráfico pertenecen a los ámbitos de lectura y de cognición.

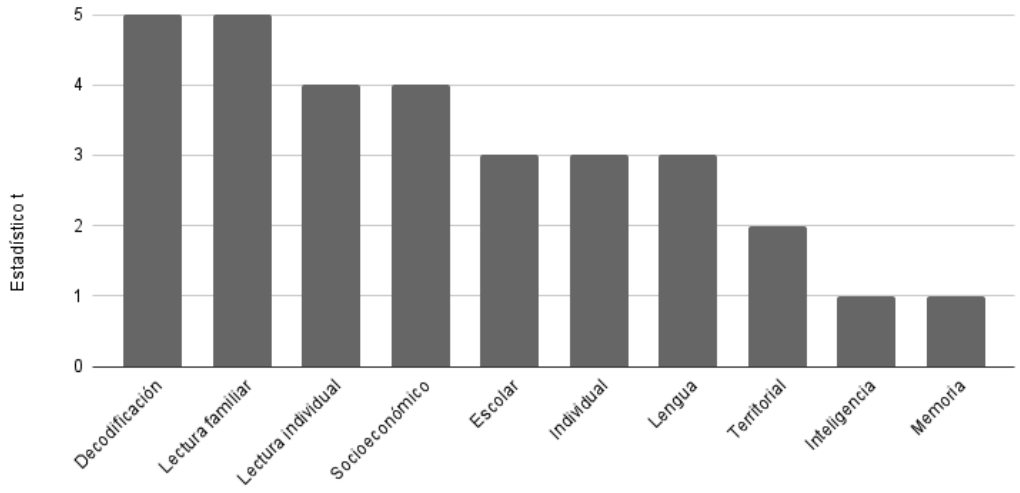

Figura 4.16: Gráfico de barras que muestra la importancia (medida a través del estadístico *t*), a la hora de predecir la variable *gradeMath*, de las variables en base al ámbito al que pertenecen, pero concretando un poco más cada ámbito. Las variables dominantes son las relacionadas con la decodificación, la lectura familiar, la lectura individual y el entorno socioeconómico.

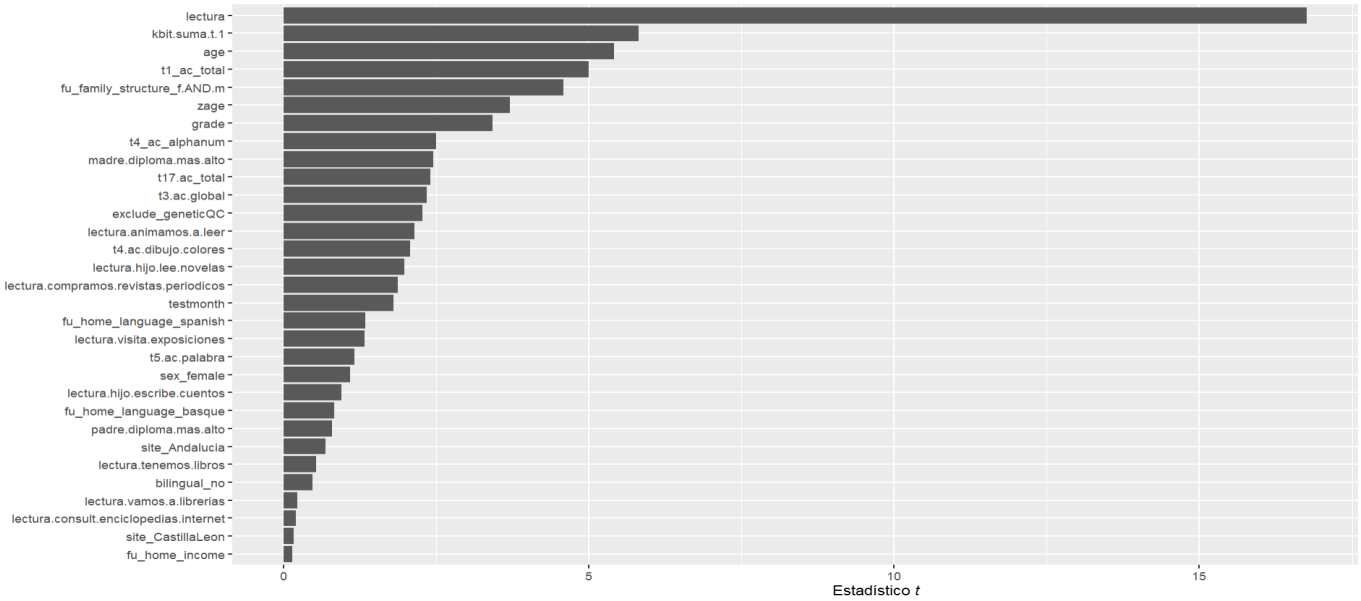

Figura 4.17: Gráfico de barras que muestra la importancia (medida a través del estadístico *t*) de cada variable a nivel individual. Es la variable *lectura*, sin lugar a dudas, la más importante para el modelo a la hora de predecir la variable *gradeMath*.

4.3.2.2. Feature effects

El impacto de las características o *feature effects* es una medida que indica cómo influye una variable predictora en la predicción de la variable dependiente [28].

En el caso de la regresión lineal, la influencia de una variable independiente se puede estimar mediante el coeficiente de regresión de dicha variable [21]. Dependiendo de qué tipo de variable (cuantitativa o nominal) estemos analizando, la interpretación del coeficiente de regresión será diferente. En el caso de las variables cuantitativas, un aumento de una unidad en la variable independiente x_k incrementa la predicción de y en β_k unidades, siempre y cuando mantengamos fijos los valores de todas las demás variables independientes. Sin embargo, en las variables nominales, al cambiar la variable independiente x_k de su categoría de referencia a otra categoría, se incrementa la predicción de y en β_k unidades, siempre y cuando mantengamos fijos los valores de todas las demás variables independientes [21]. Para comprender mejor esto que acabamos de explicar, hemos puesto un par de ejemplos en la tabla 4.7.

Tabla 4.7: Ejemplos de cómo se interpreta la influencia de una variable predictora en la predicción de la variable dependiente.

Caso	Explicación
Variable cuantitativa *age*	En el modelo que predice *gradeLanguage* la variable *age* tiene un coeficiente de regresión de -1.21 (ver gráfico 4.18). Esto implica que si incrementamos la variable *age* en una unidad (añadimos un año de edad al estudiante) y mantenemos el resto de predictores con el mismo valor que antes, a la predicción anterior de la nota *gradeLanguage* hay que restarle -1.21.
Variable nominal *sex_female*	En el modelo que predice *gradeLanguage* la variable *sex_female* tiene un coeficiente de regresión de 0.185 (ver gráfico 4.18). La variable referente es *sex_male*, por lo tanto, si cambiamos *sex_male* por *sex_female* (si cambiamos el género del estudiante de masculino a femenino) y mantenemos el resto de predictores con el mismo valor que antes, a la predicción de la nota anterior (habiendo predicho esa nota teniendo en cuenta que el estudiante tiene género masculino) hay que sumarle 0.185.

Con el objetivo de ser lo más precisos posible, hemos decidido seleccionar para cada modelo de regresión lineal únicamente las variables predictoras que contienen un coeficiente de regresión estadísticamente significativo ($p < 0.05$) y con ellos, dibujar un gráfico de puntos que muestra los coeficientes de regresión de cada variable independiente (ver figuras 4.18 y 4.19).

Predicción de *gradeLanguage*

En la figura 4.18 llama la atención la influencia que tiene el intercepto (que define la puntuación que obtendría el alumno si todas las demás variables predictoras fueran nulas) en la predicción final de la nota en lengua castellana. Además, los coeficientes de la mayoría de las predictoras son cercanos a 0, por lo que podemos intuir que en la predicción el que manda es el intercepto y que las demás predictoras sirven para calibrar y afinar la nota final. Entre estas variables, las que más destacan son *age* y *grade* con coeficientes cercanos a -1 y 1 respectivamente, y las variables *site_CastillaLeon*, *lectura*, *site_Andalucia* y *zage* con coeficientes cercanos a 0.5.

Como análisis general, podemos decir que entre todas las variables predictoras, las que más afectan en la predicción de la nota final (en lengua castellana) si aumentamos una unidad a su valor original, son la edad (*age*) y el curso (*grade*). Cuando estas dos variables se interpretan juntas, se obtienen las notas más bajas cuando confluyen edades altas con cursos bajos.

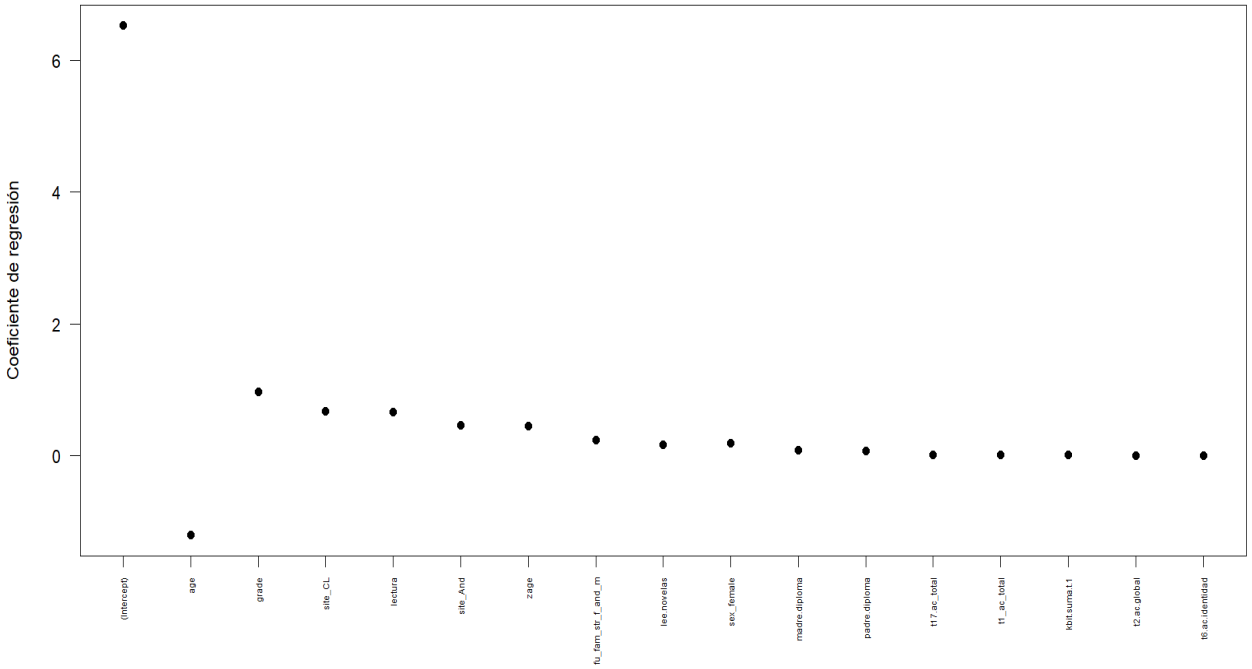

Figura 4.18: Gráfico de puntos que muestra los coeficientes de regresión de cada variable independiente en el modelo de regresión lineal múltiple (con selección de características) de *gradeLanguage*.

Predicción de *gradeMath*

Al contrario de lo que ocurre con *gradeLanguage*, tal y como podemos ver en la figura 4.19, el intercepto no es estadísticamente significativo (por ello no figura en el gráfico) y no tiene tanto peso en la predicción como antes (comparándolo con el intercepto de *gradeLanguage* su

coeficiente es 3 veces menor). Sin embargo, al igual que antes, la tendencia es que los coeficientes de las variables tiendan a ser cercanos a 0, salvo 4 que destacan sobre las demás: la variable *age* con un coeficiente muy cercano a -1 y las variables *grade*, *lectura*, *fu_family_structure_f.AND.m* y *zage*, que tienen un coeficiente cercano al 0.5.

La edad (*age*) y el curso (*grade*) siguen siendo las variables predictivas que aumentando una unidad a su valor original, más cambian el resultado de la predicción. Como trabajo futuro, podríamos analizar exactamente por qué ocurre esto estudiando cada grupo de edad o cada curso por separado.

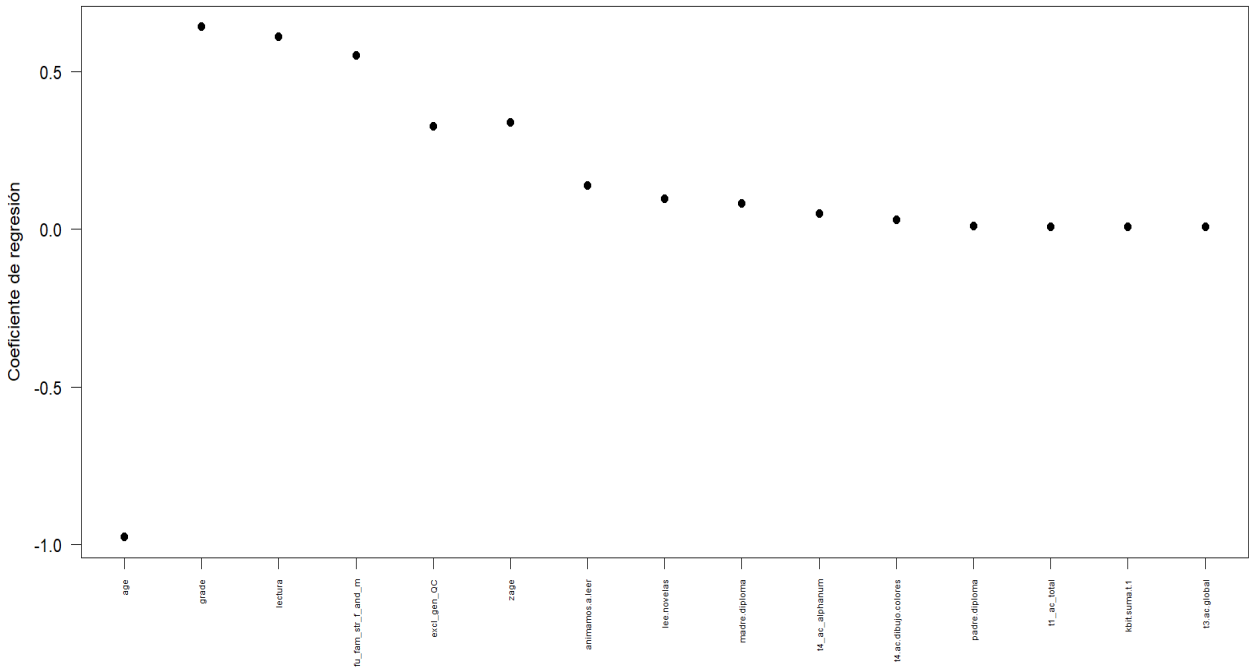

Figura 4.19: Gráfico de puntos que muestra los coeficientes de regresión de cada variable independiente en el modelo de regresión lineal múltiple (con selección de características) de *gradeMath*.

CAPÍTULO 5

Predicción de las variables *gradeLanguage* y *gradeMath* de manera simultánea

Hay una circulación común, una respiración común. Todas las cosas están relacionadas.

Hipócrates (460 a.C. - 370 a.C.)

5.1. Relación entre las variables *gradeLanguage* y *gradeMath*

Hasta ahora, hemos predicho las variables de salida de manera independiente; es decir, hemos entrenado dos modelos de aprendizaje automático diferentes para que cada uno de ellos se encargue específicamente de la predicción de una nota en concreto. Haciéndolo de esta manera, estamos actuando como si las dos notas fueran independientes. Sin embargo, la realidad es que estas dos variables tienen una correlación lineal positiva bastante fuerte. El coeficiente de correlación de *Pearson* entre ellas es de 0.75 y nos indica que a medida que aumenta una variable, es probable que la otra variable también aumente.

Esta correlación entre las dos variables dependientes nos hace pensar que si empleamos un modelo diferente para predecir cada nota, quizás estemos perdiendo información a la hora de hacer predicciones. De esta manera sólo estamos capturando las relaciones que existen entre las variables independientes y cada nota por separado; sin sacar provecho también de la relación entre las dos notas. Podemos intentar capturar no solo las relaciones entre las variables independientes y la variable dependiente correspondiente, sino que también la relación entre

las dos variables dependientes *gradeLanguage* y *gradeMath*. La solución a esto nos lo dan los modelos de regresión *multi-output* [29].

5.2. Modelos de regresión *multi-output*

La regresión *multi-output*, también conocida como regresión multivariada o multi-objetivo, tiene como objetivo predecir simultáneamente múltiples variables de salida de tipo numérico.

En diferentes estudios [30] [31], se ha demostrado que los métodos de regresión *multi-output* generan un mejor rendimiento predictivo en casos del mundo real, si los comparamos con los métodos tradicionales *single-output*. Al modelar aplicaciones del mundo real nos encontramos con problemas como la presencia de ruido debido a la complejidad de los dominios reales y sobre todo, la inherente naturaleza multivariada de dichas aplicaciones y las dependencias compuestas entre múltiples atributos y entre las múltiples variables de salida. Al contrario de lo que ocurre con los algoritmos *single-output*, los *multi-output* no solamente modelan los datos de entrenamiento teniendo en cuenta las relaciones entre los atributos y las de una única variable de salida, sino que también son capaces de capturar las relaciones entre las variables de salida, garantizando así un mejor modelado y representación de los problemas de la vida real.

Para entender exactamente lo que modelan los algoritmos de regresión *multi-output* [29], vamos a considerar el conjunto de datos de entrenamiento D de N instancias y asignaciones de valores para cada variable $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_d$. Quedando nuestro conjunto de datos de esta manera: $D = \{(\mathrm{x}^{(1)}, \mathrm{y}^{(1)}), \ldots, (\mathrm{x}^{(N)}, \mathrm{y}^{(N)})\}$. Cada instancia es representada por un vector de entrada de m características descriptivas $\mathrm{x}^{(l)} = (x_1^{(l)}, \ldots, x_j^{(l)}, \ldots, x_m^{(l)})$ y un vector de salida de d variables de salida $\mathrm{y}^{(l)} = (y_1^{(l)}, \ldots, y_i^{(l)}, \ldots, y_d^{(l)})$; con $i \in \{1, \ldots, d\}$, $j \in \{1, \ldots, m\}$ y $l \in \{1, \ldots, N\}$. La tarea consiste en aprender un modelo de regresión *multi-output* a partir de D, que consiste en encontrar una función h que asigna a cada instancia, dado el vector x, un vector y de d valores objetivo (ver figura 5.1). En dicha función Ω_{X_1} y Ω_{Y_1} representan los espacios muestrales de cada variable predictiva X_j, para $j \in \{1, \ldots, m\}$, y cada variable de salida Y_i, para $i \in \{1, \ldots, d\}$, respectivamente. Como consecuencia, el modelo de regresión *multi-output* aprendido, será capaz de predecir simultáneamente los valores $\{\hat{\mathrm{y}}^{(N+1)}, \ldots, \hat{\mathrm{y}}^{(N')}\}$ de todas las variables de salida para todas las nuevas instancias (no vistas en el entrenamiento) sin etiquetar $\{\hat{\mathrm{x}}^{(N+1)}, \ldots, \hat{\mathrm{x}}^{(N')}\}$.

$$\begin{aligned} h : \Omega_{X_1} \times \ldots \times \Omega_{X_m} &\rightarrow \Omega_{Y_1} \times \ldots \times \Omega_{Y_d} \\ \mathrm{x} = (x_1, \ldots, x_m) &\mapsto \mathrm{y} = (y_1, \ldots, y_d) \end{aligned} \tag{5.1}$$

Figura 5.1: Función que describe la tarea que debe aprender el modelo de regresión *multi-output*.

5.3. Modelos de regresión *multi-output* en *R* y *Python*

Al trabajar con el lenguaje de programación *Python* encontramos más facilidades para trabajar con modelos *multi-output*. De hecho, el lenguaje *R* no ofrece la posibilidad de implementar

el algoritmo de regresión lineal múltiple en su versión *multi-output*. En su lugar, disponemos del paquete *glmnet* que ofrece la opción de entrenar modelos lineales que permiten restringir y regularizar los coeficientes del modelo al introducir términos de penalización en la función objetivo que se optimiza durante el ajuste del modelo. Sin embargo, siguiendo el camino de los modelos fácilmente interpretables y por comparar los mismos modelos en versión *single-output* y *multiple-output*, hemos decidido seguir el estudio, al menos en este capítulo, con *Python*; que sí ofrece (en principio), gracias a la librería *scikit-learn* [32], tanto la regresión lineal múltiple como el algoritmo de árboles de clasificación y regresión en sus versiones *multi-output*.

Tal y como se describe en [29], actualmente existen diferentes métodos de regresión *multi-output*. Principalmente, se pueden dividir en dos grandes grupos: métodos de transformación del problema y métodos de adaptación del algoritmo. A continuación, desgranaremos cada método y explicaremos brevemente las características de cada uno:

1. **Transformación del problema**. Estos métodos se basan en la idea de transformar el problema de regresión *multi-output* en un problema *single-output*, después ajustar un modelo para cada variable dependiente y finalmente, concatenar todas las predicciones. Existen dos métodos de este tipo en *scikit-learn*:

    a) ***Single-target method***, llamado *MultiOutputRegressor* en *scikit-learn*, descompone el modelo *multi-output* en *d* (cantidad de variables dependientes) modelos *single-output* y entrena cada uno de ellos con una variable dependiente diferente. Así, las variables dependientes se predicen de manera independiente y no se es capaz de explotar las relaciones entre variables dependientes.

    b) ***Regressor chains***, llamado *RegressorChain* en *scikit-learn*, sí que es capaz de capturar las relaciones entre las variables dependientes. La idea es ir encadenando modelos *single-output* de manera lineal. Así, el primer modelo de la cadena es entrenado para predecir una variable dependiente. El segundo, es entrenado para predecir la segunda variable dependiente con los atributos empleados en el primer modelo más la predicción del primer modelo. El tercero, es entrenado para predecir la tercera variable dependiente con los atributos originales más las dos predicciones anteriores. Así, hasta llegar a predecir la última variable dependiente de la cadena. Es decir, siempre se usan los atributos del conjunto de datos de entrenamiento original más las predicciones de los modelos anteriores de la cadena para entrenar el siguiente modelo. El orden de la cadena, es decir, el orden en el que se van prediciendo las variables puede establecerse a través del parámetro *order*.

2. **Adaptación del algoritmo**. Estos métodos se basan en la idea de predecir simultáneamente todas las variables dependientes utilizando un solo modelo que sea capaz de capturar todas las dependencias internas entre ellas. En *scikit-learn* solo existe un método:

    a) **Algoritmos que inherentemente aceptan la versión *multi-output***. *Scikit-learn* ofrece un grupo selecto de estos algoritmos. Entre ellos, se incluyen los que nos interesan: la regresión lineal múltiple y los árboles de clasificación y regresión. Sin embargo, a pesar de que aparentemente la regresión lineal múltiple crea un

> único modelo para predecir simultáneamente las diferentes variables de salida, al mostrar los coeficientes de entrenamiento vemos que realmente modela unos coeficientes de regresión distintos para cada variable dependiente. Por lo tanto, aunque sí que acepta inherentemente múltiples variables de salida como etiquetas en el entrenamiento [33], parece que realmente está tratando cada variable objetivo de manera independiente. En principio, ocurre lo mismo con los árboles de clasificación y regresión.

Por la falta de claridad (tanto en la documentación de *scikit-learn* como en otros artículos de internet [34]) en cuanto a si los modelos que aceptan inherentemente conjuntos de datos *multi-output* capturan las relaciones entre las variables dependientes; hemos decidido actuar con prudencia y emplear el método *RegressorChain* que sí que nos asegura capturar a parte de las relaciones entre atributos y variables objetivo, las interdependencias entre las variables objetivo. De hecho, en la documentación de *scikit-learn* donde se habla específicamente sobre los modelos *multi-output* [32], comenta explícitamente el uso del método *RegressorChain* para también capturar las relaciones entre las variables objetivo.

5.4. Entrenamiento de los modelos

Al no haber algo similar a la función *train()* de *caret* en *Python*, que ajusta automáticamente los hiperparámetros de los modelos que queremos entrenar, para una aproximación inicial de los rendimientos de los modelos, hemos empleado los hiperparámetros que emplean for defecto los modelos de la librería *scikit-learn*; sin validación cruzada, únicamente dividiendo el *dataset* en datos de entrenamiento (80 %) y datos de prueba (20 %). Asimismo, por ser lo más justos posible y no encontrarnos con que el mismo modelo entrenado en *R* y *Python* obtiene diferentes resultados, hemos decidido hacer todas las pruebas (*single-output* y *multi-output*) de este capítulo en *Python*.

Para poder hacer comparaciones entre los rendimientos de los modelos *single-output* y *multiple-output*, primero hemos entrenado los mismos modelos *single-output* del anterior capítulo mediante la librería *scikit-learn*: la regresión lineal múltiple con *LinearRegression()* y los árboles de clasificación y regresión con *DecisionTreeRegressor()*. Al igual que en el capítulo anterior, hemos entrenado un modelo por cada variable dependiente.

Una vez entrenados los modelos *single-output* de referencia, hemos procedido al entrenamiento de los modelos *multi-output* mediante la técnica *regressor chains*. El problema principal de este método es que es sensible al orden seleccionado de las cadenas. En [35] se hace esta propuesta: si el número de cadenas distintas es menor que 10, crear exactamente tantos modelos como el número de cadenas distintas. De lo contrario, seleccionar aleatoriamente 10 cadenas. Nuestro caso es el primero. Al tener dos variables dependientes, hemos creado dos cadenas tanto para el modelo de regresión lineal múltiple como para el de árboles de clasificación y regresión:

1. Predecir primero *gradeLanguage* y seguidamente, añadir ésta como atributo en la predic-

ción de *gradeMath*.

2. Predecir primero *gradeMath* y seguidamente, añadir ésta como atributo en la predicción de *gradeLanguage*.

En cuanto al cálculo del rendimiento de los modelos se refiere, la única diferencia respecto al capítulo anterior es que para los modelos *multi-output* se calcula por defecto el promedio de la raíz del error cuadrático medio (*aRMSE*) [29]. El concepto es el mismo que el de *RMSE* en los modelos *single-output*, con la diferencia de que primero se calculan por separado las raíces del error cuadrático medio de cada variable dependiente y después se calcula el promedio sumándolas todas y dividiéndolas por la cantidad de variables dependientes (ver figura 5.2). Sin embargo, como nuestro objetivo es comparar el rendimiento de estos modelos *multi-output* con los *single-output*, hemos empleado la propia raíz del error cuadrático medio (*RMSE*) que el modelo *multi-output* obtiene para cada variable dependiente.

$$\text{aRMSE} = \frac{1}{d}\sum_{i=1}^{d}\sqrt{\frac{1}{N_{test}}\sum_{l=1}^{N_{test}}\left(y_i^{(l)} - \hat{y}_i^{(l)}\right)^2} \tag{5.2}$$

Figura 5.2: Fórmula del promedio de la raíz del error cuadrático medio.

5.5. Resultados

Las tablas 5.1 y 5.2 muestran los resultados de los modelos *single-output* y *multi-output* respectivamente, obtenidos tras entrenarlos de la manera explicada en el apartado anterior.

Tabla 5.1: Resultados de los modelos *single-output* empleados para predecir individualmente *gradeLanguage* y *gradeMath* en *Python*.

Modelos *single-output*	**RMSE**	
	gradeLanguage	**gradeMath**
Regresión lineal múltiple	1.158708	1.280899
Árboles de clasificación y regresión	1.663235	1.835581

Tabla 5.2: Resultados de los modelos *multiple-output* empleados para predecir simultáneamente *gradeLanguage* y *gradeMath* en *Python*. Nota: aunque parezca que los resultados de los modelos *multi-output* de la regresión lineal múltiple son exactamente iguales a los de los modelos *single-output* de regresión lineal, no lo son. A partir del decimal número 13 cambian.

Modelos *multi-output*	**Cadena**	**RMSE**	
		gradeLanguage	**gradeMath**
Regresión lineal múltiple	(*gradeLanguage, gradeMath*)	1.158708	1.280899
	(*gradeMath, gradeLanguage*)	1.158708	1.280899
Árboles de clasificación y regresión	(*gradeLanguage, gradeMath*)	1.613462	1.827308
	(*gradeMath, gradeLanguage*)	1.763353	1.838757

Tal y como ocurría en el capítulo 4, los modelos de regresión lineal siguen siendo los más precisos. A su vez, sorprende de primeras, que a penas hay diferencia alguna (difieren a partir del decimal 13) entre los modelos de regresión lineal *single-output* y *multi-output*. Una posible explicación puede ser que al tener ambas variables objetivo una correlación lineal alta y al modelar la regresión lineal relaciones lineales, puede que ocurran problemas de multicolinearidad y redundancia y que realmente no se llegan a capturar o directamente el modelo no hace caso a las relaciones entre ambas asignaturas.

En cuanto al algoritmo de árboles de clasificación y regresión, a pesar de obtener peores resultados que los algoritmos de regresión lineal, sí que se obtiene en un caso en concreto, con la cadena (*gradeLanguage, gradeMath*), mejores resultados en el modelo *multi-output* que en los modelos *single-output*. De todas formas, la diferencia no es significativa y podemos decir que en términos generales los modelos *multi-output* no mejoran los resultados de los modelos *single-output*.

CAPÍTULO 6

Conclusiones y próximos pasos

La vida es el arte de sacar conclusiones suficientes a partir de datos insuficientes.

Samuel Butler (1835 - 1902)

6.1. Conclusiones del estudio

A lo largo de estas páginas hemos querido indagar sobre los factores que afectan al rendimiento escolar. Concretamente, hemos estudiado qué tipo de características o atributos tienen importancia a la hora de predecir las notas en lengua castellana y en matemáticas.

Desde un principio suponíamos que estabamos frente a un problema multifactorial. Son muchas las circunstancias que pueden afectar a que un estudiante saque una nota u otra. De todas ellas, hemos seleccionado atributos que describen diferentes características tanto del estudiante a nivel individual como a nivel de su entorno: variables socioeconómicas, de hábitos de lectura, territoriales, escolares, de procesamiento cognitivo, etc. Con todas ellas, hemos conseguido capturar cerca del 50 % de la varianza total de las notas de lengua castellana y matemáticas. Esto implica, que hemos sido capaces de predecir la nota de lengua con un margen de error promedio de aproximadamente 1.16 puntos y la de matemáticas con un error promedio aproximado a 1.36 puntos. En una escala del 0 al 10 fallar en un rango de entre 1 y 1.5 puntos no parece descabellado, pero tenemos que tener en cuenta que estamos dejando de capturar cerca de la mitad de la varianza total de ambas notas. Es un porcentaje bastante alto.

Hemos tratado de mejorar estos resultados predeciendo ambas variables objetivo simultáneamente, pero los resultados han sido muy similares. Resulta curioso, pero podríamos decir que el eslogan de este estudio ha sido: *no compliques las cosas, lo simple funciona suficientemente bien.* Y es que, han sido los algoritmos más sencillos los más efectivos. Ni los algoritmos *multi-output*,

ni los algoritmos que capturan relaciones más complejas que las lineales han podido superar al modelo *single-output* de regresión lineal. Parece ser que las relaciones lineales entre los atributos seleccionados y las dos asignaturas son suficientes para capturar patrones y relaciones entre ellas.

Sin embargo, no todas las variables han tenido la misma importancia a la hora de modelar dichas relaciones. Sin lugar a dudas, las variables relacionadas con los hábitos/entorno de lectura y las relacionadas con procesos cognitivos han sido las más importantes a la hora de hacer las predicciones. Especialmente, el entorno familiar (variables ambientales) e individual de lectura, cómo de bien lee el estudiante y aspectos relacionados a la decodificación. Con todo esto, podemos concluir que el sistema educativo actual está canalizado a través de la lectura y que por lo tanto, cómo de bien lee un estudiante será importante para su rendimiento escolar.

6.2. Próximos pasos

El trabajo realizado ha sido una aproximación inicial al estudio de algoritmos de aprendizaje automático que ajustan el conjunto de datos seleccionado a las notas en las asignaturas de lengua castellana y matemáticas. Al ser una aproximación cuenta con ciertas limitaciones y posibles mejoras que pueden llevarse a cabo en el futuro para continuar con esta línea de investigación. Algunas de ellas se mencionan en los siguientes puntos:

- Principalmente nos hemos limitado a estudiar modelos que son inherentemente interpretables. Desde un principio hemos dejado fuera algoritmos como los del *deep learning* por el mero hecho de no ser interpretables de por sí. Sin embargo, actualmente existen muchos estudios acerca de la interpretabilidad en el aprendizaje profundo y hay disponibles métodos de interpretación agnósticos para los modelos que consideramos *black boxes* [21]. Con ellos, podríamos ampliar el abanico de modelos de aprendizaje automático y seguir cumpliendo el objetivo de analizar qué variables son importantes a la hora de hacer predicciones.
- Aprovechando que ahora no vamos *a ciegas* y que sabemos qué tipo de variables han sido útiles para capturar diferentes patrones, podríamos ampliar nuestra muestra e incidir aún más en las variables que han resultado importantes; ya sea ampliando el espectro o especificándolas aún más.
- A pesar de que los modelos *multi-output* no han cumplido con las expectativas que teníamos, sí que resulta interesante tener un único modelo de *machine learning* con el que predecir ambas asignaturas simultáneamente, en lugar de tener que ir predeciéndolas de una en una con diferentes modelos. Debido a las limitaciones en el tiempo, no se ha podido indagar lo suficientemente bien los métodos de *scikit-learn* que sí aceptan el entrenamiento de conjuntos de datos con múltiples variables de salida. Podríamos dedicar más tiempo al análisis de cómo estiman exactamente las predicciones y si realmente capturan también las relaciones entre las variables objetivo; y en el caso de hacerlo, cómo y qué tipo de relaciones buscan.

Apéndices

APÉNDICE A

Variables del conjunto de datos

Tabla A.1: Tabla que define todas las variables del conjunto de datos original.

Variable	Descripción	Tipo de variable	Ámbito de estudio
id	Identificador del estudiante.	Cualitativa nominal	Individual
sex	Género del estudiante (masculino o femenino).	Cualitativa nominal	Individual
age	Edad del estudiante computada de esta manera: la fecha del test menos la fecha de nacimiento.	Cuantitativa continua	Individual
testmonth	Mes en el que fueron hechos los tests.	Cuantitativa discreta	Escolar
school	Escuela en la que estudia el estudiante.	Cualitativa nominal	Escolar
grade	Curso en el que estudia el estudiante.	Cualitativa ordinal	Escolar

exclude	Describe si el estudiante es un *outlier* basándose en los análisis genéticos de sus ancestros.	Cualitativa nominal	Individual
site	Lugar en el que se hicieron los tests.	Cualitativa nominal	Territorial
province	Provincia de origen del estudiante.	Cualitativa nominal	Territorial
country	País de origen del estudiante.	Cualitativa nominal	Territorial
bilingual	Describe si el estudiante es bilingüe: vasco, inglés o no.	Cualitativa nominal	Lengua hablada
fu_family_structure	Describe la estructura familiar: si está compuesta por madre y padre o solo padre/madre.	Cualitativa nominal	Socioeconómica
fu_home_income	Ingresos del hogar.	Cualitativa ordinal	Socioeconómica
madre.ocupacion.isco	Código ISCO-88 [9] para la ocupación de la madre.	Cuantitativa discreta	Socioeconómica
padre.ocupacion.isco	Codigo ISCO-88 [9] para la ocupación del padre.	Cuantitativa discreta	Socioeconómica
madre.ocupacion.hsbc	Código HSBC para la educación de la madre.	Cuantitativa discreta	Socioeconómica
padre.ocupacion.hsbc	Código HSBC para la educación del padre.	Cuantitativa discreta	Socioeconómica
madre.diploma.mas.alto	Título educativo más alto de la madre.	Cualitativa ordinal	Socioeconómica

padre.diploma.mas.alto	Título educativo más alto del padre.	Cualitativa ordinal	Socioeconómica
lectura.compramos.revistas.periodicos	Cuantifica la frecuencia con la que la familia compra revistas y periódicos (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura familiar
lectura.tenemos.libros	Cuantifica la frecuencia con la que la familia tiene libros en casa (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura familiar
lectura.vamos.a.librerias	Cuantifica la frecuencia con la que la familia acude a librerías (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura familiar
lectura.animamos.a.leer	Cuantifica la frecuencia con la que la familia anima a leer a sus hijos (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura familiar
lectura.consult.enciclopedias.internet	Cuantifica la frecuencia con la que el estudiante consulta internet o enciclopedias (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura individual
lectura.visita.exposiciones	Cuantifica la frecuencia con la que la familia visita exposiciones (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura familiar

lectura.hijo.lee.novelas	Cuantifica la frecuencia con la que el estudiante lee novelas (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura individual
lectura.hijo.escribe.cuentos	Cuantifica la frecuencia con la que el estudiante escribe cuentos (nunca < a veces < casi siempre < siempre).	Cualitativa ordinal	Lectura individual
lectura	Valoración (del 0 al 6) por parte del profesor del nivel de lectura del estudiante.	Cuantitativa discreta	Lectura individual
fu_home_language	Idioma principal que se habla en casa.	Cualitativa nominal	Lengua hablada
fu_home_language_other	Idioma secundario que se habla en casa.	Cualitativa nominal	Lengua hablada
grupo	Grupo/clase de la escuela a la que pertenece el estudiante.	Cuantitativa nominal	Escolar
gradeLanguage	Nota que el estudiante ha obtenido en la asignatura de lengua castellana.	Cuantitativa continua/Cualitativa ordinal	Escolar
gradeMath	Nota que el estudiante ha obtenido en la asignatura de matemáticas.	Cuantitativa continua/Cualitativa ordinal	Escolar
zage	Edad estandarizada por curso.	Cuantitativa continua	Escolar

kbit.pc.matrices	Percentil en el que se encuentra el estudiante en la prueba de matrices *KBIT* (*Kaufman Brief Intelligence Test*) correspondiente a la inteligencia no verbal.	Cuantitativa discreta	Cognitiva-Inteligencia
kbit.suma.t.1	Suma de las puntuaciones tipificadas inteligencia verbal y no verbal en la prueba de matrices *KBIT* (*Kaufman Brief Intelligence Test*) que ha obtenido el estudiante.	Cuantitativa discreta	Cognitiva-Inteligencia
t1_ac_total	Precisión del estudiante en la tarea de eliminación de fonemas.	Cuantitativa continua	Cognitiva-Decodificación
t1_rt_total	Tiempo de reacción del estudiante en la tarea de eliminación de fonemas.	Cuantitativa discreta	Cognitiva-Decodificación
t10.ac.incong_menos_cong	Precisión del estudiante en la tarea del efecto *Stroop*.	Cuantitativa continua	Cognitiva-Atención
t10.tr.incong_menos_cong	Tiempo de reacción del estudiante en la tarea del efecto *Stroop*.	Cuantitativa continua	Cognitiva-Atención
t16.ac_spanfinal	Precisión del estudiante en la tarea de la memoria visoespacial.	Cuantitativa discreta	Cognitiva-Memoria
t17.ac_threetask	Precisión del estudiante en la tarea *3-back N-back*.	Cuantitativa discreta	Cognitiva-Memoria
t17.ac_total	Precisión total del estudiante en las tareas *(1,2,3)-N-back*.	Cuantitativa discreta	Cognitiva-Memoria

t2.ac.global	Precisión del estudiante (para todos los tiempos de reacción) en la tarea de coincidencia de fonemas.	Cuantitativa discreta	Cognitiva-Decodificación
t2.tr.global	Tiempo de reacción del estudiante en la tarea de coincidencia de fonemas.	Cuantitativa discreta	Cognitiva-Decodificación
t20_indice.congruencia	Tiempo de reacción de conflicto del estudiante en la Prueba de Red Atencional (*ANT*).	Cuantitativa continua	Cognitiva-Atención
t20_indice.congruenciaerrores	Errores de conflicto del estudiante en la Prueba de Red Atencional (*ANT*).	Cuantitativa continua	Cognitiva-Atención
t3.ac.global	Precisión del estudiante en la tarea de identificación de sílabas.	Cuantitativa continua	Cognitiva-Decodificación
t3.tr.global	Tiempo de reacción del estudiante en la tarea de identificación de sílabas.	Cuantitativa discreta	Cognitiva-Decodificación
t4.ac.dibujo.colores	Precisión del estudiante en la tarea de Denominación Automatizada Rápida (*RAN*) (no alfanuméricos).	Cuantitativa discreta	Cognitiva-Decodificación
t4.tr.dibujo.colores	Tiempo de reacción del estudiante en la tarea de Denominación Automatizada Rápida (*RAN*) (no alfanuméricos).	Cuantitativa discreta	Cognitiva-Decodificación

t4_ac_alphanum	Precisión del estudiante en la tarea de Denominación Automatizada Rápida (*RAN*) (alfanuméricos).	Cuantitativa discreta	Cognitiva-Decodificación
t4_rt_alphanum	Tiempo de reacción del estudiante en la tarea de Denominación Automatizada Rápida (*RAN*) (alfanuméricos).	Cuantitativa discreta	Cognitiva-Decodificación
t5.ac.palabra	Precisión del estudiante en la tarea de lectura de palabras.	Cuantitativa continua	Cognitiva-Decodificación
t5.ac.pseudopalabra	Precisión del estudiante en la tarea de lectura de pseudopalabras o palabras inexistentes.	Cuantitativa continua	Cognitiva-Decodificación
t5.tr.palabra	Tiempo de reacción del estudiante en la tarea de lectura de palabras.	Cuantitativa discreta	Cognitiva-Decodificación
t5.tr.pseudopalabra	Tiempo de reacción del estudiante en la tarea de lectura de pseudopalabras o palabras inexistentes.	Cuantitativa discreta	Cognitiva-Decodificación
t6.ac.identidad	Precisión del estudiante en la tarea de Efecto Identidad de Letras (*Letter Identity Effect).*	Cuantitativa continua	Cognitiva-Decodificación
t6.ac.similitud	Precisión del estudiante en la tarea de Efecto de Similitud de Letras (*Letter Similarity Effect).*	Cuantitativa continua	Cognitiva-Decodificación

t6.tr.identidad	Tiempo de reacción del efecto de identidad de letras del estudiante en la tarea Identificación de letras - Identidad.	Cuantitativa continua	Cognitiva-Decodificación
t6.tr.similitud	Tiempo del estudiante en la condición "Similitud" de la tarea de Identificación de letras Identidad.	Cuantitativa continua	Cognitiva-Decodificación
t7.ac.efecto.tl	Tiempo de reacción del efecto de transposición de letras en la tarea Identificación de letras - Posición.	Cuantitativa continua	Cognitiva-Decodificación
t7.tr.efecto.tl	Tiempo de reacción de la condición "Transposición" en la tarea Identificación de letras - Posición.	Cuantitativa continua	Cognitiva-Decodificación
t8_ac_ps_word	Precisión del estudiante en la tarea de repetición de pseudopalabras.	Cuantitativa continua	Cognitiva-Decodificación
t8_tr_ps.word	Tiempo de reacción del estudiante en la tarea de repetición de pseudopalabras.	Cuantitativa discreta	Cognitiva-Decodificación
t9_ac_index_Incongruent_minus_Congruent	Precisión del estudiante en la tarea del Efecto Stroop Numérico (*Numeric Stroop Effect*).	Cuantitativa continua	Cognitiva-Atención

t9_rt_index_Incongruent_minus_Congruent	Tiempo de reacción del estudiante en la tarea del Efecto *Stroop* Numérico (*Numeric Stroop Effect*).	Cuantitativa continua	Cognitiva-Atención

APÉNDICE B

Imputación: relaciones entre variables

B.1. Relación grande

Tabla B.1: Tabla que muestra para cada variable que debemos imputar, las variables con las que comparte una relación grande.

Variable a imputar	Variables con las que comparte una relación grande
testmonth	*school, province*
province	*school, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.vamos.a.librerias, lectura.visita.exposiciones*
fu_family_structure	*school, province, fu_home_income*
fu_home_income	*school, site, province, fu_family_structure, madre.diploma.mas.alto, padre.diploma.mas.alto*
madre.diploma.mas.alto	*school, site, province, fu_home_income, padre.diploma.mas.alto*
padre.diploma.mas.alto	*school, site, fu_home_income, madre.diploma.mas.alto*
lectura.compramos.revistas.periodicos	*school, site, province*
lectura.tenemos.libros	*school*
lectura.vamos.a.librerias	*school, province*

lectura.animamos.a.leer	*school*
lectura.consult.enciclopedias.internet	*school*
lectura.visita.exposiciones	*school, province*
lectura.hijo.lee.novelas	*school*
lectura.hijo.escribe.cuentos	*school*
fu_home_language	*fu_home_language_other*
zage	*school*
kbit.pc.matrices	*school, kbit.suma.t.1*
kbit.suma.t.1	*kbit.pc.matrices*
t1_ac_total	*age, grade, t4.tr.dibujo.colores, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra*
t10.ac.incong_menos_cong	*t4.tr.dibujo.colores, t5.tr.palabra, t5.tr.pseudopalabra*
t2.ac.global	*school, province*
t2.tr.global	*age, grade, t4.tr.dibujo.colores*
t4.tr.dibujo.colores	*age, school, grade, t1_ac_total, t10.ac.incong_menos_cong, t2.tr.global, t4_rt_alphanum, t5.tr.palabra, t5.tr.pseudopalabra*
t4_rt_alphanum	*age, school, grade, t4.tr.dibujo.colores, t5.tr.palabra, t5.tr.pseudopalabra*
t5.ac.palabra	*t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra*
t5.ac.pseudopalabra	*grade, t1_ac_total, t5.ac.palabra, t5.tr.palabra, t5.tr.pseudopalabra*
t5.tr.palabra	*age, grade, t1_ac_total, t10.ac.incong_menos_cong, t4.tr.dibujo.colores, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.pseudopalabra*
t5.tr.pseudopalabra	*age, grade, t1_ac_total, t10.ac.incong_menos_cong, t4.tr.dibujo.colores, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra*
t8_tr_ps.word	*school*

B.2. Relación media

Tabla B.2: Tabla que muestra para cada variable que debemos imputar, las variables con las que comparte una relación media.

Variable a imputar	Variables con las que comparte una relación media
testmonth	*site*
province	*sex, grade, lectura.tenemos.libros, lectura.animamos.a.leer, lectura.hijo.lee.novelas, fu_home_language_other*
fu_family_structure	*lectura.hijo.lee.novelas*
lectura.tenemos.libros	*province*
lectura.vamos.a.librerias	*lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas*
lectura.animamos.a.leer	*province, lectura.consult.enciclopedias.internet, lectura.hijo.lee.novelas*
lectura.consult.enciclopedias.internet	*lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.visita.exposiciones*
lectura.visita.exposiciones	*lectura.vamos.a.librerias, lectura.consult.enciclopedias.internet, lectura.hijo.lee.novelas*
lectura.hijo.lee.novelas	*province, fu_family_structure, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.visita.exposiciones, lectura.hijo.escribe.cuentos*
lectura.hijo.escribe.cuentos	*lectura.hijo.lee.novelas*
lectura	*school, kbit.suma.t.1*
kbit.pc.matrices	*province*
kbit.suma.t.1	*school, lectura*
t1_ac_total	*school, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.tr.global, t4_rt_alphanum, t5.ac.palabra, t8_ac_ps.word*
t10.ac.incong_menos_cong	*age, school, grade, t1_ac_total, t10.tr.incong_menos_cong, t2.tr.global, t3.tr.global, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra*
t10.tr.incong_menos_cong	*age, school, grade, t1_ac_total, t10.ac.incong_menos_cong, t2.tr.global, t4.tr.dibujo.colores, t4_rt_alphanum, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra*

t17.ac_total	*age, school, grade, t1_ac_total, t2.tr.global, t4.tr.dibujo.colores, t4_rt_alphanum, t5.tr.palabra, t5.tr.pseudopalabra*
t2.ac.global	*grade, t1_ac_total, t4.tr.dibujo.colores*
t2.tr.global	*school, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t3.tr.global, t4_rt_alphanum, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra*
t3.tr.global	*age, school, grade, t1_ac_total, t10.ac.incong_menos_cong, t2.tr.global, t4.tr.dibujo.colores, t4_rt_alphanum, t5.tr.palabra, t5.tr.pseudopalabra*
t4.tr.dibujo.colores	*t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t3.tr.global, t5.ac.palabra, t5.ac.pseudopalabra*
t4_ac_alphanum	*t5.ac.pseudopalabra*
t4_rt_alphanum	*t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.tr.global, t3.tr.global, t5.ac.palabra, t5.ac.pseudopalabra*
t5.ac.palabra	*age, school, grade, t1_ac_total, t10.ac.incong_menos_cong, t4.tr.dibujo.colores, t4_rt_alphanum*
t5.ac.pseudopalabra	*age, school, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t2.tr.global, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t8_ac_ps.word*
t5.tr.palabra	*school, t10.tr.incong_menos_cong, t17.ac_total, t2.tr.global, t3.tr.global*
t5.tr.pseudopalabra	*school, t10.tr.incong_menos_cong, t17.ac_total, t2.tr.global, t3.tr.global*
t8_ac_ps.word	*age, school, grade, t1_ac_total, t5.ac.pseudopalabra*
t8_tr_ps.word	*province*

B.3. Relación pequeña

Tabla B.3: Tabla que muestra para cada variable que debemos imputar, las variables con las que comparte una relación pequeña.

Variable a imputar	Variables con las que comparte una relación pequeña
testmonth	*sex, age, grade, exclude, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
province	*exclude, lectura.consult.enciclopedias.internet, fu_home_language*
fu_family_structure	*sex, grade, exclude, site, bilingual, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.escribe.cuentos, fu_home_language, fu_home_language_other, fu_family_structure*

fu_home_income	*sex, age, testmonth, grade, exclude, bilingual, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
madre.diploma.mas.alto	*sex, age, testmonth, grade, exclude, bilingual, fu_family_structure, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

padre.diploma.mas.alto	*sex, age, testmonth, grade, exclude, bilingual, fu_family_structure, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
lectura.compramos.revistas.periodicos	*sex, age, testmonth, grade, exclude, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

lectura.tenemos.libros	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
lectura.vamos.a.librerias	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.animamos.a.leer, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

lectura.animamos.a.leer	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.visita.exposiciones, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
lectura.visita.exposiciones	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.animamos.a.leer, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

lectura.hijo.lee.novelas	*sex, age, testmonth, grade, exclude, site, bilingual, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.consult.enciclopedias.internet, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
lectura.hijo.escribe.cuentos	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

lectura	*sex, age, testmonth, grade, exclude, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
fu_home_language	*sex, school, grade, exclude, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, fu_home_language*
zage	*sex, age, testmonth, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

kbit.pc.matrices	*sex, age, testmonth, grade, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
kbit.suma.t.1	*sex, age, testmonth, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, fu_home_language, fu_home_language_other, zage, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t1_ac_total	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t10.ac.incong_menos_cong	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t17.ac_total, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t10.tr.incong_menos_cong	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t17.ac_total, t2.ac.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4_ac_alphanum, t5.ac.palabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent, t10.tr.incong_menos_cong*
t17.ac_total	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t2.ac.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4_ac_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t2.ac.global	*sex, ge, testmonth, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t2.tr.global	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t5.ac.palabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent, t2.tr.global*

t3.ac.global	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t3.tr.global	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t5.ac.palabra, t5.ac.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t4.ac.dibujo.colores	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t4.tr.dibujo.colores	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other. zage, kbit.pc.matrices, kbit.suma.t.1, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t4_ac_alphanum	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_rt_alphanum, t5.ac.palabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent, t4_ac_alphanum*
t4_rt_alphanum	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t5.ac.palabra	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.globa, t3.tr.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t5.ac.pseudopalabra	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t17.ac_total, t2.ac.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t5.tr.palabra	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t5.tr.pseudopalabra	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t2.ac.global, t3.ac.global, t4.ac.dibujo.colores, t4_ac_alphanum, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t6.ac.identidad	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t6.ac.identidad	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t6.tr.identidad	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t6.tr.similitud	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t7.ac.efecto.tl	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.similitud, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t7.tr.efecto.tl	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.similitud, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t7.tr.efecto.tl	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.similitud, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*
t8_ac_ps.word	*sex, testmonth, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t8_tr_ps.word	*sex, age, testmonth, grade, site, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9º_rt_index_Incongruent_minus_Congruent*
t9_ac_index_Incongruent_minus_Congruent	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

t9_rt_index_Incongruent_minus_Congruent	*sex, age, testmonth, school, grade, site, province, bilingual, fu_family_structure, fu_home_income, madre.diploma.mas.alto, padre.diploma.mas.alto, lectura.compramos.revistas.periodicos, lectura.tenemos.libros, lectura.vamos.a.librerias, lectura.animamos.a.leer, lectura.consult.enciclopedias.internet, lectura.visita.exposiciones, lectura.hijo.lee.novelas, lectura.hijo.escribe.cuentos, lectura, fu_home_language, fu_home_language_other, zage, kbit.pc.matrices, kbit.suma.t.1, t1_ac_total, t10.ac.incong_menos_cong, t10.tr.incong_menos_cong, t17.ac_total, t2.ac.global, t2.tr.global, t3.ac.global, t3.tr.global, t4.ac.dibujo.colores, t4.tr.dibujo.colores, t4_ac_alphanum, t4_rt_alphanum, t5.ac.palabra, t5.ac.pseudopalabra, t5.tr.palabra, t5.tr.pseudopalabra, t6.ac.identidad, t6.ac.similitud, t6.tr.identidad, t6.tr.similitud, t7.ac.efecto.tl, t7.tr.efecto.tl, t8_ac_ps.word, t8_tr_ps.word, t9_ac_index_Incongruent_minus_Congruent, t9_rt_index_Incongruent_minus_Congruent*

Bibliografía

[1] Alice Miller. *El drama del niño dotado*. Tusquets Editores, 2020. Ver página 1.

[2] Jacqueline Lettau. The impact of children's academic competencies and school grades on their life satisfaction: What really matters? *Child Indicators Research*, 14(6):2171–2195, 2021. Ver página 1.

[3] Lex Borghans, Bart Golsteyn, James J. Heckman, and John Eric Humphries. What Grades and Achievement Tests Measure. Working Papers 2016-022, Human Capital and Economic Opportunity Working Group, November 2016. Ver página 1.

[4] D. Goleman. *Emotional Intelligence*. A Bantam book. Bantam Books, 1995. Ver páginas 1, 2.

[5] H. Gardner. *Intelligence Reframed: Multiple Intelligences For The 21st Century*. Basic Books, 1999. Ver página 2.

[6] K Davis, JA Christodoulou, S Seider, and H. Gardner. *The Theory of Multiple Intelligences*, pages 485–503. Cambridge University Press, New York, 2011. Ver página 2.

[7] J.W. Tukey. *Exploratory Data Analysis*. Number v. 2 in Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, 1977. Ver página 15.

[8] Mirian Sánchez-Morán, Juan Andrés Hernández, Jon Andoni Duñabeitia, Adelina Estévez, Laura Bárcena, Aintzane González-Lahera, María Teresa Bajo, Luis J. Fuentes, Ana M. Aransay, and Manuel Carreiras. Genetic association study of dyslexia and adhd candidate genes in a spanish cohort: Implications of comorbid samples. *PLOS ONE*, 13(10):1–17, 10 2018. Ver página 15.

[9] International Labour Organization. International standard classification of occupations. Ver páginas 20, 22, and 62.

[10] Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data,*. John Wiley & Sons, 1987. Ver páginas 23, 24.

[11] Edgar Acuña and Caroline Rodriguez. The treatment of missing values and its effect on classifier accuracy. In David Banks, Frederick R. McMorris, Phipps Arabie, and Wolfgang Gaul, editors, *Classification, Clustering, and Data Mining Applications*, pages 639–647, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. Ver página 25.

[12] Julián Luengo, Salvador García, and Francisco Herrera. A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfns and eventcovering method. *Neural Networks*, 23(3):406–418, 2010. Ver página 25.

[13] RDocumentation. mice: Multivariate imputation by chained equations. Ver página 25.

[14] Google Cloud. What is machine learning? Ver página 31.

[15] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229, jul 1959. Ver página 31.

[16] Kaggle. Dogs vs cats. Ver página 31.

[17] T.M. Mitchell. *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill, 1997. Ver página 32.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. Ver página 32.

[19] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37, 11 2019. Ver página 33.

[20] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences, 2018. Ver página 34.

[21] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. Ver páginas 34, 35, 37, 41, 42, 47, and 58.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009. Ver páginas 34, 35, and 36.

[23] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984. Ver página 35.

[24] University of Cincinnati. Regression trees. Ver páginas 36, 37.

[25] Beth Atkinson Terry Therneau and Brian Ripley. *rpart: Recursive Partitioning and Regression Trees*. Ver página 37.

[26] Kuhn and Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008. Ver páginas 38, 39, and 40.

[27] Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002. Ver página 40.

[28] Iñaki Inza and Borja Calvo. Interpretable machine learning. Ver páginas 42, 47.

[29] Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larrañaga. A survey on multi-output regression. *Data Mining and Knowledge Discovery*, 5:216–233, September 2015. Ver páginas 52, 53, and 55.

[30] Alison J. Burnham, John F. Macgregor, and Román Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48:167–180, 1999. Ver página 52.

[31] Zhongyang Han, Ying Liu, Jun Zhao, and Wei Wang. Real time prediction for converter gas tank levels based on multi-output least square support vector regressor. *Control Engineering Practice*, 20(12):1400–1409, 2012. Ver página 52.

[32] scikit learn. Multiclass and multioutput algorithms. Ver páginas 53, 54.

[33] scikit learn. sklearn.linear_model.linearregression. Ver página 54.

[34] Machine Learning Mastery. Multi-output regression models with python. Ver página 54.

[35] Eleftherios Spyromitros-Xioufis, Grigorios Tsoumakas, William Groves, and Ioannis Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, feb 2016. Ver página 54.