



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Desarrollo de un modelo de aprendizaje automático para la mejora de la detección temprana de casos de TDAH en niños y adolescentes.

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Sánchez Torres, Alejandra

Tutor/a: Vázquez Barrachina, Elena

Cotutor/a: Chirivella González, Vicente

Cotutor/a: Alcover Arandiga, Rosa María

CURSO ACADÉMICO: 2022/2023

Desarrollo de un modelo de aprendizaje automático para la mejora de la detección temprana de casos de TDAH en niños y adolescentes.

# Resumen

---

El Trastorno por Déficit de Atención con Hiperactividad (TDAH) es un trastorno neuropsiquiátrico que se caracteriza por la dificultad para mantener la atención, la hiperactividad y la impulsividad. El TDAH es un trastorno del neurodesarrollo que afecta principalmente a niños y adolescentes, pero también puede persistir en la edad adulta y es por ello por lo que se trata de un tema de gran preocupación para la sociedad. En España, se estima que entre un 2 y un 5% de la población infantil padece TDAH, afectando directamente al rendimiento académico, las relaciones interpersonales y a la calidad de vida de los afectados. Por esta razón, es de vital importancia abordar este trastorno desde diferentes enfoques para mejorar su detección y tratamiento precoz. Actualmente, ya existen algunos estudios que están empezando a aplicar técnicas de aprendizaje automático en este campo, no obstante, muchos de los modelos desarrollados se basan intrínsecamente en modelos de diagnóstico estandarizados y datos sesgados, que hacen que la calidad de predicción del modelo disminuya. Teniendo esto en mente, el presente trabajo final de grado pretende desarrollar un modelo de aprendizaje automático, el cuál será entrenado a través de unos datos recopilados directamente por personas que ya sufren la enfermedad o tienen síntomas de ella. Por lo tanto, el objetivo final es el de conseguir una temprana detección del TDAH tanto en niños como adolescentes, gracias al uso del modelo desarrollado.

Como ya se ha comentado, se utilizarán datos, de niños y jóvenes en edad escolar, recopilados a través de un cuestionario, el cual ha sido distribuido tanto a gran parte de las asociaciones de TDAH en España como de forma personal. Los datos obtenidos a través de este cuestionario serán analizados mediante técnicas estadísticas y de aprendizaje automático para identificar las variables más relevantes para crear y entrenar el modelo.

Gracias a un diagnóstico más rápido, tanto los profesionales de la salud como de la educación podrán, por un lado, brindar el apoyo y la atención necesarias para mejorar la calidad de vida y el rendimiento académico de los afectados, y por otro, reducir significativamente sus respectivas cargas de trabajo. En resumen, este trabajo final de grado tiene como objetivo contribuir al desarrollo de herramientas efectivas y eficientes para la detección temprana del TDAH en niños y adolescentes.

**Palabras clave:** TDAH; hiperactividad; neurodesarrollo; aprendizaje automático; predicción; datos; estadística; mejora.



# Abstract

---

Attention Deficit Hyperactivity Disorder (ADHD) is a neuropsychiatric disorder with symptoms such as difficulty in maintaining attention, hyperactivity, and impulsivity. It is a neurodevelopmental disorder that mainly affects children and adolescents, but can also persist into adulthood, which is why it is a matter of great concern for society. In Spain, it is estimated that between 2 and 5% of the child population suffers from ADHD, directly affecting academic performance, interpersonal relationships, and the quality of life of those affected. For this reason, it is vitally important to address this disorder from different approaches to improve its early detection and treatment. Currently, there are already some studies that are starting to apply machine learning techniques in this field, however, many of the models developed are based on standardised diagnostic models and biased data, which makes the predictive quality of the model decrease. Taking all of this into consideration, this final thesis aims to develop a machine learning model, which will be trained on data collected directly from people who already suffer from the disease or have symptoms of the disease. Therefore, the main goal is to achieve an early detection of ADHD in both children and adolescents, thanks to the use of the developed model.

As already mentioned, data will be used, from school-aged children and young people, collected through a questionnaire, which has been distributed both to most of the ADHD associations in Spain and personally. The data obtained through this questionnaire will be analysed using statistical and machine learning techniques to identify the most relevant variables to create and train the model.

Thanks to a quicker diagnosis, both health and education professionals will be able, on the one hand, to provide the necessary support and care to improve the quality of life and academic performance of those affected, and on the other hand, to significantly reduce their respective workloads. In summary, this final thesis aims to contribute to the development of effective and efficient tools for the early detection of ADHD in children and adolescents.

**Keywords:** ADHD; hyperactivity; neurodevelopment; machine learning; prediction; data; statistics; improvement.

# AGRADECIMIENTOS

---

Quisiera expresar mi profundo agradecimiento a todas las personas que contribuyeron de manera significativa en la realización de este trabajo de fin de grado. Sus valiosas aportaciones y apoyo han sido fundamentales para el éxito de este proyecto.

En primer lugar, quiero agradecer a mis tutores, Elena Vázquez Barrachina, Rosa María Alcover Arandiga y Vicente Chirivella González, por su orientación experta y dedicación. Su guía y disposición para resolver mis dudas han sido de gran ayuda a lo largo del proceso de desarrollo del trabajo.

En segundo lugar, agradecer a todas las personas y asociaciones que decidieron apostar por este proyecto y que me brindaron toda la información necesaria para poder crear y entrenar el modelo.

Además, quiero expresar mi profundo agradecimiento a mi familia, especialmente a mi madre, Flora, por su constante motivación y apoyo incondicional durante todo el tiempo que ha durado este trabajo.

Por último, pero no menos importante, quisiera expresar mi más sincero agradecimiento a Adrián Santiago Sánchez quien ha sido una fuente inagotable de inspiración y motivación en el desarrollo de mi trabajo de fin de grado. Su influencia ha sido fundamental para impulsarme a alcanzar mis metas y superar los desafíos que se me presentaron durante este proyecto.



# Tabla de contenidos

---

1. <b>INTRODUCCIÓN</b> .....	12
1.1 Motivación .....	12
1.2 Objetivos del trabajo .....	13
1.3 Impacto esperado.....	14
1.4 Estructura .....	14
1.5 Planificación del trabajo.....	15
2. <b>EL TDAH</b> .....	17
2.1 Causas y consecuencias.....	17
2.2 Breve historia del TDAH .....	19
2.3 Métodos para el diagnóstico.....	21
2.4 Terapias y tratamientos .....	22
3. <b>ESTADO DEL ARTE</b> .....	23
3.1 Análisis del problema y solución .....	24
3.2 Análisis del marco legal y ético .....	25
4. <b>ANÁLISIS DE LOS DATOS</b> .....	26
4.1 El cuestionario.....	26
4.1.1 Preguntas generales .....	27
4.1.2 Hábitos .....	27
4.1.3 Síntomas .....	28
4.2 Distribución del cuestionario .....	29
4.3 Variables resultantes .....	30
4.4 Preparación de los datos.....	32
4.4.1 Limpieza de datos.....	33
4.4.1.1 Datos irrelevantes.....	33
4.4.1.2 Datos faltantes .....	34
4.4.2 Transformación de variables .....	34
4.5 Análisis exploratorio de los datos .....	35
4.5.1 Datos atípicos .....	35
4.5.2 Análisis de las variables numéricas.....	36
4.5.3 Análisis de las variables categóricas .....	41
4.5.4 Análisis de la variable "Diagnostico_TDAH" .....	43
4.5.4.1 Relación entre la variable "Diagnostico_TDAH" y las variables numéricas ..	44

4.5.4.2	Relación entre las variables presentes en la sección de hábitos y la variable "Diagnostico_TDAH" .....	45
4.5.4.2.1	Relación entre las variables presentes en la sección de hiperactividad y la variable "Diagnostico_TDAH" .....	48
	A continuación, se muestra la distribución de las diferentes variables relacionadas con la hiperactividad en su respectivo histograma.....	48
4.5.4.2.2	Relación entre las variables presentes en la sección de atención y la variable "Diagnostico_TDAH" .....	50
4.5.4.3	Relación entre la variable "Diagnostico_TDAH" y las variables categóricas.	51
<b>5.</b>	<b>ANÁLISIS PREDICTIVO</b> .....	<b>55</b>
5.1	Selección y optimización de modelos predictivos .....	55
5.1.1	Máquina de Soporte Vectorial (SVM) .....	55
5.1.2	Árboles de clasificación .....	56
5.1.3	Random Forest .....	57
5.1.4	Redes neuronales (RNA).....	58
5.2	Preparación de los datos para los modelos predictivos .....	59
5.2.1	Codificación de las variables categóricas para los modelos.....	59
5.2.2	Estandarización de los datos .....	60
5.2.3	Desbalanceo de la variable a predecir .....	60
5.3	Optimización de hiperparámetros .....	61
5.4	Evaluación y comparación de los modelos predictivos.....	62
5.4.1	Análisis de los resultados obtenidos.....	65
5.5	Modelo final .....	66
<b>6.</b>	<b>CONCLUSIONES DEL TRABAJO</b> .....	<b>69</b>
6.1	Conclusiones del análisis .....	69
6.2	Trabajos futuros.....	70
6.3	Relación del trabajo desarrollado con los estudios cursados.....	71
<b>7.</b>	<b>BIBLIOGRAFÍA</b> .....	<b>72</b>
<b>8.</b>	<b>ANEXOS</b> .....	<b>75</b>
8.1	Cuestionario .....	75
8.2	Librerías utilizadas en Python.....	102
8.3	Curva ROC de los modelos analizados .....	103
8.4	Relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).....	105



# LISTA DE IMÁGENES


Imagen 1.- Diagrama de Gantt seguido para el desarrollo de este trabajo final de grado. Fuente: (Haz Una Infografía - Acceder, n.d.) .....	16
Imagen 2.- Representación de la conexión entre neuronas a través de neurotransmisores. Fuente: (Neurotransmisores: Qué Son, Tipos y Descripción de Los Más Conocidos   Psyciencia, n.d.).	18
Imagen 3.- Portado del estudio "An Inquiry into the Nature and Origin of Mental Derangement: Comprehending a Concise System of the Physiology and Pathology of the Human Mind and a History of the Passions and Their Effects". Fuente: (Sir Alexander Crichton - Google Libros, n.d.) .....	20
Imagen 4.- Portada del DSM V. Fuente: (DSM-5: Trastorno Específico Del Aprendizaje, n.d.) .....	21
Imagen 5.- Comparación por neuroimagen de un cerebro con TDAH y otro que no. Fuente: (Neuroimagen En El Diagnóstico Del TDAH - Neuropsicólogo Ulises Espino Rodríguez, n.d.) .....	23
Imagen 6.- Ejemplo de una pregunta de la sección general del cuestionario creado para recopilar los datos. Fuente: Propia. ....	27
Imagen 7.- Ejemplo de preguntas de la sección de hábitos del cuestionario creado para recopilar los datos. Fuente: Propia. ....	28
Imagen 8.- Ejemplo de preguntas de la sección de síntomas del cuestionario creado para recopilar los datos. Fuente: Propia. ....	29
Imagen 9.- Diagramas de caja para las variables numéricas del estudio. Fuente: Propia. ....	36
Imagen 10.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la falta de atención. Fuente: Propia. ....	37
Imagen 11.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la hiperactividad. Fuente: Propia.....	38
Imagen 12.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con hábitos. Fuente: Propia.....	40
Imagen 13.- Conjunto de gráficos de barras para las variables categóricas del estudio. Fuente: Propia. ....	41
Imagen 14.- Distribución de la variable "Diagnostico_TDAH". Fuente: Propia. ....	43
Imagen 15.- Gráficas comparativas de las diferentes variables dentro de la sección de hábitos con la variable "Diagnostico_TDAH". Parte 1. Fuente: Propia. ....	46
Imagen 16.- Gráficas comparativas de las diferentes variables dentro de la sección de hábitos con la variable "Diagnostico_TDAH". Parte 2. Fuente: Propia. ....	47
Imagen 17.- Gráficas comparativas de las diferentes variables dentro de la sección de hiperactividad con la variable "Diagnostico_TDAH". Fuente: Propia.....	49
Imagen 18.- Gráficas comparativas de las diferentes variables dentro de la sección de atención con la variable "Diagnostico_TDAH". Fuente: Propia. ....	50
Imagen 19.- Gráficas comparativas de las diferentes variables categóricas con la variable "Diagnostico_TDAH". Fuente: Propia.....	53
Imagen 20.- Gráfica explicativa del funcionamiento de un modelo SVM. Fuente: (Máquinas Vectores de Soporte Clasificación – Teoría -  Aprende IA, n.d.).....	56
Imagen 21.- Ejemplo de un árbol de decisión. Fuente:(¿Qué Es Un Diagrama de Árbol de Decisión?   Lucidchart, n.d.) .....	57
Imagen 22.- Ejemplo gráfico de un modelo Random Forest. Fuente: (Random Forests Definition   DeepAI, n.d.).....	57



Imagen 23.- Explicación gráfica del funcionamiento del modelo de Red Neuronal. Fuente: (Clasificación de Redes Neuronales Artificiales - Diego Calvo, n.d.).....	58
Imagen 24.- Distribución de la variable "Diagnostico_TDAH" al aplicar la técnica SMOTE. Fuente: Propia. ....	61
Imagen 25.- Imagen explicativa de la matriz de confusión. Fuente: (Confusion Matrix   Everything About Data Science, n.d.).....	63
Imagen 26.- Imagen explicativa del funcionamiento de la validación cruzada con 5 folds. Fuente: (Cómo Realizar La Validación Cruzada y Cruce Seccional de Datos - Blog US.NUMERICA.MX, n.d.) .....	64
Imagen 27.- Ejemplo de Curva ROC. Fuente:(Compare Deep Learning Models Using ROC Curves - MATLAB & Simulink, n.d.). ....	65
Imagen 28.- Curva ROC del modelo final. Fuente: Propia. ....	67
Imagen 29.- Distribución de la importancia de las variables para el modelo final con todas las variables. Fuente: Propia. ....	68
Imagen 30.- Diferentes librerías utilizadas en Python. Fuente: Propia. ....	102
Imagen 31.- Curva ROC del modelo de árbol de clasificación. Fuente: Propia. ....	103
Imagen 32.- Curva ROC del modelo Random Forest. Fuente: Propia.....	104
Imagen 33.- Curva ROC del modelo de Redes Neuronales. Fuente: Propia.....	104



## LISTA DE TABLAS

---

Tabla 1.- Tabla explicativa de los colores asociados a las tareas planificadas en el Diagrama de Gantt de la imagen 1. Fuente: Propia. ....	16
Tabla 2.- Tabla explicativa de los diferentes tratamientos y terapias convencionales para el TDAH. Fuentes: (National Institute of Mental Health, 2016; Treatment of ADHD   CDC, n.d.). ....	22
Tabla 3.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 1. Fuente: Propia.....	30
Tabla 4.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 2. Fuente: Propia.....	31
Tabla 5.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 3. Fuente: Propia.....	31
Tabla 6.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 4. Fuente: Propia.....	31
Tabla 7.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 5. Fuente: Propia.....	32
Tabla 8.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 6. Fuente: Propia.....	32
Tabla 9.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la falta de atención con su media correspondiente. Fuente: Propia. ....	37
Tabla 10.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la hiperactividad con su media correspondiente. Fuente: Propia .....	38
Tabla 11.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con los hábitos con su media correspondiente. Fuente: Propia. ....	39
Tabla 12.- Explicación de los resultados obtenidos al analizar la distribución de las variables categóricas de la imagen 13. Fuente: Propia. ....	43
Tabla 13.- Tabla que muestra las variables numéricas significativas en relación con la variable "Diagnostico_TDAH". Fuente: Propia.....	45
Tabla 14.- Tabla resultado de aplicar el estadístico Chi-cuadrado a las variables categóricas. Fuente: Propia. ....	51
Tabla 15.- Tabla resultante de aplicar la técnica OneHotEncoder a la variable categórica "Sexo". Fuente: Propia. ....	59
Tabla 16.- Tabla resultante al aplicar la técnica Z-score a las variables categóricas .....	60
Tabla 17.- Tabla que muestra los resultados obtenidos al optimizar los hiperparámetros del conjunto de datos balanceado para los diferentes modelos predictivos estudiados. Fuente: Propia. ....	62
Tabla 18.- - Tabla que muestra los resultados obtenidos al optimizar los hiperparámetros del conjunto de datos balanceado para los diferentes modelos predictivos estudiados. Fuente: Propia. ....	62
Tabla 19.- Resultados de las métricas al evaluar los modelos con un conjunto de datos balanceados. Fuente: Propia. ....	65
Tabla 20.- Resultados de las métricas al evaluar los modelos con un conjunto de datos no balanceados. Fuente: Propia. ....	66



# 1. INTRODUCCIÓN

---

El presente trabajo final de grado correspondiente al grado en Ciencia de Datos detalla cómo los avances tecnológicos pueden ayudar a mejorar la calidad de vida de las personas. Concretamente, la actual memoria se centra en el desarrollo de un modelo de aprendizaje automático capaz de detectar casos de Trastorno por Déficit de Atención e Hiperactividad (TDAH) en niños y adolescentes.

El TDAH es un trastorno neuropsiquiátrico común que afecta tanto a niños como adultos, pero que tiene una especial incidencia en la población joven debido a que se encuentran en una etapa de constante aprendizaje y asimilación de nueva información. Este hecho hace que los síntomas más comunes del TDAH, como pueden ser la dificultad de atención, la hiperactividad o la impulsividad, interfieran en el desarrollo personal de los afectados, incidiendo tanto en su rendimiento académico como en su vida personal. (Qué Es El TDAH - Feadah, n.d.).

La falta de detección precoz de los casos de TDAH puede conllevar una tardía intervención por parte de los profesionales de la salud provocando que los efectos del TDAH incidan en todos los ámbitos de la vida diaria del afectado. Es por eso por lo que, el desarrollo de un modelo predictivo que permita identificar a los alumnos con mayor probabilidad de padecer TDAH podría ayudar a realizar diagnósticos más rápidos y efectivos, para así poder mejorar la calidad de vida del afectado.

Por todo lo anterior expuesto, este estudio pretende aprovechar el potencial de las tecnologías relacionadas con el *machine learning* para conseguir una detección precoz del TDAH en niños y adolescentes. Primeramente, se recogerán una muestra de datos a través de un cuestionario, los cuales serán preparados para analizar mediante técnicas estadísticas y de aprendizaje automático para identificar las variables más relevantes para el modelo predictivo. Una vez analizados los datos, se procederá a entrenar al modelo, el cuál propondrá una serie de conclusiones con relación al TDAH sobre las muestras analizadas.

Por lo tanto, en el presente capítulo se pretende mostrar al lector cuales han sido los verdaderos motivos detrás del estudio y qué objetivos se pretenden cumplir con él. Además, se procederá a explicar cuál es el impacto esperado y deseado del proyecto. Finalmente, se explicará que estructura se ha seguido junto a la planificación inicial que ha sido necesaria seguir para poder finalizar el trabajo final de grado.

## 1.1 Motivación

La principal motivación del actual trabajo final de grado nace de una experiencia personal. Durante mi infancia siempre he tenido muchas dificultades académicas que no solo afectaron mi rendimiento académico, sino que también impactaron en mi bienestar personal, generando sentimientos de inferioridad en relación con el resto de mis compañeros y limitando mi confianza en mis propias habilidades. Como resultado, mi autoestima se vio afectada negativamente. Gracias a mi constancia y al apoyo incondicional de mi familia, conseguí finalizar mis estudios obligatorios y obtener una destacada calificación en la selectividad. Inicié mi primer año de

carrera en Matemáticas en la Universidad de Murcia, enfrentándome a las dificultades que se me plantearon tanto a la hora de seguir el ritmo de las clases como de comprender el contenido de las diferentes asignaturas y sumándole una falta de motivación personal para continuar, me llevo a tomar la decisión de abandonar mis estudios.

Afortunadamente, me enteré de que en Valencia se estaba impartiendo un nuevo grado, Ciencia de Datos. Me pareció muy interesante el contenido de éste y decidí inscribirme en él. El primer año no noté tanto los efectos del TDAH, ya que coincidió que las clases se impartían de manera online debido a las consecuencias de la pandemia provocada por el COVID-19, por lo que me fue más sencillo seguir las clases. Durante el segundo año de carrera, y reanudando la presencialidad en las clases, comencé a tener muchas dificultades para poder seguir las clases e incluso me llegué a plantear volver a dejar los estudios.

No obstante, antes de tomar dicha decisión, decidí acudir a los profesionales de la salud correspondientes para que analizaran mi caso. Después de una serie de pruebas, me diagnosticaron con un elevado grado de TDAH, justificando así todo el sufrimiento personal y académico que había sentido toda mi vida. A partir de ese momento, y gracias a la medicación recetada, pude volver a seguir las clases sin los mismos problemas anteriores, siendo capaz de finalizar mis estudios con el fin de poder desarrollar mi carrera profesional en un campo como es el del análisis de datos, que tanto me apasiona.

## 1.2 Objetivos del trabajo

Antes de proceder a mostrar el desarrollo del proyecto, es necesario dejar claros cuáles han sido los objetivos por cumplir y que han marcado y justificado las técnicas y procedimientos empleados. Como ya se ha comentado, el objetivo principal del presente TFG es el de la creación de un modelo de aprendizaje automático capaz de detectar posibles casos de TDAH en estudiantes. No obstante, para poder cumplir dicho objetivo, fue necesario marcar una serie de objetivos más específicos, los cuáles se muestran a continuación:

- Realizar una búsqueda de información relacionada con el TDAH y analizar qué técnicas se han utilizado para su diagnóstico y qué tipos de tecnologías similares a las utilizadas en el presente trabajo ya están en funcionamiento.
- Recopilar los datos necesarios para el estudio a través del diseño de un cuestionario con preguntas adicionales a las habituales que permitan predecir el TDAH.
- Analizar los datos obtenidos con el fin de detectar las variables más interesantes para predecir el TDAH utilizando técnicas de análisis estadístico.
- Desarrollar y entrenar el modelo de aprendizaje automático que pueda llegar a predecir en función de unas variables el tanto por ciento de probabilidad que tienes de padecer TDAH.
- Analizar los resultados obtenidos para mostrar unas conclusiones claras y concisas en el ámbito del problema estudiado.



### 1.3 Impacto esperado

El impacto esperado de este trabajo final de grado es beneficiar el bienestar social de tres grupos de personas. En primer lugar, los claros beneficiados de este estudio serán todos aquellos estudiantes que sufren TDAH y no han sido diagnosticados, ya que gracias al modelo desarrollado podrán detectárselo precozmente. Esto facilitará que se tomen las medidas necesarias de una forma más rápida y efectiva por parte de los profesionales de la salud, mejorando sustancialmente la calidad de vida de los estudiantes afectados y su rendimiento académico.

Por otro lado, los profesionales de la salud también se podrían beneficiar, pues el modelo desarrollado permitirá identificar más fácilmente a los estudiantes que necesiten una evaluación más detallada y un tratamiento adecuado. Además, en el ámbito educativo los docentes pueden verse claramente beneficiados con este estudio pues podrán adaptar de forma adecuada las metodologías y técnicas de enseñanza para así poder educar mejor y de forma más efectiva a estudiantes con este trastorno.

### 1.4 Estructura

El proyecto ha sido estructurado en diferentes capítulos, los cuales explican el procedimiento que se ha seguido para obtener el modelo final. Para que el lector tenga una visión global de lo que se va a encontrar en la memoria, se procede a hacer un breve recorrido sobre cómo se ha estructurado el trabajo. El proyecto se divide en los siguientes capítulos:

- **El TDAH:** Gracias a este capítulo, se muestran las bases teóricas para entender qué es el TDAH, qué causas tiene y qué tratamientos puede tener. A pesar de no tener tanto que ver con el grado cursado, es un apartado de vital importancia ya que pone al lector en el contexto necesario para entender la necesidad de buscar nuevas soluciones para el diagnóstico y tratamiento del TDAH.
- **Estado del arte:** Posteriormente al marco teórico, se procede a hacer un breve análisis de cómo se está utilizando la tecnología estudiada, el *machine learning*, en el campo del diagnóstico y tratamiento del TDAH. Además, se presentan apartados que muestran un análisis del problema que se intenta solucionar y la solución propuesta, además de un apartado en el que abordan los aspectos legales y éticos del tratamiento de datos.
- **Análisis de los datos:** Este apartado es el primer apartado de desarrollo práctico del trabajo final de grado. Aquí se podrá entender que procedimientos se han seguido para obtener los datos, su tratamiento y finalmente su preparación para aplicarlos al modelo de aprendizaje automático.
- **Análisis predictivo:** En este segundo apartado práctico, el objetivo es el de mostrar los diferentes pasos que se han seguido para elegir el modelo final y los resultados obtenidos.
- **Conclusiones del trabajo:** Finalmente, este apartado sirve para destacar las principales conclusiones obtenidas en este trabajo, las líneas futuras o futuras aplicaciones y de hacer un ejercicio de retro inspección sobre cómo los conocimientos aprendidos durante el grado de Ciencia de Datos han ayudado al desarrollo del proyecto.

Cabe destacar que después de este último apartado también se podrán encontrar las referencias utilizadas para el estudio y el siguiente contenido en forma de anexo:

- Anexo 1: Cuestionario completo usado para obtener los datos.
- Anexo 2: Librerías utilizadas en Python.
- Anexo 3: Curva ROC de los modelos analizados.
- Anexo 4: Relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

## **1.5 Planificación del trabajo**

Antes de poder empezar el desarrollo del trabajo, fue necesario realizar una planificación de todas las tareas necesarias para su realización. Para ello, se decidió utilizar un Diagrama de Gantt, el cual se basa en un diagrama de barras en que cada tarea se representa con una barra horizontal. Estas barras están marcadas en función del tiempo esperado de realización de la tarea, para así poder tener una visión clara del progreso del proyecto. Se decidió utilizar esta herramienta ya que es una herramienta visual muy útil para planificar y gestionar proyectos. En la imagen 1 se puede ver el Diagrama de Gantt para el presente trabajo final de grado y en la Tabla 1, una breve explicación de lo que significa cada color.



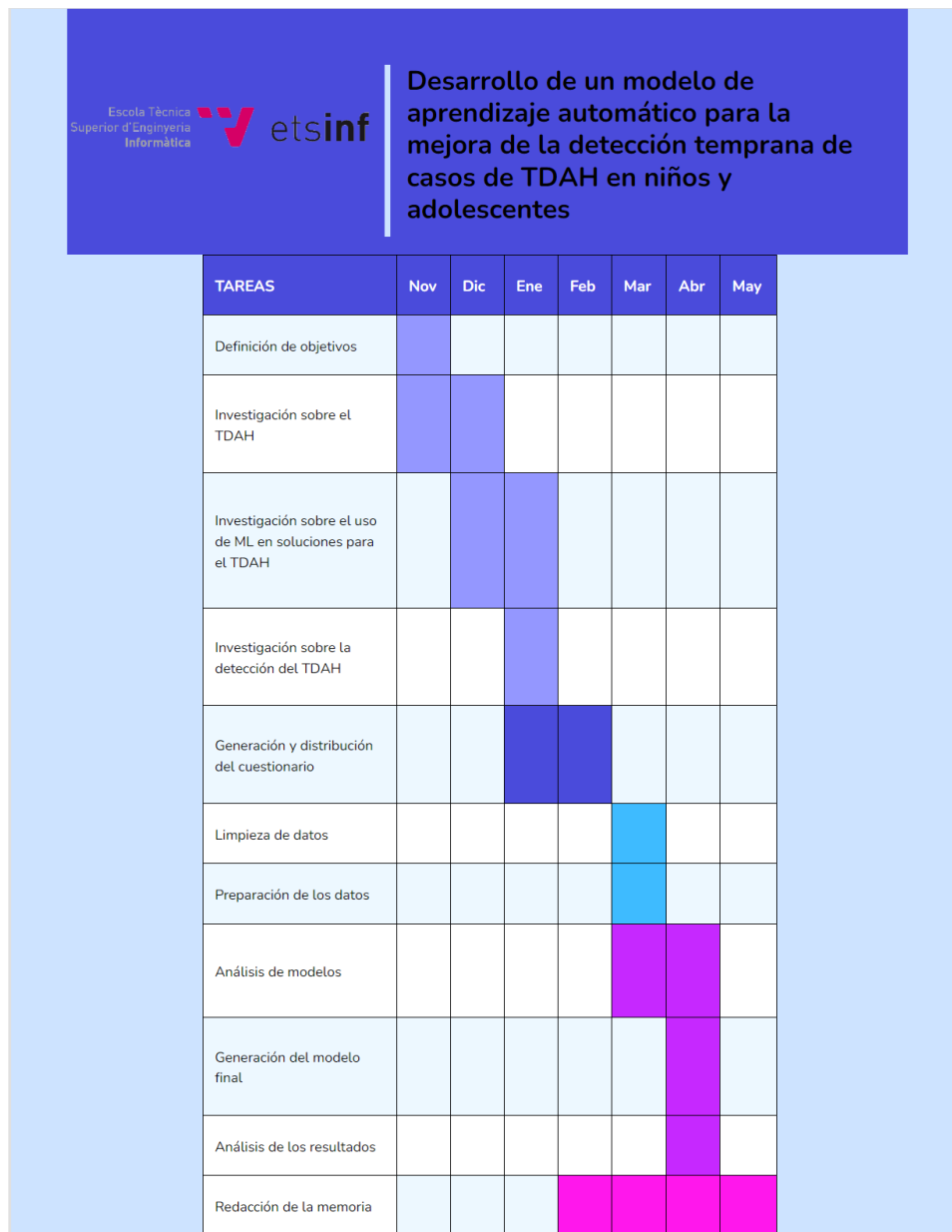


Imagen 1.- Diagrama de Gantt seguido para el desarrollo de este trabajo final de grado. Fuente: (Haz Una Infografía - Acceder, n.d.)

	Las tareas con este color fueron las realizadas en la fase inicial de investigación, en la que se establecieron los objetivos a cumplir y se empezó a estudiar más sobre el tema del proyecto.
	La tarea con este color es la relacionada con la creación, distribución y obtención del cuestionario para generar una base de datos.
	Las tareas con este color son las relacionadas con la preparación de los datos después de haber generado la base de datos para su posterior aplicación al modelo predictivo.
	Las tareas con este color son las que estuvieron relacionadas con el análisis de los diferentes modelos posibles, la aplicación de los datos a ellos y el análisis final en el que se concluye que modelo es mejor para el estudio actual.
	La tarea con este color es la relacionada a la redacción de la memoria, la cual ha sido hecha en paralela a las otras tareas, a medida que se iban finalizando.

Tabla 1.- Tabla explicativa de los colores asociados a las tareas planificadas en el Diagrama de Gantt de la imagen 1. Fuente: Propia.



## 2. EL TDAH

---

El trastorno por Déficit de Atención e Hiperactividad (TDAH) es un trastorno de tipo neurobiológico que se debe al neurodesarrollo de las personas, es decir, este trastorno se genera en el desarrollo del cerebro y el sistema nervioso. Este hecho hace que no sea posible adquirir el trastorno en el transcurso de la vida del afectado, sino que su origen se encuentra en la infancia. No obstante, a pesar de que generalmente se presenta sobre todo en niños y adolescentes, la manifestación de esta enfermedad también puede darse en la etapa adulta de los pacientes. (Biederman & Faraone, 2005). El TDAH es uno de los trastornos más comunes en nuestra sociedad actual ya que se estima que entre el 5 y el 10 % de la población infantil y joven sufren esta enfermedad. Esto hace que haya una media de uno o dos alumnos con TDAH en cada aula. (DuPaul & Stoner, 2014).

El TDAH se caracteriza por la presencia de síntomas como hiperactividad, impulsividad y falta de atención. Como resultado, los niños con TDAH muestran niveles de actividad superiores a los típicos de su edad, encuentran problemas al intentar controlar sus acciones, sentimientos y pensamientos, y tienen problemas para concentrarse y prestar atención. Es importante recordar que existen varios subtipos de TDAH y que estos síntomas no siempre están presentes al mismo tiempo. Los diferentes tipos de TDAH se dividen en distintas categorías en función de los síntomas más frecuentes:

- **Hiperactivo-impulsivo:** Destacan los síntomas de hiperactividad e impulsividad.
- **Inatento:** Destacan los síntomas de falta de atención.
- **Combinado:** Destacan síntomas de hiperactividad, impulsividad e inatención.

(Qué Es El TDAH - Feaadah, n.d.)

### 2.1 Causas y consecuencias

Debido a la gran incidencia de este trastorno, el estudio de éste es algo de vital importancia para los profesionales de la salud. Las últimas investigaciones afirman que el origen del trastorno está totalmente relacionado con la baja producción de los neurotransmisores de dopamina y noradrenalina. (Qué Es El TDAH - Feaadah, n.d.). Para entender estos conceptos es necesario explicar cómo funciona la comunicación de las neuronas del cerebro humano. Los neurotransmisores son los encargados de que se realice esta comunicación y para ello es necesario que exista un nivel determinado de dopamina y noradrenalina, por lo que la ausencia de estos niveles genera una irregularidad dentro del cerebro y sistema nervioso del afectado. (Qué Es El TDAH - Feaadah, n.d.).





*Imagen 2.- Representación de la conexión entre neuronas a través de neurotransmisores. Fuente: (Neurotransmisores: Qué Son, Tipos y Descripción de Los Más Conocidos | Psyciencia, n.d.).*

A pesar de ser un trastorno tan estudiado, no está claro que exista una causa común para todos los afectados, sino que más bien existen varios factores tanto biológicos como sociales que influyen en esta falta de neurotransmisores. Por lo que respecta a los factores biológicos, cabe destacar que el momento de mayor importancia para desarrollar este trastorno se encuentra en el período prenatal o cuando el infante nace. Los siguientes factores son los más importantes a la hora de incrementar el riesgo de padecer TDAH:

- **Genética:** Si uno de los padres del infante ya sufre TDAH, hay un 75% de posibilidades de que el infante pueda sufrirlo.
- **Bajo peso al nacer:** La desnutrición en los primeros meses de vida del infante puede llegar a multiplicar por tres el riesgo de sufrir TDAH.
- **Malos hábitos durante el embarazo:** El consumo de alcohol, tabaco o sustancias estupefacientes por parte de la madre durante el embarazo puede elevar por tres el riesgo de que el infante desarrolle TDAH.

(American Psychiatric Association, 2013).

No obstante, los factores biológicos no son los únicos que influyen en el desarrollo del trastorno, sino que los factores sociales también pueden jugar un gran papel en la aparición del TDAH. Algunos de estos factores son:

- **Pobreza:** La falta de poder adquisitivo para tener una nutrición correcta durante el crecimiento, puede ser un detonante para el desarrollo de la enfermedad. Además, muchos barrios marginales no cuentan con la higiene adecuada, hecho que también puede ser un detonante de la aparición del TDAH.
- **Violencia doméstica y abuso infantil:** Los niños que experimentan este tipo de comportamientos durante su crecimiento son más propensos a desarrollar problemas emocionales y de comportamiento, derivando en algún tipo de TDAH.
- **Falta de supervisión de los padres:** Los niños que tienen padres que no establecen límites claros y consistentes, que no supervisan adecuadamente su comportamiento y que

no ofrecen apoyo emocional, pueden tener más dificultades para desarrollar habilidades de autocontrol y regular sus emociones, lo que a su vez puede llevar al desarrollo del TDAH.

(Rutter & Taylor, 2002).

Uno de los motivos por los que el TDAH es un trastorno tan estudiado, es porque las consecuencias de éste inciden directamente en todos los aspectos de la vida cotidiana de los afectados, llegando a dificultar hasta las tareas más sencillas. Las consecuencias pueden derivar tanto en problemas en el desarrollo social y emocional como en el ámbito académico. Esto puede tener un claro impacto durante toda la vida del afectado, derivando en problemas de ansiedad, depresión y baja autoestima. No obstante, no todos los síntomas son iguales para todos los afectados por lo que una temprana detección del trastorno es necesaria para evitar estas posibles consecuencias.(Biederman, 2005).

## **2.2 Breve historia del TDAH**

Antes de seguir hablando sobre el trastorno en sí, se procede a realizar un breve recorrido sobre la evolución histórica del TDAH. A pesar de ser un trastorno definido y aceptado durante el siglo XX, ya existen registros desde la antigua Grecia en la que se describían comportamientos similares a los síntomas relacionados con el TDAH. Personajes históricos como Hipócrates o el mismo Aristóteles hablan en sus obras sobre personas con comportamientos hiperactivos e impulsivos y una falta de control en sus acciones. A pesar de ello, al no haber registros médicos, esto no se puede afirmar con rotundidad, pero sí que puede dejar ver que el TDAH lleva presente en la sociedad mucho más de lo que se podría pensar inicialmente. (Barkley, 2014).

Si se analizan los diferentes estudios que se han llevado a cabo sobre pacientes con síntomas similares a los del TDAH, es necesario remontarse al siglo XVIII, en el que el médico británico Alexander Crichton publicó un ensayo titulado "An Inquiry into the Nature and Origin of Mental Derangement: Comprehending a Concise System of the Physiology and Pathology of the Human Mind and a History of the Passions and Their Effects". En este ensayo, Alexander Crichton estudia a una serie de niños, los cuales el médico los describió como débiles mentalmente y con deficiencias a la hora de prestar atención, sobre todo en las tareas escolares. (Sir Alexander Crichton - Google Libros, n.d.).



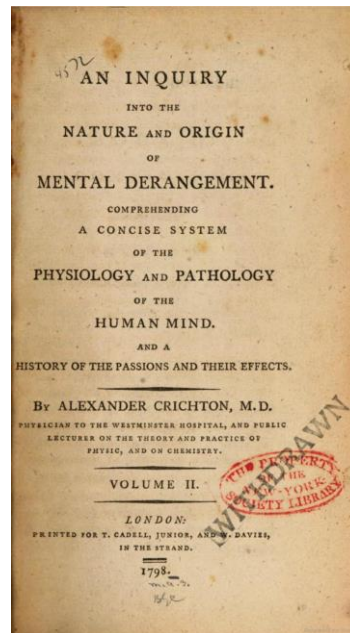


Imagen 3.- Portado del estudio "An Inquiry into the Nature and Origin of Mental Derangement: Comprehending a Concise System of the Physiology and Pathology of the Human Mind and a History of the Passions and Their Effects". Fuente: (Sir Alexander Crichton - Google Libros, n.d.)

Desde entonces, surgieron diferentes estudios sobre niños que coincidían con los mismos patrones de conducta; dificultad para controlar su comportamiento y atención además de una actividad demasiado elevada comparada con el resto de los niños. No obstante, no fue hasta el siglo XX en que surge el término TDAH, ya que el trastorno comenzó a ser considerado y fruto de múltiples estudios.

En la década de los 60 sucede un gran avance para el TDAH ya que es la primera vez que se incluye en el Manual Diagnóstico y Estadístico de los Trastornos Mentales (DSM-II) el término de "síndrome hiperkinético infantil". EL DSM es una obra psiquiátrica que se utiliza para clasificar y diagnosticar los trastornos mentales. Es publicado por la Asociación Estadounidense de Psiquiatría (APA) y actualmente ya va por su quinta edición (DSM-V). No fue hasta la tercera edición del DSM en que se introdujo por primera vez el término "Trastorno por Déficit de Atención" y en la cuarta fue cuando se añadió la subcategoría "tipo hiperactivo-impulsivo" al diagnóstico. En la última edición del documento se han incluido nuevas subcategorías y criterios de diagnósticos con objetivo de mejorar su tratamiento e identificación. (American Psychiatric Association, 2013).

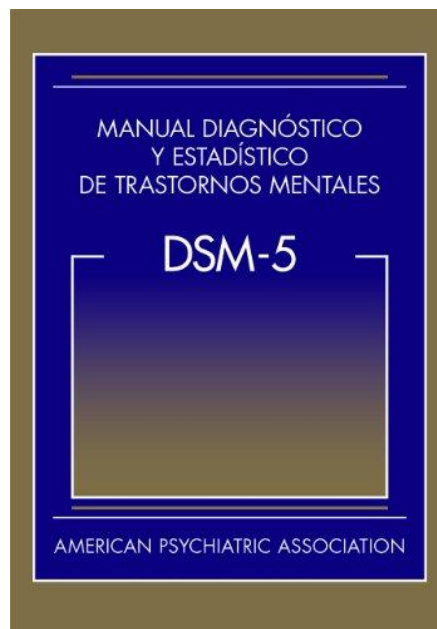
Otro de los momentos más destacables en la historia reciente del TDAH fue cuando la Organización Mundial de la Salud en 1992 aceptó el TDAH dentro de la clasificación internacional de enfermedades (CIE-10), más concretamente, en la categoría de enfermedades mentales. Desde entonces, se han llevado a cabo numerosos estudios sobre el TDAH, lo que ha llevado a una mejor comprensión de su etiología, síntomas, diagnóstico y tratamiento.

## 2.3 Métodos para el diagnóstico

Habiendo revisado el TDAH históricamente, es momento de analizar los diferentes métodos para su diagnóstico.

Como es de esperar, estos métodos han ido evolucionando a la par que se investigaba más sobre el trastorno. Además, la aparición de nuevas tecnologías ha influido en los diferentes métodos para su detección.

Por lo que respecta a la actualidad, el diagnóstico se realiza mediante la combinación de diversas técnicas, entre ellas, la observación clínica, la evaluación psicológica, las entrevistas y la recopilación de información de diferentes fuentes, como los padres, maestros y el propio paciente. Todo este diagnóstico está estandarizado a través del ya comentado DSM-V o del CIE-11, revisión del documento CIE-10 en el que se aceptó el TDAH como una enfermedad mental. (Cortese, 2018).



*Imagen 4.- Portada del DSM V. Fuente: (DSM-5: Trastorno Específico Del Aprendizaje, n.d.).*

El uso de técnicas de neuroimagen para ayudar en el diagnóstico del TDAH es uno de los nuevos métodos que se están investigando y empezando a usar con más frecuencia. Los profesionales de la salud pueden ver en tiempo real la actividad cerebral del paciente, pudiendo así detectar patrones de actividad cerebral anómalos y evaluar la eficacia del tratamiento. La resonancia magnética (RM) y la tomografía por emisión de positrones (PET) son dos de los métodos de neuroimagen más empleados para la detección del TDAH. Estos métodos permiten visualizar la estructura y el funcionamiento del cerebro, lo que ayuda a detectar cualquier anomalía o mal funcionamiento en regiones relacionadas con el TDAH, como el córtex prefrontal, el sistema límbico y el cerebelo. No obstante, como todas nuevas técnicas, tienen desventajas como el coste, su baja disponibilidad en el centro médico o que son técnicas invasivas que pueden requerir del uso de radiación, por lo que es necesario seguir investigando este campo para obtener los mejores resultados posibles. (Rubia, 2018).

En el capítulo del estado del arte (3), se investigarán las diferentes soluciones que actualmente se pueden encontrar utilizando técnicas de aprendizaje automático, con el fin de poder entender el estado actual de la tecnología, aplicada en el diagnóstico y tratamiento del TDAH.

## 2.4 Terapias y tratamientos

Existen diferentes tipos de terapias y tratamientos en relación con el tratamiento del TDAH. Para poder entenderlas, es necesario realizar un análisis de las técnicas actuales en uso. Como ya se ha destacado en apartados anteriores, el TDAH es un trastorno que afecta de forma distinta a cada persona ya que influyen diversos factores en su desarrollo. Es por eso, que cada tratamiento debe ser focalizado a los síntomas de la persona, con el fin de obtener el mejor posible en cada caso concreto.

Siguiendo esta última afirmación, es necesario diferentes enfoques en función del paciente. Es por eso por lo que existen las siguientes opciones para los afectados:

<b>Tratamiento o terapia</b>	<b>Contexto</b>
<b>Tratamiento con medicación</b>	En algunos casos, el TDAH afecta altamente a todos los aspectos de la vida cotidiana del paciente y en estos casos, el profesional de la salud encargado de tratarlo puede decidir medicar al afectado. Estos medicamentos son estimulantes que ayudan a controlar la impulsividad, la hiperactividad y a centrar la mente.
<b>Terapia conductual</b>	Esta terapia se centra en cambiar los patrones de comportamiento del afectado y enseñar nuevas habilidades sociales y de comunicación, siendo a veces complementaria con un tratamiento medicado.
<b>Terapia cognitivo-conductual</b>	Esta terapia se centra en enseñar habilidades de pensamiento y comportamiento que ayudan a manejar los síntomas del TDAH. Se centra en la enseñanza de estrategias de organización, planificación y establecimiento de metas.
<b>Terapia ocupacional</b>	Esta terapia se basa en mejorar habilidades necesarias para la vida diaria, como puede ser la organización y planificación o a prestar atención a conversaciones cotidianas.
<b>Terapia de apoyo familiar</b>	Esta terapia necesita la implicación del núcleo familiar ya que el objetivo es el de mejorar la comunicación y el manejo del estrés en el hogar.

Tabla 2.- Tabla explicativa de los diferentes tratamientos y terapias convencionales para el TDAH. Fuentes: (National Institute of Mental Health, 2016; Treatment of ADHD | CDC, n.d.).

### 3. ESTADO DEL ARTE

---

En este capítulo referente al estado del arte de la tecnología utilizada, *machine learning*, se analizará cómo se utiliza este tipo de tecnología aplicada en el diagnóstico del TDAH. Al ser una tecnología tan novedosa, su uso está en constante evolución por lo que en función de cuando se lea este apartado, la información puede quedar desactualizada.

Es importante aclarar algunas ideas fundamentales antes de seguir adelante con este análisis. El *machine learning* o aprendizaje automático es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a los ordenadores aprender de los datos que se les suministran y mejorar con el tiempo sin tener que ser programados específicamente para cada tarea. Aunque hay muchos tipos distintos de algoritmos de aprendizaje automático, en general, todos tratan de encontrar patrones y relaciones en los datos y utilizar ese conocimiento para sacar conclusiones en función de los datos introducidos. Para ello se construyen modelos matemáticos y estadísticos que pueden aprender a partir de datos anteriores y evaluarse con datos nuevos para determinar su eficacia. Como la base de este sistema es la recopilación y procesamiento de datos, a medida que los modelos se entrenen con más y mejores datos, los sistemas creados serán más precisos y eficaces. (Alpaydin, 2010).

Dejando claro estos conceptos, se procede a analizar sus aplicaciones dentro del campo de estudio del actual trabajo final de grado. Por lo que respecta al diagnóstico y tratamiento del TDAH, se está comenzando a aplicar modelos y sistemas que se basan en el análisis de grandes conjuntos de datos, como registros médicos y de comportamiento, con el fin de poder generar modelos de aprendizaje automático para identificar patrones y predecir el diagnóstico y el resultado del tratamiento. Por ejemplo, algunos estudios han utilizado técnicas de aprendizaje automático para analizar imágenes cerebrales de niños con TDAH y compararlas con imágenes de niños sin el trastorno. Estos estudios han identificado patrones únicos en la actividad cerebral de los niños con TDAH, lo que demuestra que las técnicas de *machine learning* pueden ser una herramienta efectiva para el diagnóstico del TDAH. (Neuroimagen En El Diagnóstico Del TDAH - Neuropsicólogo Ulises Espino Rodríguez, n.d.).

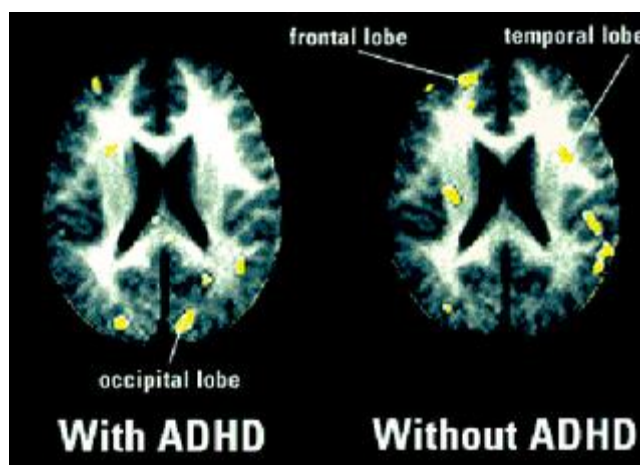


Imagen 5.- Comparación por neuroimagen de un cerebro con TDAH y otro que no. Fuente: (Neuroimagen En El Diagnóstico Del TDAH - Neuropsicólogo Ulises Espino Rodríguez, n.d.).

Además del diagnóstico, se ha propuesto el uso de técnicas de *machine learning* para el tratamiento del TDAH. Algunos estudios han utilizado algoritmos de aprendizaje automático para personalizar los tratamientos en función de las características individuales de los pacientes, como su edad, género, gravedad del trastorno y otros factores clínicos. Algunos ejemplos de estos estudios se pueden encontrar en algunas revistas médicas como la *Journal of Psychiatric Research*, en la que se presentó un estudio en el que los investigadores utilizaron técnicas de *machine learning* para identificar subgrupos de pacientes con trastornos depresivos que podrían beneficiarse más de ciertos tipos de terapias. (Ng, 2017).

A pesar de todas estas ventajas que puede ofrecer el uso del *machine learning*, cabe destacar que como toda tecnología que está en auge, es necesario tener en cuenta tanto sus ventajas como sus desventajas. Por lo que respecta al *machine learning*, es altamente necesario que los datos no estén sesgados o sean incompletos, ya que sino el diagnóstico puede ser impreciso. Es importante tener en cuenta que los algoritmos de *machine learning* son tan buenos como los datos que utilizan y, la calidad del modelo dependerá de la cantidad y calidad de los datos de entrenamiento. La falta de datos confiables y bien documentados puede llevar a resultados imprecisos y por ende a un tratamiento ineficaz o incluso perjudicial. (Brinkman et al., 2019).

### 3.1 Análisis del problema y solución

A continuación, se procede a identificar algunas de las posibles áreas de mejora dentro del campo de la detección y tratamiento del TDAH. Actualmente, la detección del TDAH se basa principalmente en una evaluación clínica en la que se estudian y observan los síntomas, en función de las estandarizaciones de los reglamentos de psiquiatría pertinentes. En muchos casos, cuando el paciente es una persona joven, este análisis u observación es contestado por los responsables del sujeto analizado, por lo que las respuestas pueden estar sesgadas. Por ello, este trabajo propone el generar un cuestionario el cuál debe ser respondido siempre por los afectados para que la información proporcionada al profesional de la salud sea lo más ajustada posible.

Añadido a eso, las evaluaciones clínicas suelen basarse intrínsecamente en las estandarizaciones de los manuales de psiquiatría. En este trabajo se propone utilizar aparte de esos criterios, las preguntas basadas en la experiencia de sufrir la enfermedad, para así darle un enfoque más personal y que el encuestado pueda ofrecer unas respuestas más honestas al sentirse identificado con ellas. (Barkley & Murphy, 2010).

Por lo tanto, este estudio propone desarrollar un modelo de aprendizaje automático, el cuál será creado y entrenado a través de unos datos obtenidos de primera mano por afectados y posibles afectados, haciendo que la información obtenida sea lo más veraz posible y que las predicciones del modelo sean realistas.

En conclusión, se propone el desarrollo de un modelo de aprendizaje automático para la mejora de la detección temprana de casos de TDAH en niños y adolescentes para hacer que esta detección sea mucho más eficaz y se pueda, por un lado, enfocar el tratamiento de una forma más personal, y por otro, ahorrar recursos médicos ya que todo el proceso de detección y tratamiento puede ser más fluido.



### 3.2 Análisis del marco legal y ético

El análisis del marco legal y ético es una de las partes más importantes en cualquier investigación científica que involucra la recolección de datos de los participantes. Como todo estudio que se basa en el uso y tratamiento de datos de terceros, es necesario tener en cuenta cómo se van a tratar esos datos y asegurar que estos van a estar protegidos en todo momento. En este estudio, en el que los datos han sido recopilados a través de un cuestionario anónimo, es especialmente importante asegurar que se cumplan todas las regulaciones legales y éticas aplicables.

Uno de los principales aspectos a considerar es el consentimiento informado de los participantes, o en este caso al tratarse de menores, de sus tutores legales. Este cuestionario fue enviado tanto a asociaciones relacionadas con el TDAH como de forma personal, por lo que en todo momento los participantes estuvieron completamente informados sobre los objetivos de la investigación, la naturaleza de las preguntas y cómo se utilizarán los datos. De este modo se tuvo el consentimiento de los participantes para colaborar en el estudio.

Además, es importante garantizar que se va a respetar la privacidad de los participantes. Por lo que, para garantizar la confidencialidad de los datos recopilados, por un lado, todas las cuestiones fueron respondidas de forma anónima, y por otro, solo tiene acceso a los datos la creadora del estudio y nadie puede modificarlos ni extraer información de ellos más allá de las conclusiones y resultados del presente trabajo final de grado. Es por ello, que se puede afirmar que este estudio respeta las normativas de protección de datos.

Otro aspecto importante para considerar es el cumplimiento de las regulaciones y normas legales que rigen la investigación científica donde se está realizando el estudio. En el caso de España, es necesario tener en cuenta el Reglamento General de Protección de Datos (RGPD) y la Ley Orgánica de Protección de Datos y Garantía de Derechos Digitales (LOPDGDD). En estos reglamentos se estipula que a no ser que las personas participantes en el estudio den su consentimiento, está totalmente prohibido que sea posible identificar de forma directa o indirecta a los participantes y dicha información debe manejarse con confidencialidad y almacenarla en un lugar seguro. (Ley Orgánica de Protección de Datos - LOPDGDD 3/2018 | Grupo Atico34, n.d.; RGPD - Resumen General Del Reglamento | Sage España, n.d.).

En conclusión, para garantizar que el estudio cumpla con el marco legal y ético, ha sido esencial obtener el consentimiento informado de los participantes o de sus tutores legales para respetar su privacidad y la confidencialidad de los datos recopilados, de modo que se cumpliese con todas las regulaciones y normas aplicables. De este modo se asegura que la investigación sea ética y legalmente responsable, lo que a su vez mejorará la credibilidad y la calidad de los resultados obtenidos.



## 4. ANÁLISIS DE LOS DATOS

---

En este trabajo de fin de grado de Ciencia de Datos, uno de los puntos más clave es el tratamiento de los datos. En el caso actual, estos han sido obtenidos gracias a la generación de un cuestionario en línea a través de la aplicación *Google Forms* para posteriormente, ser tratados a través de diversas técnicas estadísticas utilizando el lenguaje de programación *Python* en la aplicación *Jupyter Notebook*. La principal característica de este cuestionario es que se ha creado teniendo en cuenta tanto los criterios específicos del DSM-V, cómo una serie de criterios añadidos a raíz de haber sufrido la enfermedad y que no estaban contemplados.

### 4.1 El cuestionario

Para poder generar un modelo de predicción autónoma, es necesario contar con suficientes datos que sirvan tanto como base de entrenamiento como de validación. Una de las primeras aproximaciones para resolver este problema, fue el de intentar encontrar una base de datos ya creada que proporcionase la información necesaria para el estudio. No obstante, la mayoría de los cuestionarios utilizados para la detección del TDAH son respondidos por padres o maestros, proporcionando información sobre el comportamiento y las habilidades del niño en el hogar y en la escuela. Si bien esta información es útil y necesaria, hay limitaciones en la precisión y objetividad de los datos recopilados. La percepción de la conducta y la capacidad del niño por parte de padres, tutores legales o profesores puede no reflejar la realidad del afectado, produciendo resultados inconsistentes o sesgados. Además, los niños con TDAH pueden presentar síntomas en contextos distintos de la familia y la escuela, lo que puede pasar desapercibido si los datos se recogen exclusivamente de estas fuentes. Con esto en mente, se quiso diseñar un cuestionario con diversas secciones que deben responder los propios niños o jóvenes, en lugar de sus padres, tutores legales o profesores. (Sánchez-Muñoz et al., 2020).

Respecto al contenido del cuestionario, se diseñó con la intención de recopilar información completa y detallada acerca del TDAH y su impacto en la vida de las personas que lo padecen. Para ello, se siguió como base los diferentes criterios de diagnóstico establecido en el DSM-V. La utilización de preguntas del DSM-V es fundamental para este tipo de estudios ya que es el criterio de diagnóstico utilizado por los profesionales de la salud mental. Estas preguntas permiten evaluar la presencia y la gravedad de los síntomas del TDAH, lo que es básico para el diagnóstico y tratamiento del trastorno.

Por otra parte, se añadieron una serie de preguntas, relacionadas con hábitos y síntomas relacionados con el TDAH, basadas en la experiencia personal de sufrir la enfermedad permitiendo obtener una visión más completa y detallada de la experiencia de vivir con el trastorno. Además, las preguntas sobre los hábitos, intereses y actitudes personales pueden mostrar detalles importantes sobre cómo los síntomas del TDAH afectan a la vida diaria. Asimismo, estas preguntas personales podrían aumentar el compromiso y la motivación de los encuestados ya que tiene un toque más personal, consiguiendo un ambiente más relajado y

cómodo para el participante y, por ende, unas respuestas más honestas y precisas. (Barkley & Murphy, 2010).

Por lo tanto, el objetivo final del cuestionario es el de generar una base de datos con las respuestas obtenidas para poder entrenar al modelo predictivo. Este cuestionario, se dividió en tres grandes secciones, para facilitar la recopilación de datos y abordar diferentes aspectos del TDAH. Tal y como se ha comentado en el apartado 1.4, el cuestionario se puede encontrar en el anexo 1 de esta memoria. Comentar también qué, gracias al uso de *Google Forms*, se pudo distribuir de forma sencilla ya que solo era necesario tener una cuenta de correo electrónico de *Google* y acceso a internet.

A continuación, se procede a explicar el contenido de las diferentes secciones que se encuentran dentro del cuestionario.

#### 4.1.1 Preguntas generales

Para esta primera sección, el objetivo era el de realizar una serie de preguntas generales tales como el sexo, lugar de procedencia o estudios actuales, para generar un primer perfil del usuario. Dentro de estas, la más importante y a tener en cuenta dentro de esta sección es la que se muestra en la imagen 6.

Soy una persona que: \*

- Ha sido diagnostica con Déficit de Atención
- No ha sido diagnostica con Déficit de Atención, pero creo que puedo presentar algunos síntomas
- No ha sido diagnostica con Déficit de Atención y considero que no presento ningún síntoma

Imagen 6.- Ejemplo de una pregunta de la sección general del cuestionario creado para recopilar los datos. Fuente: Propia.

La importancia de esta pregunta se debe a que, en el momento de analizar los datos, las respuestas de todas aquellas personas que ya sufren TDAH van a tener una relevancia mayor, ya que, gracias a sus respuestas en el cuestionario, se facilita la obtención de un modelo lo más preciso posible.

#### 4.1.2 Hábitos

Dentro de la sección de hábitos, se centralizan la mayoría de las preguntas creadas a partir de la experiencia personal. Estas preguntas abordan temas sobre el tiempo libre, deporte, actividad diaria, empatía, independencia, facilidad para aprender idiomas, constancia y comunicación con

los profesores. El encuestado puede escoger cuán de acuerdo esta con la afirmación que se propone, tal y como se puede observar en la imagen 7.

The image shows a portion of a questionnaire with four questions, each followed by five radio button options ranging from 'Totalmente en desacuerdo' to 'Totalmente de acuerdo'. The second and third options are selected in each question.

**Soy una persona que en mi tiempo libre disfruto de series, películas o libros: \***

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

---

**Soy una persona que disfruta haciendo deporte, me sirve para desconectar y no me supone mucho sacrificio: \***

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

---

**Me considero una persona activa en mi día a día: \***

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

---

**Me considero una persona empática. Me pongo en la piel de los demás para sentir de verdad lo que la otra persona está experimentando, sobre todo, si está pasando un mal momento: \***

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo

Imagen 7.- Ejemplo de preguntas de la sección de hábitos del cuestionario creado para recopilar los datos. Fuente: Propia.

El objetivo de esta sección es el de saber qué impacto puede llegar a tener el TDAH en diferentes áreas de la vida y, la percepción sobre el diagnóstico recibido por parte del individuo.

### 4.1.3 Síntomas

Dentro de la sección de síntomas, el grueso principal se basa en los criterios de diagnóstico del DSM-V. Los síntomas en los que se basa el cuestionario son la falta de atención y la hiperactividad, y es por ello, por lo que la mayoría de las preguntas se basan en problemas de atención, seguimiento de instrucciones, organización, distracción, olvido e hiperactividad, entre

otras. Además, se preguntó acerca de la duración de estos síntomas, así como de su impacto en la vida cotidiana.

**Síntomas** 0 de 0 puntos

En las siguientes afirmaciones se propondrán varias situaciones cotidianas donde tienes que elegir cómo de acuerdo está con ellas. La escala es la siguiente:

- 1: Totalmente en desacuerdo.
- 2: En desacuerdo.
- 3: Ni de acuerdo ni en desacuerdo.
- 4: De acuerdo.
- 5: Totalmente de acuerdo

---

**No presto atención a detalles o cometo errores por descuido: \***

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

---

**No mantengo la atención durante largos períodos de tiempo: \***

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

---

**No escucho cuando me hablan: \***

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

---

**No sigo instrucciones, no finalizo tareas: \***

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>

---

**Encuentro dificultades para organizarme: \***

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>

Imagen 8.- Ejemplo de preguntas de la sección de síntomas del cuestionario creado para recopilar los datos.

Fuente: Propia.

La gran diferencia de esta sección respecto a las secciones previas es que, en ésta, las respuestas son numéricas. El baremo utilizado se basa en una escala del 1 al 5, donde 1 representa "totalmente en desacuerdo" y 5 representa "totalmente de acuerdo".

## 4.2 Distribución del cuestionario

El cuestionario se ha distribuido de varias formas para maximizar su alcance y aumentar la participación de los encuestados. Por un lado, se ha distribuido de forma personal a través de

contactos directos, como familiares, amigos y conocidos que han sido diagnosticados con TDAH, así como a través de grupos de apoyo y terapia.

Por otro lado, se ha distribuido a través de diferentes asociaciones de toda España relacionadas con el TDAH, como asociaciones de pacientes, asociaciones de familiares y asociaciones de profesionales de la salud mental. Las asociaciones que han decidido participar en el estudio son:

- **FEADAHA:** Federación Española de Asociaciones de Ayuda al Déficit de Atención e Hiperactividad.
- **APADAHCAS:** Asociación de Padres Afectados por Déficit de Atención e Hiperactividad de la provincia de Castellón.
- **ADAHBI:** Asociación Déficit de Atención del Bierzo.
- Asociación TDAH más 16 Valencia.
- **TDAH Santa Coloma de Gramanet.**
- **ANHDA:** Asociación de Elda y comarca del medio Vinalopó de niños/as con TDAH y dificultades del aprendizaje.
- **AHIDA:** Asociación de Hiperactividad e Déficit de Atención Vasca.

Gracias a haber distribuido el cuestionario de forma personal y a diversas asociaciones, ha sido posible que éste llegue a un amplio abanico de afectados por el TDAH, desde pacientes recién diagnosticados hasta aquellos con experiencia en el manejo de la enfermedad. Además, como el cuestionario también ha sido distribuido a personas no diagnosticadas con TDAH, se ha conseguido una muestra más amplia, permitiendo comparar y analizar las respuestas de personas con y sin TDAH. En definitiva, gracias a este cuestionario se ha podido obtener una base de datos con una gran diversidad en cuanto a género, edad y contextos personales.

### 4.3 Variables resultantes

Una vez que se cerró el período de recopilación de los datos, se pudieron generar las variables correspondientes a cada pregunta que se realiza en el cuestionario. Estas variables son necesarias para poder entrenar al modelo. En las siguientes tablas se muestran todas las variables resultantes.

Pregunta del cuestionario	Variable
Indica tu sexo:	Sexo
Indica en qué Comunidad Autónoma resides:	Comunidad
Indica tu nivel de educación actual:	Educación
Indica a qué centro educativo asististe o has asistido en tu etapa escolar (Primaria):	Centro_Educativo
Soy una persona que:	Diagnostico_TDAH
Indica cómo consideras que es tu relación familiar:	Relación_familiar
Indica si sabes de algún familiar directo que presente déficit de atención:	Familiar_TDAH
Indica si eres hijo/o único:	Hijo_unico
Indica el número de hermanos/as que tienes:	Hermanos

Tabla 3.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 1. Fuente: Propia

Pregunta del cuestionario	Variable
Soy una persona que en mi tiempo libre disfruto de series, películas o libros:	Series_Pelis_Libros
Soy una persona que disfruta haciendo deporte, me sirve para desconectar y no me supone mucho sacrificio:	Deporte
Me considero una persona activa en mi día a día:	Activa
Me considero una persona empática. Me pongo en la piel de los demás para sentir de verdad lo que la otra persona está experimentando, sobre todo, si está pasando un mal momento:	Empatica
Me considero una persona independiente, no dependo de otros para hacer algo:	Independiente
Me considero una persona con facilidad para aprender idiomas:	Idiomas
Me considero una persona constante, soy responsable y me esfuerzo por lograr mis metas:	Constante
Me considero una persona que en el caso de tener alguna duda pregunto o preguntaba a los profesores sin ningún tipo de problema:	Pregunta_Profesores
El tipo de asignatura en la que más suelo o solía destacar es:	Asignatura

Tabla 4.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 2. Fuente: Propia

Pregunta del cuestionario	Variable
Le dedico diariamente a las tareas escolares:	Tiempo_Tareas
No presto atención a detalles o cometo errores por descuido:	Atencion_Detalles_A
No mantengo la atención durante largos períodos de tiempo:	Atencion_Tiempo_A
No escucho cuando me hablan:	Escuchar_A
No sigo instrucciones, no finalizo tareas:	Instrucciones_A
Encuentro dificultades para organizarme:	Organización_A
Evito tareas de esfuerzo mental:	Tareas_Esfuerzo_Mental_A
Pierdo objetos con mucha facilidad:	Objetos_A
Me distraigo con estímulos externos:	Distraccion_A

Tabla 5.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 3. Fuente: Propia

Pregunta del cuestionario	Variable
Interrumpo a los demás en sus actividades:	Interrumpir_H
Tengo dificultades para encontrar soluciones a los problemas de la vida diaria:	Soluciones
Tengo dificultades para aprender una nueva tarea, como por ejemplo, aprender cómo llegar a un nuevo lugar:	Nueva_tarea
Tengo dificultades para comenzar y mantener una conversación:	Conversación
Tengo dificultades para relacionarme con personas que no conozco:	Relacionarse
Tengo dificultades para mantener una amistad:	Amistad
Tengo dificultades para llevarme bien con personas cercanas a mi:	Llevarse_Bien

Tabla 6.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 4. Fuente: Propia

Pregunta del cuestionario	Variable
Tengo dificultades para hacer nuevos amigos:	Amigos
Tengo dificultades para llevar a cabo mi trabajo diario o las actividades escolares diarias:	Trabajo_Diario
Tengo facilidad para posponer las tareas diarias:	Tareas_Diarias
Tengo dificultades para participar en actividades en grupo:	Actividades_Grupo
Tengo dificultades para participar en actividades individuales:	Actividades_Ind
Respecto las afirmaciones anteriores, ¿has marcado algún 4 o un 5?	Afirmaciones
¿Estos síntomas han estado presentes antes de los 12 años?	Sintomas

Tabla 7.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 5. Fuente: Propia.

Pregunta del cuestionario	Variable
Soy olvidadizo/a en las actividades diarias:	Olvidar_A
Muevo manos y pies en situaciones en las que debería estar tranquilo/a:	Mover_H
Me levanto constantemente en situaciones en las que debería estar sentado/a:	Levantarse_H
Corro y salto en situaciones inadecuadas:	Correr_H
Tengo dificultad para jugar tranquilamente:	Jugar_tranquilamente_H
Tengo la necesidad de estar siempre en movimiento:	Siempre_movimiento_H
Suelo hablar en exceso:	Hablar_Exceso_H
Me precipito al responder preguntas:	Precipito_Preguntas_H
Tengo dificultad para respetar los turnos:	Turnos_H

Tabla 8.- Tabla comparativa que muestra a la derecha la variable asociada a cada pregunta del cuestionario. Parte 6. Fuente: Propia.

#### 4.4 Preparación de los datos

La preparación de los datos es una etapa fundamental en cualquier proyecto de investigación que involucre la obtención de información o bien el análisis de datos, por lo que, para este estudio, el procesamiento de los datos obtenido a través del cuestionario es crucial para posteriormente entrenar al modelo. Para lograrlo, es necesario limpiar los datos, identificando aquellos irrelevantes para el estudio, encontrar valores faltantes y transformar los datos si fuera necesario con el fin de utilizarlos adecuadamente y conseguir una mayor interpretación y comprensión del problema a partir de ellos.



#### 4.4.1 Limpieza de datos

La primera parte de la preparación de los datos es la limpieza de éstos. La limpieza de datos es una etapa fundamental ya que, a pesar de tener fuentes de recopilación de datos fiables, estos datos han de ser limpiados y preparados para entrenar el modelo. A continuación, se procede a describir las diferentes acciones que se han llevado a cabo para limpiar los datos.

##### 4.4.1.1 Datos irrelevantes

Una de las tareas más importantes en el preprocesamiento de datos es identificar los datos que son significativos y los que no. Por lo tanto, identificar y eliminar los datos que no son relevantes para el estudio es el primer paso por realizar. Los datos irrelevantes son aquellos que no contribuyen a la toma de decisiones o a la comprensión del problema en cuestión, y pueden ser una distracción o incluso obstaculizar la identificación de patrones o tendencias importantes. Analizar los datos irrelevantes implica aplicar técnicas de filtrado y limpieza para eliminar aquella información que no es necesaria para el análisis o que puede generar ruido en los resultados. Esta etapa puede ayudar a mejorar la calidad de los datos y obtener unos resultados más exactos a la realidad. Teniendo esto en cuenta, a través del cuestionario de *Google Forms* se crearon varias variables automáticamente que no aportaban información al estudio, por lo que se decidió eliminarlas del análisis de datos. Las variables que cumplían estas características fueron "Marca temporal" y "Puntuación".

La siguiente variable que no se ha tenido en cuenta en el estudio fue la creada a través de la pregunta: "Respecto las afirmaciones anteriores, ¿has marcado algún 4 o un 5?". Esta pregunta se refiere a si en alguna de las preguntas sobre síntomas se había marcado la casilla 4 o 5, afirmando que los encuestados han sufrido alguno de los síntomas relacionados con el TDAH. Por lo tanto, al seleccionar las muestras que habían marcado algún 4 o 5 en esas afirmaciones, se está identificando a aquellos individuos que se sienten bastante de acuerdo con los síntomas de falta de atención y de hiperactividad. Esto es realmente importante ya que estos síntomas son clave en la evaluación del TDAH, y al seleccionar solo a aquellos individuos que se sienten bastante de acuerdo con estas afirmaciones, se puede estar seguros de que se está trabajando con una muestra que probablemente tenga un mayor grado de severidad del trastorno. Además, esta selección también reduce el ruido en los datos, ya que se eliminan aquellos individuos que no están de acuerdo con las afirmaciones y por lo tanto no son relevantes para el estudio. Asimismo, es importante destacar que al seleccionar únicamente las muestras que respondieron "Si" en la variable "Afirmaciones", se observa que se sigue manteniendo la misma proporción de individuos diagnosticados con TDAH en el conjunto de datos, es decir, la selección de muestras no afecta a la distribución de la variable objetivo.



#### 4.4.1.2 Datos faltantes

Encontrar los valores que faltan es un paso más en el proceso de limpieza de datos. Los datos que no se han registrado o que no están presentes en el conjunto de datos se denominan valores faltantes. Los datos que faltan pueden afectar al estudio ya que pueden introducir sesgos en el análisis y la interpretación de los datos, lo que a su vez puede reducir la precisión del modelo predictivo.

En este caso, después de revisar el *dataframe*, se observó que solamente la variable "Hermanos" mostraba valores faltantes. La variable "Hermanos" contenía valores faltantes porque era una pregunta que solo se podía acceder si en la pregunta "Indica si eres hija/o único:" se había respondido que sí. Por lo tanto, todos aquellos encuestados hijos únicos no llegaron a responder esta pregunta. Para resolver este problema, se decidió reemplazar todos los valores faltantes en esta variable con el valor 0 y seguidamente se eliminó la variable "Hijo\_unico".

#### 4.4.2 Transformación de variables

Habiendo finalizado la limpieza de datos, el siguiente paso es el de la transformación de las variables para una mejor comprensión de los datos y facilitar el análisis de éstos. Debido a que las variables se identificaban por un código alfanumérico poco descriptivo, se renombraron las variables con nombres más descriptivos los cuáles reflejan el contenido de las preguntas. Por ejemplo, la variable "Afirmaciones" es la reformulación de la variable obtenida en la pregunta "Respecto las afirmaciones anteriores, ¿has marcado algún 4 o un 5?".

En adición de lo anterior, se decidió no solo renombrar las variables para hacerlas más comprensibles, sino también cambiar los valores en algunas de ellas para, por un lado, una mejor interpretación y por otro, el de facilitar su análisis posteriormente. Por ejemplo, se reemplazaron los valores de la variable "Diagnostico\_TDAH" por los siguientes términos:

- No ha sido diagnosticado con Déficit de Atención y considero que no presento ningún síntoma → 0.
- No ha sido diagnosticado con Déficit de Atención, pero creo que puedo presentar algunos síntomas → 0.
- Ha sido diagnosticado con Déficit de Atención → 1.

Finalmente, se realizó una transformación en las variables "Series\_Pelis\_Libros", "Deporte", "Activa", "Empatica", "Independiente", "Idiomas", "Constante" y "Pregunta\_Profesores", presentes en la sección 4.1.2, para que estuvieran medidas del mismo modo que aquellas preguntas presentes en la sección 4.1.3.

Para lograr esto, se reemplazaron los valores del siguiente modo:

- Totalmente de acuerdo → 5.
- De acuerdo → 4.
- Ni de acuerdo ni en desacuerdo → 3.
- En desacuerdo → 2.

- Totalmente en desacuerdo → 1.

De esta manera, se logró que todas las variables de la sección hábitos estuvieran medidas del mismo modo que aquellas de la sección de síntomas, lo que aumenta la claridad en el contenido de las variables y facilita su interpretación y análisis.

## 4.5 Análisis exploratorio de los datos

El siguiente punto del análisis de datos es el análisis exploratorio de los datos (EDA). En esta sección se procede a explicar los pasos seguidos para conseguir este análisis que ha tenido como objetivo final el comprender mejor la distribución de las variables y determinar cuál de éstas está más relacionada con la presencia del TDAH. Este paso es totalmente necesario para que posteriormente se pueda generar el modelo de aprendizaje automático.

### 4.5.1 Datos atípicos

Durante el análisis de los datos, se encontró una serie de datos atípicos a tener en cuenta. Los datos atípicos son valores que difieren significativamente del resto de los datos en una variable. Estos resultados pueden deberse a una introducción de datos inexacta, a problemas de medición o simplemente a cambios aleatorios en los datos. Dado que los valores atípicos pueden alterar la distribución de los datos y afectar a la precisión del modelo, es necesario tratarlos y tenerlos en cuenta para garantizar que los resultados del análisis sean fiables.

Para el estudio actual, se empezó utilizando la técnica del rango *intercuartil*, la cual se basa en buscar la diferencia entre el tercer y el primer cuartil de los datos ya que cualquier valor que se encuentre fuera de este rango puede considerarse un valor atípico.



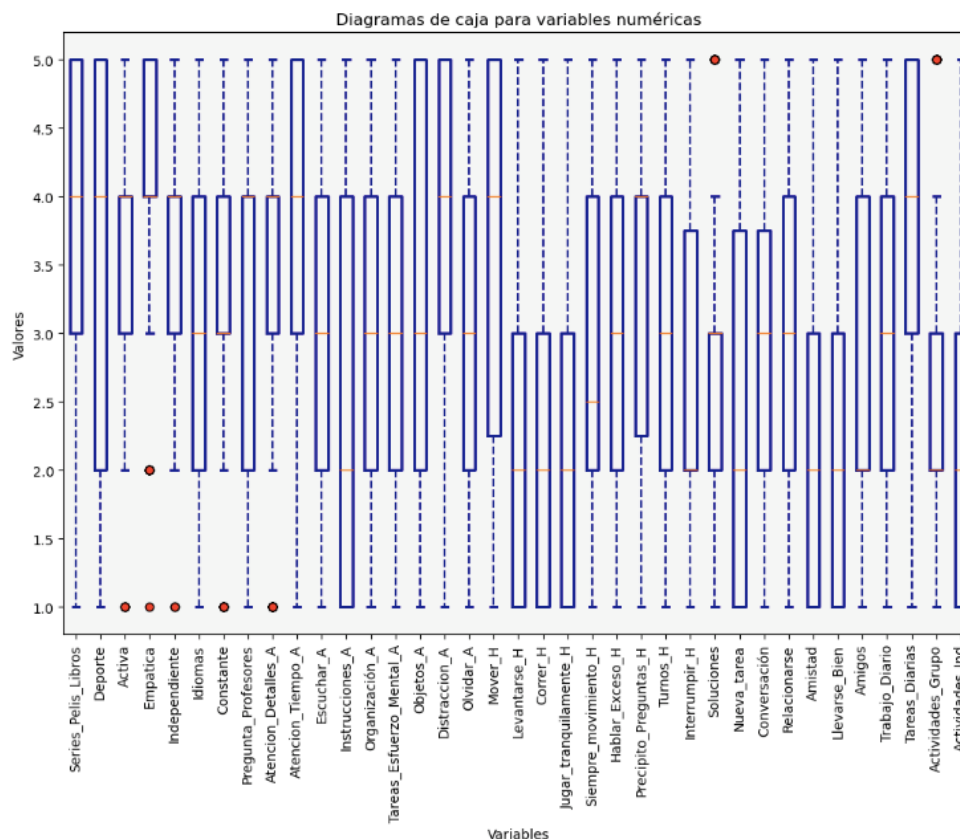


Imagen 9.- Diagramas de caja para las variables numéricas del estudio. Fuente: Propia.

En la imagen 9, se puede apreciar los diferentes valores atípicos que se encontraron después de utilizar la técnica del rango *intercuartil*. Una vez identificados estos datos atípicos, lo normal sería comprobarlos o eliminarlos. No obstante, las variables "Activa", "Empatica", "Independiente", "Constante", "Atención\_Detalles\_A", "Soluciones" y "Actividades\_Grupo" son variables ordinales, por lo que en el caso de estas variables la eliminación o imputación de valores atípicos puede no ser significativa. Esto se debe a que las variables ordinales son variables categóricas, es decir, toman valores discretos y no continuos. En este tipo de variables los valores atípicos no son tan relevantes como lo pueden ser en variables numéricas continuas. La eliminación o imputación de valores atípicos en variables categóricas puede distorsionar la distribución de la variable y, por lo tanto, afectar la validez de cualquier análisis posterior. Por lo que la decisión que se tomó fue la de mantener los valores atípicos encontrados en las variables ordinarias y tenerlos en cuenta para el análisis posterior.

#### 4.5.2 Análisis de las variables numéricas

Para realizar un análisis más detallado y organizado de las variables numéricas presentes en el conjunto de datos, se dividen las variables en las diferentes secciones en las que se presentan en el cuestionario. Por un lado, están las variables que hacen referencia a los síntomas de falta de atención, por otro las variables que hacen referencia a los síntomas de hiperactividad y finalmente las variables que hacen referencia a hábitos del día a día que pueden verse afectados por el trastorno. Los valores para estas variables se encuentran entre el 1 y el 5 y para cada una de estas

secciones, se calcularon las medias de las variables numéricas y se generaron gráficos de barras para visualizar la distribución de los valores.

Síntomas de falta de atención	Media
Instrucciones_A	2.459
Escuchar_A	2.726
Tareas_Esfuerzo_Mental_A	2.938
Olvidar_A	3.151
Organización_A	3.178
Objetos_A	3.308
Atencion_Tiempo_A	3.493
Atencion_Detalles_A	3.521
Distraccion_A	3.808

Tabla 9.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la falta de atención con su media correspondiente. Fuente: Propia.

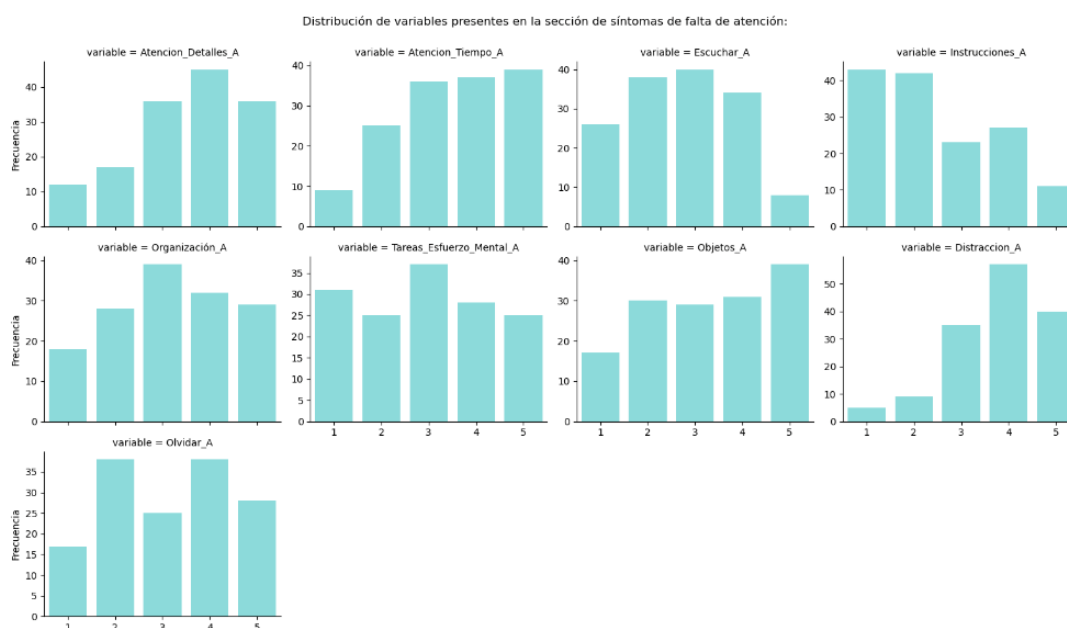


Imagen 10.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la falta de atención. Fuente: Propia.

En la Tabla 9 se puede observar las medias para las variables de síntomas relacionados con la falta de atención. Estas medias oscilan entre 2.46 y 3.81, indicando que los participantes presentaron un grado moderado de dificultades en estas áreas. En concreto, los participantes informaron mayores dificultades en la variable "Distracción\_A", con una media de 3.81 y donde se refleja menos dificultades es en la variable "Instrucciones\_A", en la que la media es de 2.5.

Desarrollo de un modelo de aprendizaje automático para la mejora de la detección temprana de casos de TDAH en niños y adolescentes.

Síntomas de hiperactividad	Media
Escuchar_A	1.938
Instrucciones_A	2.130
Atencion_Tiempo_A	2.363
Olvidar_A	2.582
Organización_A	2.616
Distraccion_A	2.733
Tareas_Esfuerzo_Mental_A	3.103
Objetos_A	3.390
Atencion_Detalles_A	3.479

Tabla 10.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la hiperactividad con su media correspondiente. Fuente: Propia

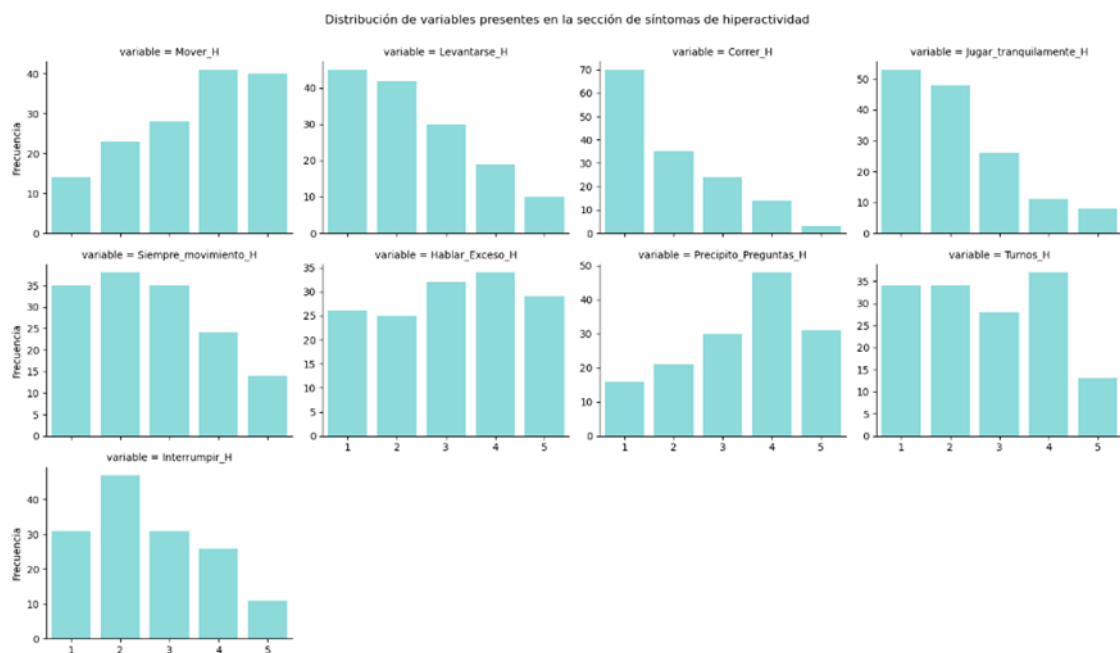


Imagen 11.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con la hiperactividad. Fuente: Propia

En la Tabla 10 se puede observar las medias para las variables de síntomas relacionados con la hiperactividad. Las medias obtenidas en esta sección son en general más bajas que las encontradas en la Tabla 9, ya que los valores oscilan entre la variable "Escuchar\_A" con una media de 1.9 y la variable "Atención\_Detalles\_A" con l mayor media de esta sección, 3.5.

Hábitos	Media
Llevarse_Bien	2.110
Actividades_Ind	2.274
Amistad	2.329
Actividades_Grupo	2.459
Nueva_tarea	2.473
Soluciones	2.562
Conversación	2.644
Amigos	2.678
Trabajo_Diario	2.692
Relacionarse	2.829
Idiomas	2.863
Pregunta_Profesores	3.130
Constante	3.390
Deporte	3.397
Independiente	3.705
Activa	3.726
Series_Pelis_Libros	3.747
Tareas_Diarias	3.767
Empatica	4.253

Tabla 11.- Tabla con las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con los hábitos con su media correspondiente. Fuente: Propia.



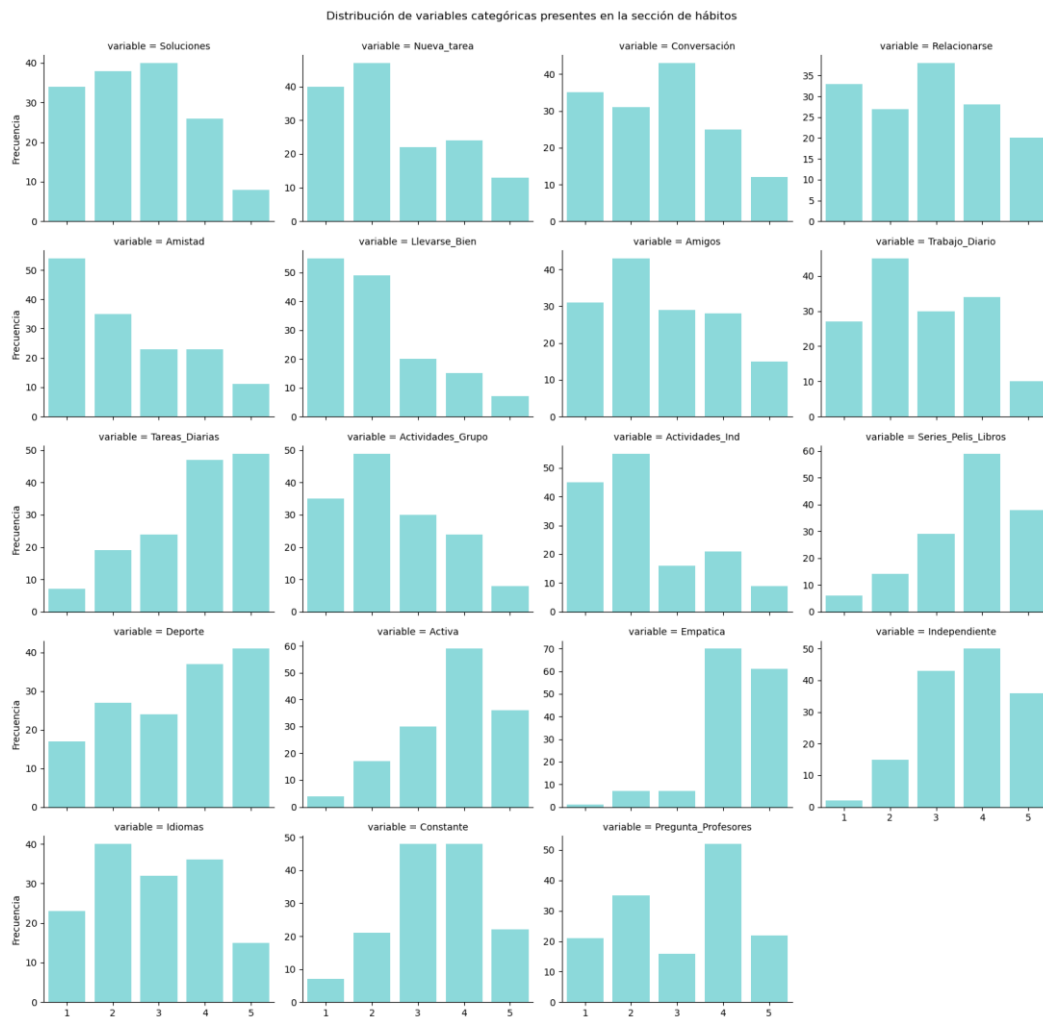


Imagen 12.- Conjunto de gráficos de barras para las variables numéricas obtenidas de la sección de síntomas del cuestionario relacionadas con hábitos. Fuente: Propia

En la Tabla 11 se puede observar las medias para las variables de síntomas relacionados con los hábitos del día a día. En la sección de hábitos, las medias oscilaron entre 2.11 y 4.25, indicando que los participantes presentaron una variabilidad de habilidades en diferentes áreas, siendo la variable "Empatica" la que obtuvo la media más alta y la variable "Llevarse\_Bien" la que menos.

Como primera conclusión de este análisis, se puede afirmar que los participantes del estudio tienen un grado moderado de dificultades en las áreas evaluadas, lo que puede ser relevante para evaluaciones clínicas con el fin de buscar el tratamiento que más se adecue al paciente. Además, al visualizar los gráficos de barras correspondientes a cada una de las variables, se puede observar que éstas presentan una alta variabilidad en sus valores. Específicamente, la mayoría de los participantes presentan valores bajos en algunas de las variables, lo cual indica que no experimentaron con tanta frecuencia los síntomas asociados a dichas variables. Por otro lado, también se observa a algunos participantes con valores muy altos en las mismas variables, lo que sugiere que experimentaron con mayor frecuencia los síntomas relacionados con esas variables. En consecuencia, estos resultados señalan la presencia de una gran variabilidad en los síntomas del TDAH, tal y como se esperaba.



### 4.5.3 Análisis de las variables categóricas

Una vez acabado el análisis de las variables numéricas, es turno de las variables categóricas. Para ello, se crearon gráficos de barras para observar la distribución de cada variable.

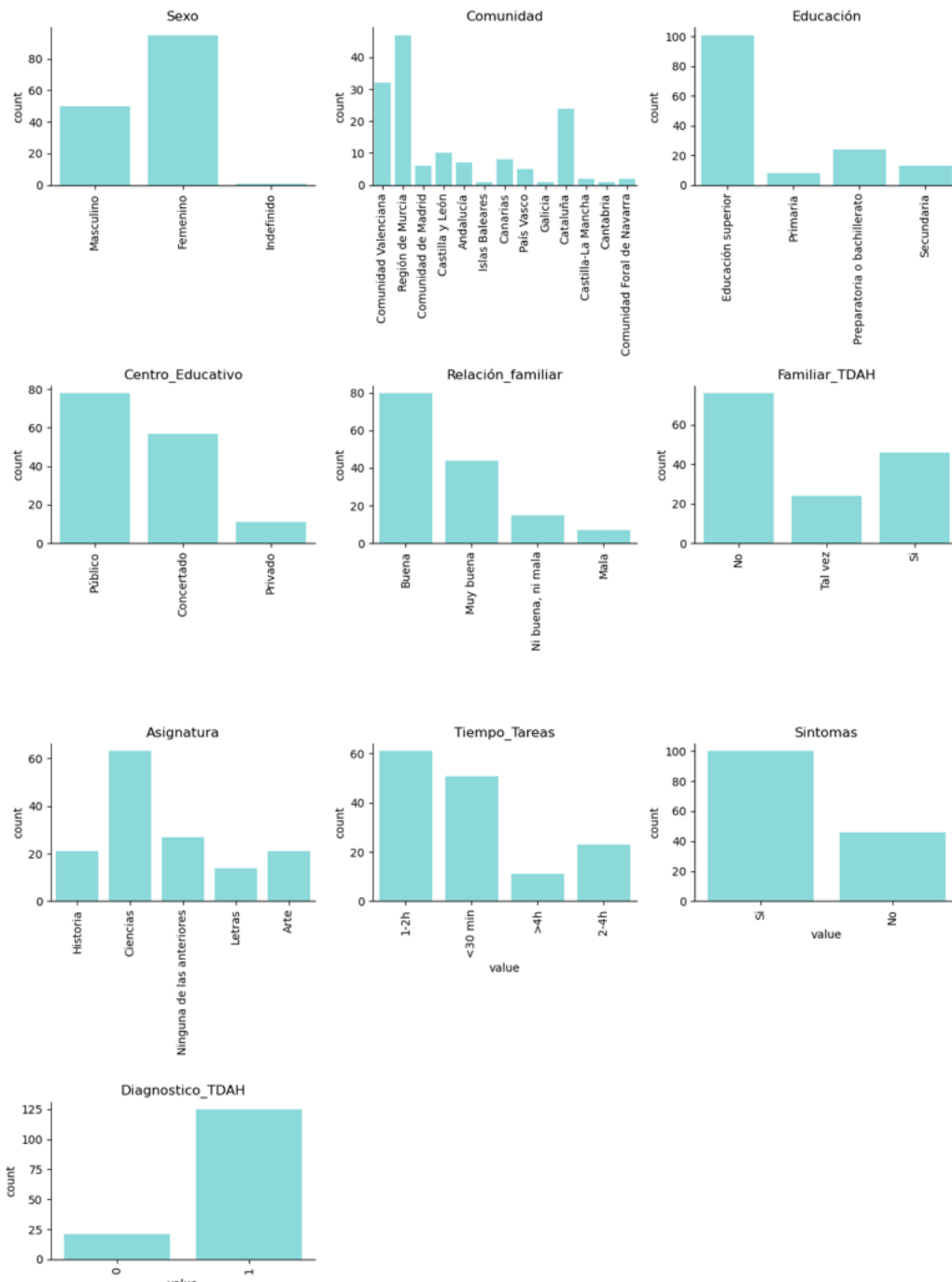


Imagen 13.- Conjunto de gráficos de barras para las variables categóricas del estudio. Fuente: Propia.

Al analizar las gráficas, se observan varios comportamientos los cuales son especificados en la Tabla 12. En general, los resultados obtenidos muestran una gran variabilidad en los datos de las variables categóricas y proporcionan una descripción general de la población de participantes en el estudio.

<b>Variable</b>	<b>Comportamiento</b>
Sexo	De esta variable se puede afirmar que tiene una distribución desigual ya que, de las 146 muestras estudiadas, 95 son femeninas, 50 masculinas y 1 indefinida. Esto era algo esperado debido a la forma en la que se distribuyó el cuestionario, en la que no se tenía control para conseguir una muestra equitativa con relación al sexo de los participantes.
Comunidad	Para esta variable pasa algo similar a la anterior, pero con una distribución más diversa, destacando las comunidades de la Región de Murcia, seguida de la Comunidad Valenciana y de Cataluña.
Educación	La variable "Educación" revela que la mayoría de las observaciones están en la categoría de educación superior, seguida de preparatoria o bachillerato y secundaria, lo que podría sugerir una correlación entre el nivel de educación y los síntomas de TDAH.
Centro_Educativo	Para esta variable se observa una clara predominancia de estudios en centros públicos o concertados por parte de los encuestados.
Relación_familiar	Para esta variable se observa como la relación familiar de los encuestados es predominantemente buena o muy buena.
Familiar_TDAH	Esta variable muestra como la mayoría de los encuestados no tienen ningún familiar directo diagnosticado con TDAH.
Asignatura	En el caso de la variable referente al tipo de asignatura en la que el participante destacaba, las asignaturas relacionadas con la ciencia son aquellas en las que los participantes se sentían más cómodos mientras que las asignaturas relacionadas con letras destacan por ser las que menos.
Tiempo_Tareas	Respecto al tiempo empleado en las tareas escolares, la gran mayoría de los participantes le dedican de 1 a 2 horas o menos de 30 minutos.
Sintomas	La variable "Síntomas" tiene un valor añadido ya que se relaciona con la pregunta de si el encuestado ha sufrido alguno de los síntomas antes de los 12 años. En el gráfico, se puede observar una gran predominancia del Sí, lo que concuerda con lo dicho en el DSM-V. (American Psychiatric Association, 2013).

Diagnostico_TDAH	Esta variable se relaciona con si el encuestado ha sido diagnosticado con TDAH o no. En caso afirmativo se categoriza con un 0 y si es negativo con un 1.
------------------	---

Tabla 12.- Explicación de los resultados obtenidos al analizar la distribución de las variables categóricas de la imagen 13. Fuente: Propia.

#### 4.5.4 Análisis de la variable "Diagnostico\_TDAH"

La variable "Diagnostico\_TDAH" ha sido tratada como una variable numérica binaria que toma el valor 0, si el individuo ha sido diagnosticado previamente con TDAH, o el valor 1 en caso contrario. Una vez analizada la variable "Diagnostico\_TDAH" se observa que, de todos los individuos presentes en la base de datos, 21 individuos (14,38%) habían sido diagnosticados previamente con TDAH, mientras que los 125 individuos restantes (85,62%) no lo habían sido. La primera conclusión de este análisis es que, en la muestra analizada, el TDAH no es una condición común. No obstante, se aproximan a lo visto en el marco teórico, en el que se estimaba que entre un 5 y 10% de la población padece TDAH.

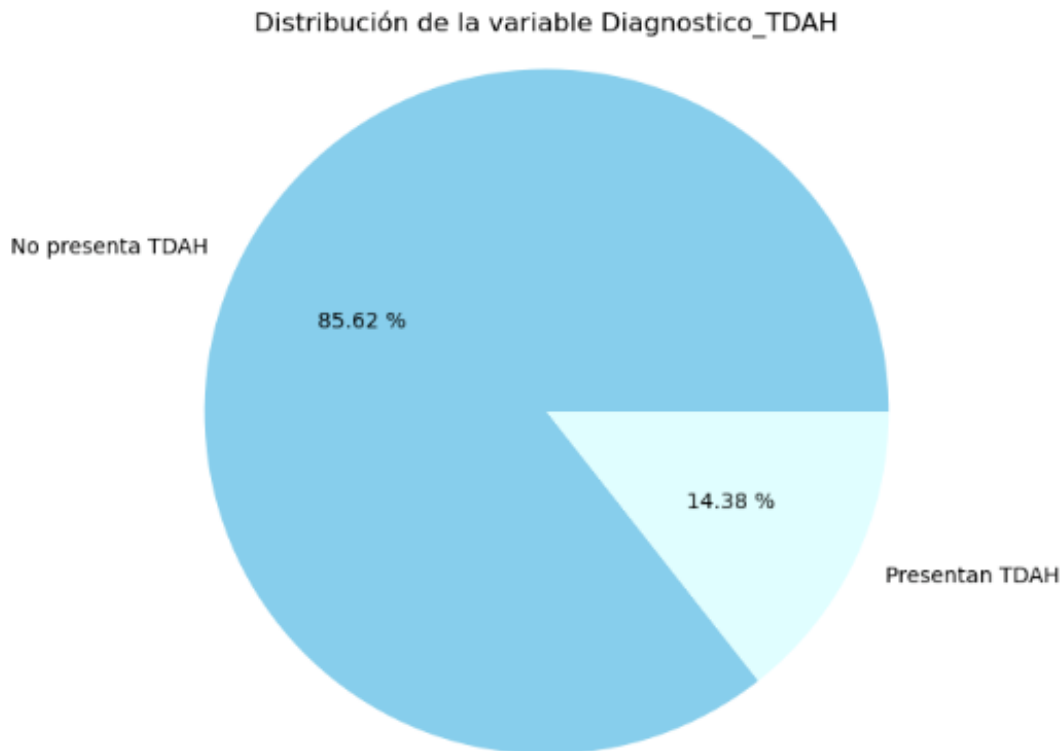


Imagen 14.- Distribución de la variable "Diagnostico\_TDAH". Fuente: Propia.

#### 4.5.4.1 Relación entre la variable "Diagnostico\_TDAH" y las variables numéricas

Para evaluar la relación entre las variables numéricas de la base de datos y la variable "Diagnostico\_TDAH", la cual también es numérica, se utiliza el estadístico de correlación de *Spearman*. Se decidió utilizar esta técnica estadística ya que por un lado no se encontró una distribución normal, comportamiento esperado ya que el TDAH afecta de forma diferente en función de la persona, y por otro, ya que la muestra obtenida no se trataba de un gran conjunto de datos, perfecto para el uso del coeficiente de *Spearman*.

El coeficiente de *Spearman* es una medida de correlación no paramétrica que evalúa la relación monótonica entre dos variables. Este coeficiente puede variar en rango de -1 a 1, donde -1 indica una correlación negativa perfecta, 0 indica ausencia de correlación y 1 indica una correlación positiva perfecta. Además, también se puede obtener el p-valor que indica la significancia estadística de la correlación observada. Un p-valor menor a un nivel de significancia previamente establecido (comúnmente 0.05) sugiere que la correlación observada es estadísticamente significativa, es decir, que es poco probable que se haya obtenido por azar. En cambio, un p-valor mayor a ese nivel de significancia indica que la correlación observada podría deberse al azar y no es estadísticamente significativa.

Por lo tanto, al aplicar el coeficiente de *Spearman* a las variables numéricas presentes en la base de datos, se encontraron varias variables con una correlación significativa con la variable "Diagnostico\_TDAH". Estas variables se pueden ver agrupadas en la Tabla 13.

Variable	Spearman Coefficient	p-value	Significativa
Independiente	0.23	0.01	True
Idiomas	0.34	0.00	True
Atencion_Detalles_A	-0.28	0.00	True
Atencion_Tiempo_A	-0.29	0.00	True
Escuchar_A	-0.30	0.00	True
Instrucciones_A	-0.19	0.02	True
Organización_A	-0.18	0.03	True
Tareas_Esfuerzo_Mental_A	-0.23	0.01	True
Objetos_A	-0.24	0.00	True
Distraccion_A	-0.28	0.00	True
Olvidar_A	-0.36	0.00	True
Mover_H	-0.23	0.01	True
Jugar_tranquilamente_H	-0.29	0.00	True
Siempre_movimiento_H	-0.30	0.00	True
Precipito_Preguntas_H	-0.25	0.00	True
Turnos_H	-0.24	0.00	True
Interrumpir_H	-0.23	0.00	True
Soluciones	-0.20	0.02	True
Nueva_tarea	-0.23	0.00	True
Conversación	-0.18	0.03	True
Trabajo_Diario	-0.18	0.03	True
Actividades_Ind	-0.23	0.00	True

Tabla 13.- Tabla que muestra las variables numéricas significativas en relación con la variable "Diagnostico\_TDAH". Fuente: Propia.

En este caso, las variables numéricas mostradas en la Tabla 13 muestran que existe una correlación estadísticamente significativa con la variable "Diagnostico\_TDAH", ya que tienen p-valores menores a 0.05. Por lo tanto, el resto de las variables numéricas que no aparecen en la tabla no presentan una correlación significativa con la variable "Diagnostico\_TDAH". Cabe decir, que una variable no explicativa en solitario puede ser explicativa en una interacción, al matizar el efecto de otra variable. No obstante, con el objetivo de conseguir un modelo con el menor ruido posible, se decidió no tenerlas en cuenta para el estudio.

A continuación, se han dividido las variables numéricas en tres subconjuntos para compararlas gráficamente con la variable "Diagnostico\_TDAH".

#### 4.5.4.2 Relación entre las variables presentes en la sección de hábitos y la variable "Diagnostico\_TDAH"

Para observar la distribución de los valores numéricos de cada variable para cada nivel de la variable numérica "Diagnostico\_TDAH", se generaron histogramas en los que se pueden observar ciertas diferencias de dichas distribuciones.



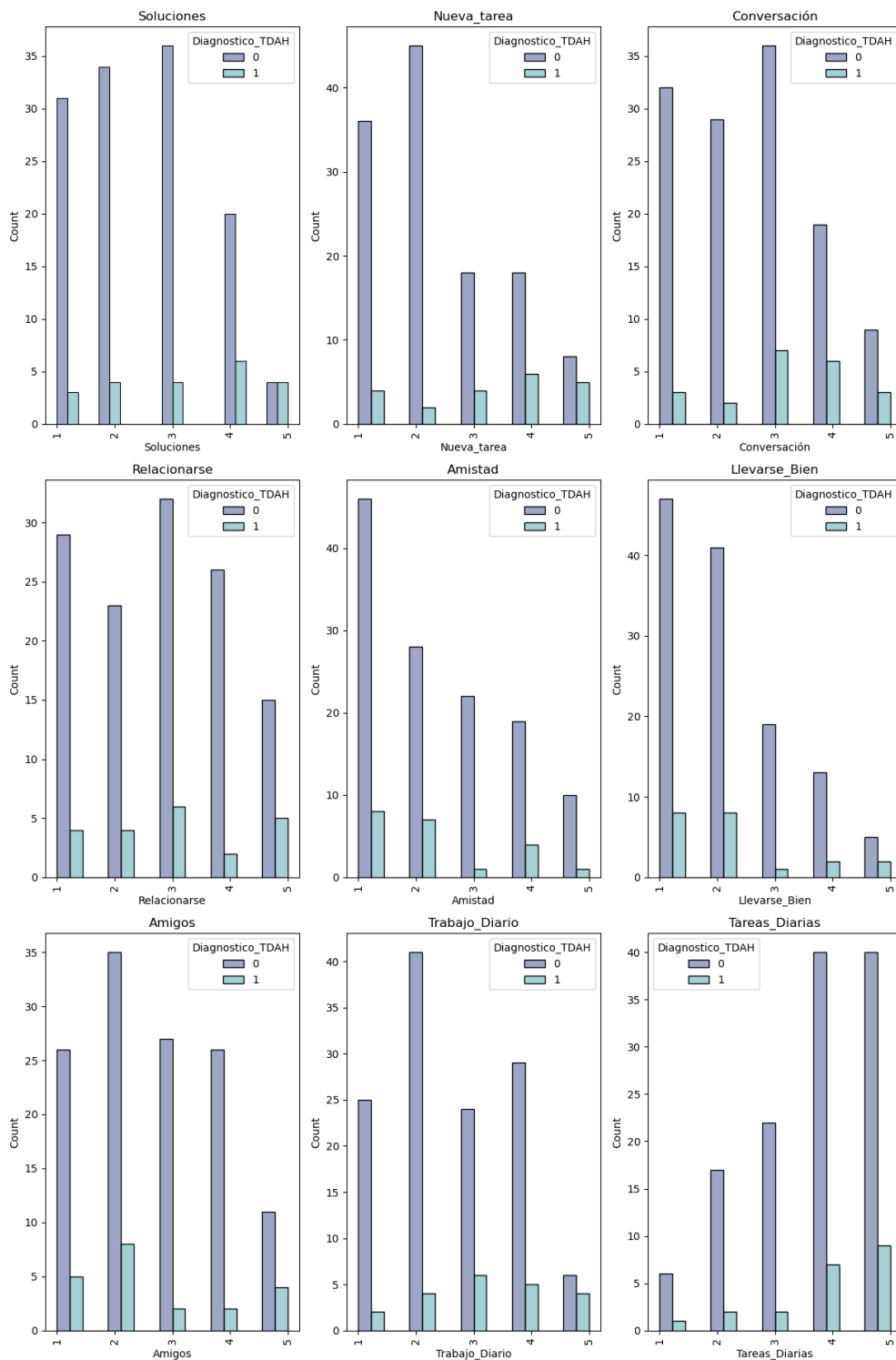


Imagen 15.- Gráficas comparativas de las diferentes variables dentro de la sección de hábitos con la variable "Diagnostico\_TDAH". Parte 1. Fuente: Propia.

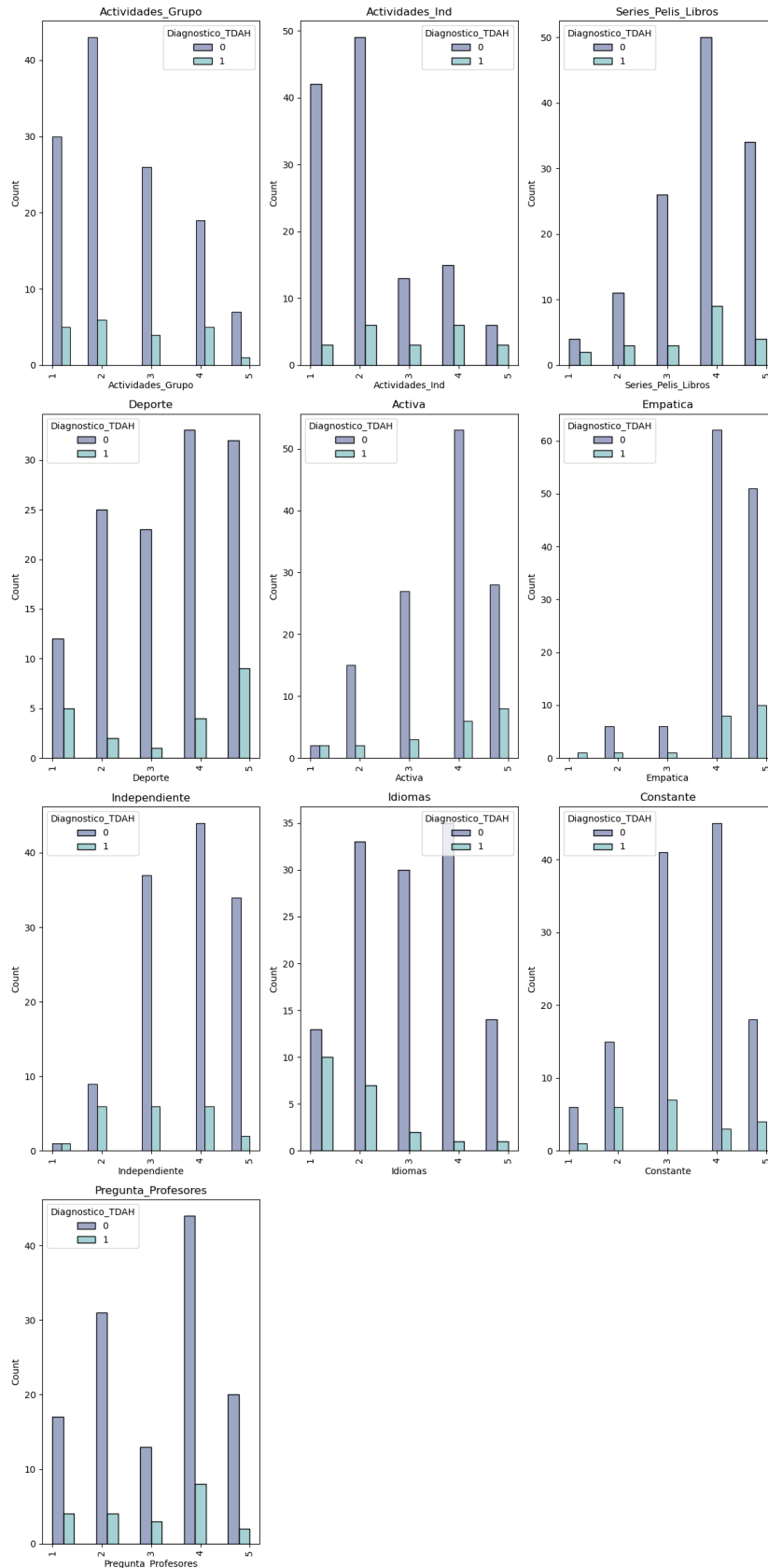


Imagen 16.- Gráficas comparativas de las diferentes variables dentro de la sección de hábitos con la variable "Diagnostico\_TDAH". Parte 2. Fuente: Propia.



A partir de los histogramas se pueden observar que, en general, parece haber una mayor concentración de valores más altos en las variables relacionadas con el rendimiento académico y habilidades sociales en aquellos individuos diagnosticados con TDAH.

Por ejemplo, en la variable "Nueva\_tarea", se puede ver que los individuos diagnosticados con TDAH tienden a tener una mayor concentración de valores altos (niveles 4 y 5) en comparación con los individuos no diagnosticados. Esto puede indicar que la presencia de este trastorno puede dificultar el aprendizaje de nuevas tareas. Lo mismo ocurre en la variable "Tareas\_Diarias", donde se puede observar que los individuos ya diagnosticados presentan una predisposición mayor al resto de participantes en el estudio a procrastinar las tareas diarias. También se puede observar una tendencia similar en las variables "Deporte", "Activa" y "Empática", donde los individuos diagnosticados con TDAH tienden a tener una mayor concentración de valores altos. Por otro lado, en las variables relacionadas con la interacción social, como "Amistad", "Llevarse\_Bien" y "Amigos", parece haber una menor concentración de valores altos en los individuos diagnosticados con TDAH.

Por lo que los resultados sugieren que los individuos diagnosticados con TDAH pueden tener un mejor desempeño en tareas que requieren actividad física, pero pueden tener más dificultades en tareas que requieren atención o en habilidades sociales y/o emocionales. Es importante tener en cuenta que estos resultados se basan en un conjunto de datos limitado y se necesitarían más investigaciones para confirmar estas observaciones.

#### **4.5.4.2.1 Relación entre las variables presentes en la sección de hiperactividad y la variable "Diagnostico\_TDAH"**

A continuación, se muestra la distribución de las diferentes variables relacionadas con la hiperactividad en su respectivo histograma.



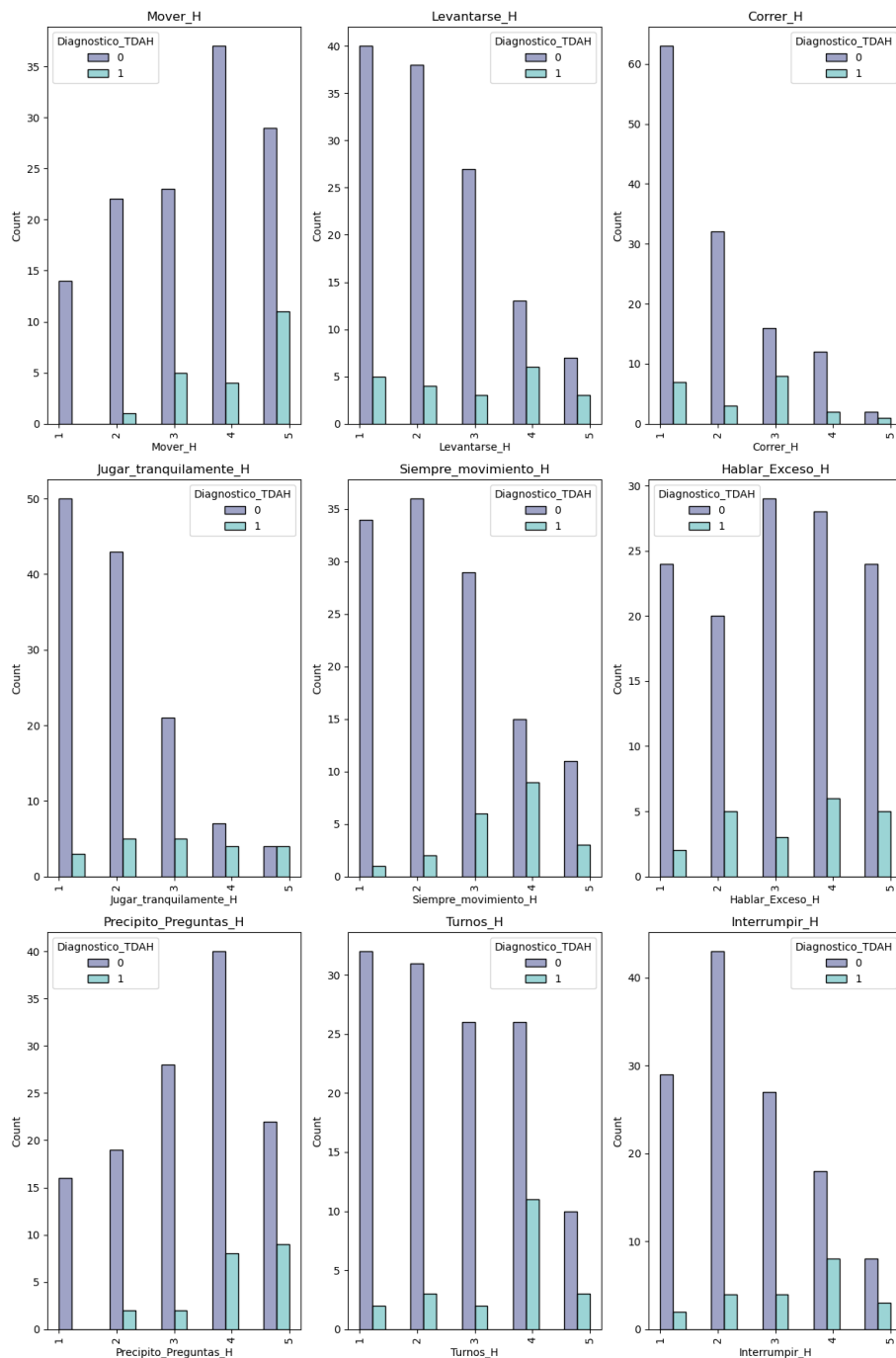


Imagen 17.- Gráficas comparativas de las diferentes variables dentro de la sección de hiperactividad con la variable "Diagnostico\_TDAH". Fuente: Propia.

En la imagen 17 se puede observar que aquellos individuos que fueron diagnosticados con TDAH presentan valores más altos en la mayoría de las variables analizadas que aquellos que individuos que no fueron diagnosticadas con TDAH.

Específicamente, se observa una mayor frecuencia de valores altos en las variables "Mover\_H", "Correr\_H", "Siempre\_movimiento\_H", "Hablar\_Exceso\_H", "Precipito\_Preguntas\_H" e "Interrumpir\_H" para aquellos individuos que tienen TDAH. En la variable "Jugar\_tranquilamente\_H" se puede ver una distribución similar entre los valores de ambas categorías, mientras que en la variable "Turnos\_H" se observa una distribución más homogénea



en ambos grupos, aunque con una mayor concentración de valores altos en aquellos con TDAH. Estos resultados confirman que la hiperactividad está especialmente relacionada con el trastorno.

#### 4.5.4.2.2 Relación entre las variables presentes en la sección de atención y la variable "Diagnostico\_TDAH"

Finalmente, se muestra la distribución de las diferentes variables relacionadas con la falta de atención en su respectivo histograma.

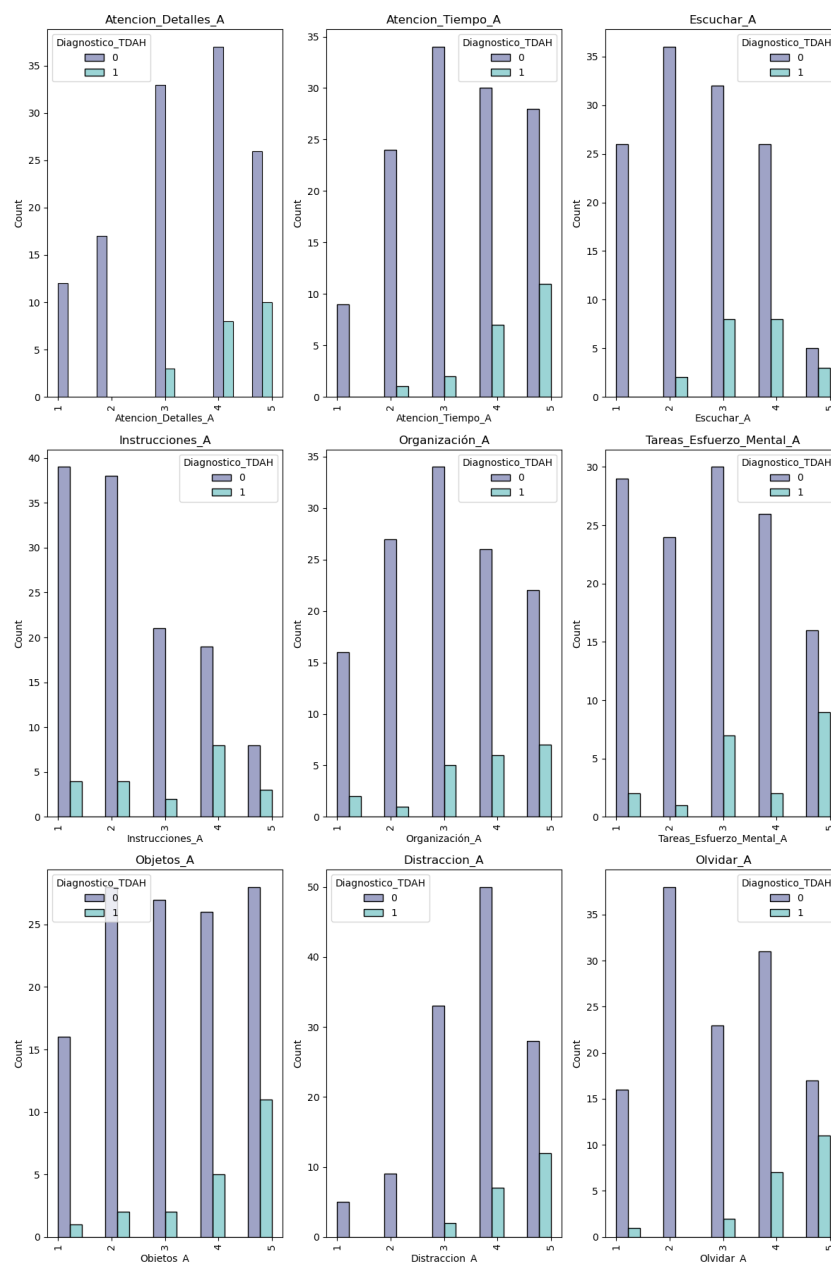


Imagen 18.- Gráficas comparativas de las diferentes variables dentro de la sección de atención con la variable "Diagnostico\_TDAH". Fuente: Propia.

A partir de los resultados presentados, se puede decir que los individuos con diagnóstico de TDAH presentan valores numéricos más altos en la mayoría de las variables, lo que sugiere una mayor dificultad para prestar atención y concentrarse en tareas específicas en comparación con aquellos no diagnosticados.

#### 4.5.4.3 Relación entre la variable "Diagnostico\_TDAH" y las variables categóricas

Para observar la relación entre la variable "Diagnostico\_TDAH" y las diferentes variables categóricas presentes en la base de datos se utilizó una prueba no paramétrica, más concretamente, el test de Chi-cuadrado ( $\chi^2$ ). La elección del test Chi-cuadrado es similar al caso de las variables numéricas, ya que los datos no son normales y se tiene una muestra relativamente pequeña. Además, este test es útil para analizar la relación entre variables categóricas y variables numéricas binarias.

El funcionamiento del test Chi-cuadrado se basa en comparar la frecuencia observada de cada categoría con la frecuencia que se esperaría si no hubiera relación entre las variables. Se calcula un estadístico de prueba ( $\chi^2$ ) que indica la discrepancia entre las frecuencias observadas y esperadas, y se compara con un valor crítico de la distribución chi-cuadrado para determinar si hay una relación significativa entre las variables. En el caso del estudio actual, la construcción de la tabla de contingencia y la aplicación de la prueba de chi-cuadrado permitieron analizar si la distribución de frecuencia observada en la tabla de contingencia difería significativamente de la distribución de frecuencia que se esperaría si las variables fueran independientes. Además, la prueba de chi-cuadrado permitió evaluar la fuerza de la asociación entre las variables categóricas y la variable "Diagnostico\_TDAH". El valor de la estadística de chi-cuadrado y el valor p asociado a ella permitieron cuantificar la significancia de la relación observada.

Variable	Estadístico chi-cuadrado	P-valor	Relación
Sexo	7.090199	0.028866	Significativa
Comunidad	31.069614	0.001922	Significativa
Educación	4.097167	0.251161	No significativa
Centro_Educativo	0.346706	0.840841	No significativa
Relación_familiar	2.421734	0.489602	No significativa
Familiar_TDAH	11.021894	0.004042	Significativa
Asignatura	8.769858	0.067117	No significativa
Tiempo_Tareas	5.187386	0.158579	No significativa
Sintomas	9.642033	0.001902	Significativa

Tabla 14.- Tabla resultado de aplicar el estadístico Chi-cuadrado a las variables categóricas. Fuente: Propia.

El resultado del test se presenta en la Tabla 14 en forma de estadístico chi-cuadrado y p-valor, donde un p-valor menor a 0.05 indica que existe una relación significativa entre las variables. Por lo que como se puede observar en la tabla, las variables "Sexo", "Comunidad", "Familiar\_TDAH" y "Sintomas" presentan una relación significativa con la variable "Diagnostico\_TDAH", mientras que aquellas variables que presentan un p-valor mayor a 0.05, como las variables "Educación",



"Centro\_Educativo", "Relación\_familiar", "Asignatura" y "Tiempo\_Tareas", se consideran como no significativas, por lo que no se tienen en cuenta para el estudio. El motivo por el que se decidió eliminar las variables no significativas viene porque así se puede reducir la complejidad del modelo para así centrarse en aquellas variables que realmente aportan información valiosa para la construcción del modelo. Además, mantener variables no significativas puede resultar en un aumento del ruido o variabilidad en los resultados, disminuyendo la capacidad predictiva del modelo y dificultando la interpretación de los resultados finales.

No obstante, fue necesario un análisis diferente para la variable "Comunidad". A pesar de que ésta presenta una relación significativa con la variable objetivo "Diagnostico\_TDAH", al examinar los datos se observó que las muestras no cubren todas las comunidades en igual medida, por lo que los resultados obtenidos podrían estar sesgados. Además, la inclusión de esta variable no aportaría información valiosa al modelo predictivo, ya que éste necesitaría datos de todas las comunidades para ser útil en la predicción de casos de TDAH. Por lo tanto, se optó por eliminarla del estudio para asegurar la calidad y precisión del modelo.

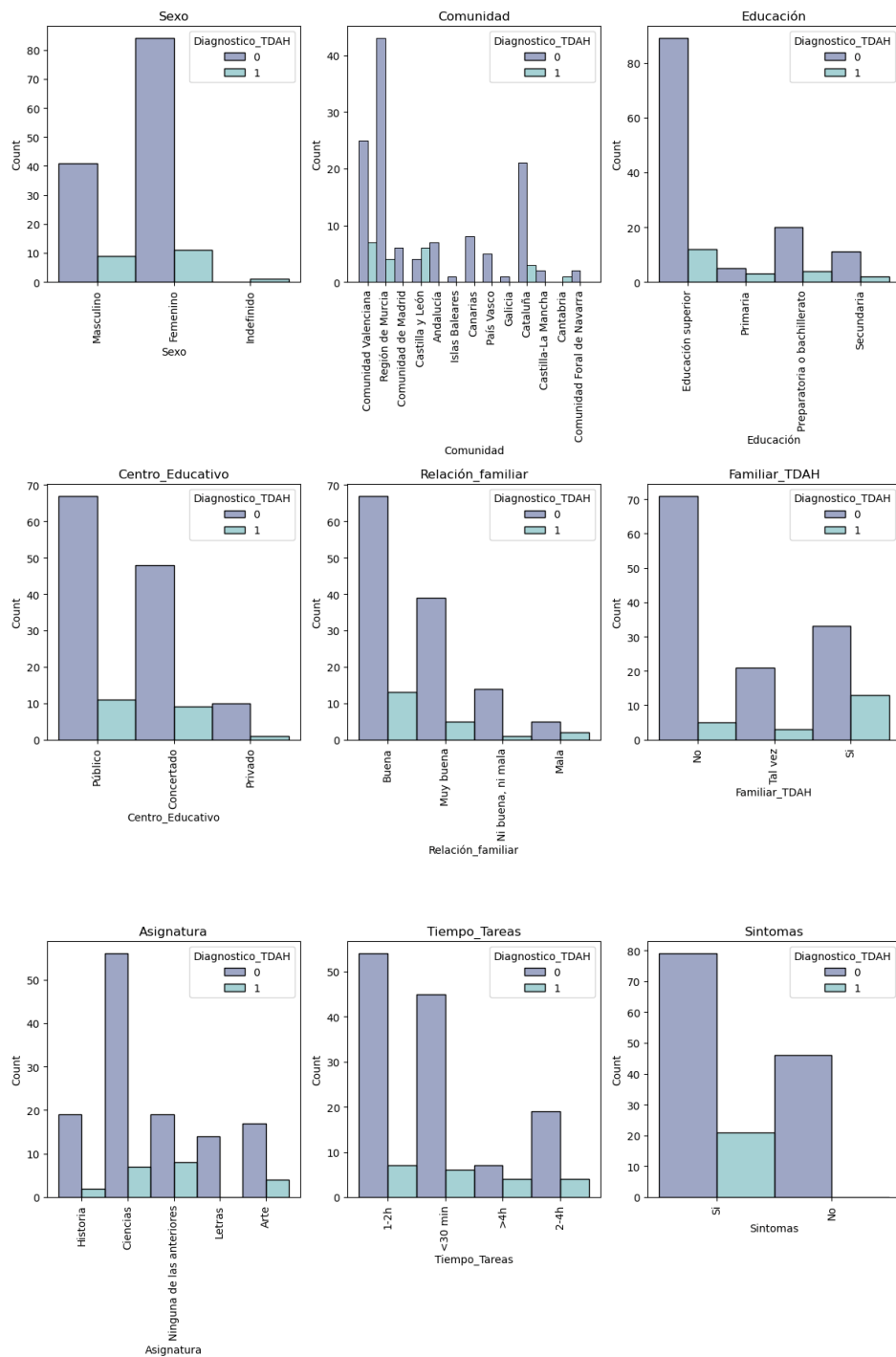


Imagen 19.- Gráficas comparativas de las diferentes variables categóricas con la variable "Diagnostico\_TDAH".  
Fuente: Propia.

A nivel gráfico, se observa que el número de individuos con TDAH es bastante similar entre hombres y mujeres. También se puede ver que hay una mayor proporción de individuos con TDAH en la educación superior. Esto puede deberse a que probablemente haya casos que todavía no hayan sido diagnosticados en niveles de estudio inferiores. En el ámbito educativo, se observa que la mayoría de los individuos que padecen TDAH estudian en centros educativos públicos, mientras que en los centros concertados y privados hay menos casos de TDAH.

En cuanto a la relación familiar, hay una proporción significativa de individuos que tienen una buena relación con sus familiares, especialmente para aquellos



que padecen TDAH. En cuanto a la variable "Familiar\_TDAH", se puede observar que la mayoría de los individuos con TDAH tienen un familiar directo que ha sido diagnosticado previamente con dicho trastorno. En la variable "Asignatura", se puede ver que los pacientes que padecen TDAH tienen más facilidad en las asignaturas de ciencias y arte. Por otra parte, también se observa que la mayoría de los pacientes invierten de 1 a 2 horas en hacer tareas, independientemente de si tienen TDAH o no. Finalmente, en la variable "Síntomas" se puede ver que todos los individuos con TDAH han sufrido los síntomas antes de los 12 años.

## 5. ANÁLISIS PREDICTIVO

---

En esta sección, se procede a explicar todos los pasos que han sido necesarios para la creación de un modelo predictivo con el fin de detectar el TDAH en niños y adolescentes. Esto ha sido posible gracias al uso de la técnica del aprendizaje automático, la cual permite a las máquinas aprender de manera autónoma a partir de los datos.

El proceso para obtener el modelo explicado pasa por la preparación de los datos para ser utilizados en los modelos predictivos para luego seguir con la selección y optimización de los modelos, con su evaluación y comparación pertinente. Finalmente, se presentará una descripción detallada de los modelos entrenados con el fin de encontrar un modelo que tenga una buena capacidad de predicción y que pueda ayudar a entender mejor los factores que influyen en el TDAH. De esta manera, se busca determinar cuál de estos modelos es el más adecuado para una detección temprana del TDAH en niños y adolescentes.

### 5.1 Selección y optimización de modelos predictivos

Antes de seleccionar los modelos, es importante tener en cuenta que no todos los algoritmos funcionan igual de bien con todos los tipos de datos. Por lo tanto, se evaluaron varios algoritmos de aprendizaje automático para encontrar el que mejor se ajusta a los datos estudiados.

En este estudio, se busca crear un modelo predictivo capaz de detectar el TDAH utilizando datos que presentan relaciones no lineales, es decir, las relaciones entre las variables de entrada y la variable objetivo no siguen una forma lineal, lo que hace que el análisis de los datos sea más complejo. Para lograr este objetivo, se han seleccionado varios modelos de aprendizaje automático, entre ellos SVM, Árboles de clasificación, *Random Forest* y Redes neuronales. La razón de elegir estos modelos se debe a que son capaces de aprender patrones y relaciones complejas entre las variables de entrada y la variable objetivo, incluso si los datos no tienen una relación lineal clara.

#### 5.1.1 Máquina de Soporte Vectorial (SVM)

El algoritmo máquinas de soporte vectorial (SVM) consiste en un algoritmo de aprendizaje supervisado que se utiliza para predecir la clase a la que pertenece un conjunto de datos. El objetivo del SVM es encontrar un hiperplano que maximice el margen, es decir, la distancia entre los puntos de datos de varias clases. El algoritmo funciona a través de una serie de parámetros que permiten que los resultados puedan ser más ajustados al conjunto de datos introducido. Uno de ellos es el parámetro *kernel*, el cual se basa en la clasificación de los puntos en una clase u otra, y por tanto le da forma al hiperplano. Además, otro de los parámetros a tener en cuenta es el  $C$ , ya que permite controlar el margen y la cantidad de error permitida en la clasificación. Si el parámetro  $C$  es alto, el margen para que el modelo clasifique correctamente la mayoría de los



datos de entrenamiento será menor. Por otro lado, un valor bajo del parámetro  $C$  dará lugar a un modelo que puede clasificar incorrectamente algunos puntos de entrenamiento, pero producirá un margen mayor. Una de las principales ventajas del SVM es que puede manejar datos con exceso de ruido y llenos de errores sin afectar significativamente a la predicción ya que es resistente a la presencia de valores atípicos. Como consecuencia, el SVM es capaz de manejar conjuntos de datos con un gran número de variables o atributos. (Cherkassky & Ma, 2004).

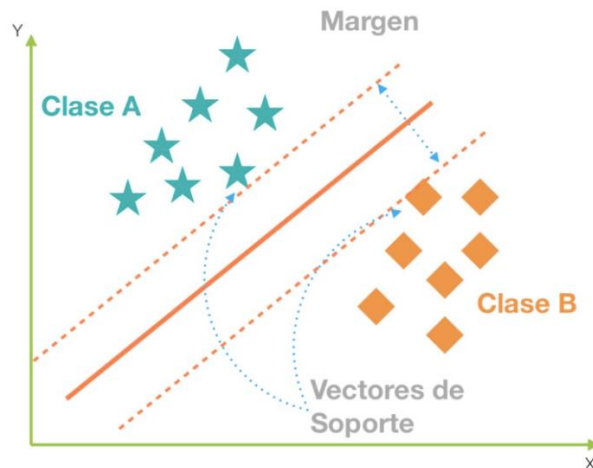


Imagen 20.- Gráfica explicativa del funcionamiento de un modelo SVM. Fuente: (Máquinas Vectores de Soporte Clasificación – Teoría - Aprende IA, n.d.).

### 5.1.2 Árboles de clasificación

Los árboles de clasificación son modelos de aprendizaje supervisado que se utilizan para predecir la clase o categoría a la que pertenece un objeto, a partir de características o atributos de ese objeto. En este modelo, los datos de entrenamiento se utilizan para construir un árbol de decisión que predice la clase de una observación. Para crear el árbol, el conjunto de datos se divide en grupos más pequeños utilizando criterios de división que dependen de las características del conjunto de datos.

El objetivo del modelo es construir un árbol que divida con éxito el conjunto de datos en subconjuntos, y que las observaciones de cada subconjunto compartan propiedades comparables con respecto a la variable objetivo. En cada nodo del árbol se plantea una pregunta relativa a una variable, y el conjunto de datos se divide en función de la respuesta. El árbol final puede representarse gráficamente, en la que cada nodo interno representa una división en la colección de datos y cada hoja una clase. Para determinar la clase de una nueva observación, se recorre el árbol haciendo preguntas en cada nodo interno hasta llegar a una hoja. Una de las ventajas de los Árboles de Clasificación es que son fácilmente interpretables y permiten identificar las variables más importantes en la clasificación. Sin embargo, también pueden ser propensos al sobreajuste si se construyen árboles muy complejos con muchos nodos y divisiones. (Hastie et al., 2009).



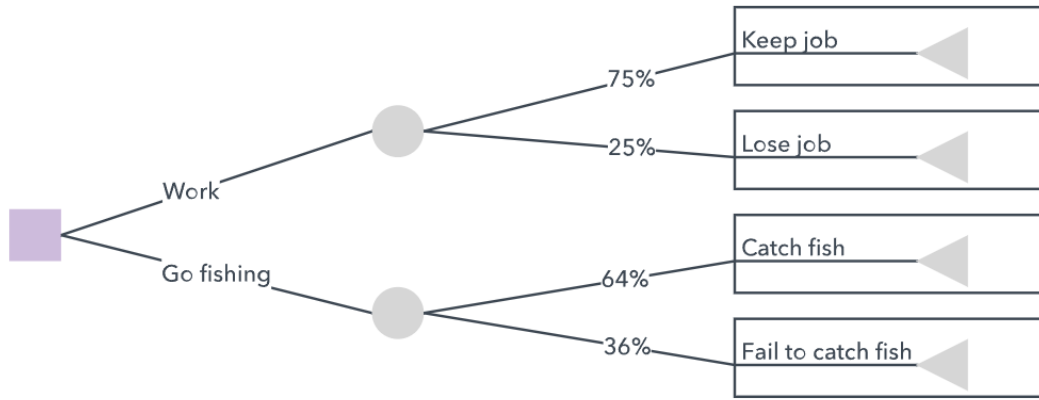


Imagen 21.- Ejemplo de un árbol de decisión. Fuente: (¿Qué Es Un Diagrama de Árbol de Decisión? | Lucidchart, n.d.)

### 5.1.3 Random Forest

El modelo *Random Forest* es un algoritmo de aprendizaje automático utilizado tanto para la clasificación como para la regresión. Es una técnica de conjunto que combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste. Al clasificar los datos, cada árbol genera una predicción de clase para una instancia de entrada determinada. La clase final de la instancia de entrada se decide entonces por votación de las predicciones de cada árbol individual. En otras palabras, se elige la clase elegida por la gran mayoría de los árboles. (Breiman, 2001).

La imagen 22 muestra un ejemplo simplificado de un bosque aleatorio de 3 árboles de decisión para la clasificación de dos clases (A y B). Basándose en un subconjunto de características elegidas al azar, cada árbol determina de qué tipo de objeto se trata. El resultado de una votación mayoritaria sobre las decisiones de cada árbol determina el resultado final de la categorización para una instancia de entrada dada. En este caso, la clase final proyectada sería azul porque los tres árboles indican que la instancia de entrada pertenece a la clase A.

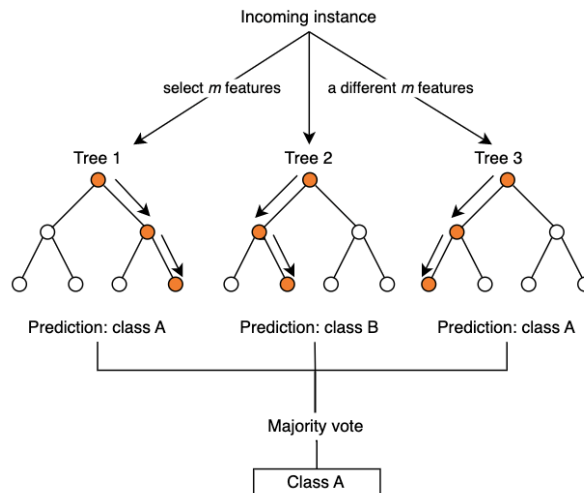


Imagen 22.- Ejemplo gráfico de un modelo Random Forest. Fuente: (Random Forests Definition | DeepAI, n.d.).



Dado que cada árbol se produce de forma independiente, el entrenamiento y la predicción pueden realizarse en paralelo, lo que aumenta la eficiencia computacional. Es por ello por lo que este modelo fue tomando en cuenta ya que es extremadamente adaptable a todo tipo de problemas gracias a que puede gestionar datos que faltan y variables categóricas.

#### 5.1.4 Redes neuronales (RNA)

El modelo de Redes Neuronales Artificiales (RNA) es un modelo de aprendizaje profundo utilizado para resolver problemas de clasificación y regresión. Se compone de numerosas capas interconectadas de neuronas que procesan y cambian los datos de entrada para producir una salida. En el caso de la clasificación binaria, la RNA utiliza la función de activación *sigmoid* para generar una probabilidad de pertenencia a la clase positiva o negativa. La arquitectura más comúnmente utilizada para la clasificación con RNA es la red neuronal multicapa, que consta de una capa de entrada, una o varias capas ocultas y una capa de salida. La capa de entrada recibe los datos de entrada, las capas ocultas procesan y transforman la información y la capa de salida genera una probabilidad de pertenencia a la clase positiva o negativa. (Clasificación de Redes Neuronales Artificiales - Diego Calvo, n.d.).

La imagen número 23 muestra un ejemplo de una red neuronal multicapa utilizada para la clasificación binaria. La capa de entrada recibe los datos de entrada y la capa de salida genera la probabilidad de pertenencia a la clase positiva o negativa.

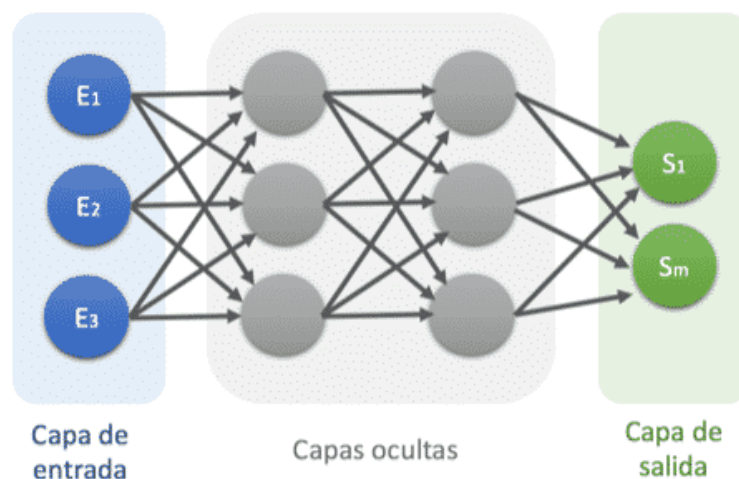


Imagen 23.- Explicación gráfica del funcionamiento del modelo de Red Neuronal. Fuente: (Clasificación de Redes Neuronales Artificiales - Diego Calvo, n.d.).

Durante el entrenamiento de la RNA, se utiliza una técnica llamada retropropagación del error para ajustar los pesos de las conexiones entre las neuronas y minimizar la función de pérdida. Una vez que se ha entrenado el modelo, se utiliza para hacer predicciones sobre nuevos datos de entrada.

## 5.2 Preparación de los datos para los modelos predictivos

Antes de crear un modelo predictivo, es necesario preparar los datos para que puedan ser utilizados adecuadamente. En esta sección, se describen los pasos llevados a cabo para la preparación de los datos, incluyendo la codificación de las variables categóricas, la estandarización de los datos y el tratamiento del desbalanceo de la variable a predecir.

### 5.2.1 Codificación de las variables categóricas para los modelos

El primer paso para conseguir entrenar al modelo predictivo es la codificación de variables categóricas. Como ya se ha comentado, las variables categóricas son aquellas variables que representan una cierta categoría y estas tienen que estar codificadas correctamente para que puedan ser utilizadas en los modelos predictivos. En el caso del estudio actual, se ha optado por utilizar la técnica de codificación *OneHotEncoder*, la cual se basa en transformar las variables categóricas en una matriz de características numéricas para entrenar así a los modelos predictivos. Dado que es más sencillo trabajar con datos numéricos a la hora de generar los algoritmos de aprendizaje, el uso de la técnica de codificación *OneHotEncoder* permite mejorar el rendimiento de los modelos. Además, el valor numérico dado a cada categoría impide que el modelo suponga que una es mayor que otra, lo que podría dar lugar a resultados de predicción inexactos. (Kumar & Rai, 2021).

En el caso de la base de datos del estudio, existen una serie de variables categóricas a tratar, como puede ser la variable "Sexo" o "Familiar\_TDAH". En el caso de la variable "Sexo", tiene tres categorías distintas: "Masculino", "Femenino" y "Indefinido". Al aplicar la codificación *OneHotEncoder*, se crean tres columnas en la matriz de características, una para cada categoría. Si un registro tiene el valor "Masculino" para la variable "Sexo", la columna correspondiente a "Masculino" tendrá un valor de 1, mientras que las otras dos columnas tendrán un valor de 0. Este proceso es necesario realizarlo para todas las variables categóricas.

Sexo_Femenino	Sexo_Indefinido	Sexo_Masculino	Familiar_TDAH_Si	Sintomas_No	Sintomas_Si
0	0	1	0	0	1
1	0	0	0	0	1
0	0	1	0	0	1
0	0	1	0	0	1
1	0	0	0	0	1

Tabla 15.- Tabla resultante de aplicar la técnica *OneHotEncoder* a la variable categórica "Sexo". Fuente: Propia.



### 5.2.2 Estandarización de los datos

El siguiente paso en el proceso de creación de los modelos es el de la estandarización de los datos. La estandarización se basa en ajustar los datos para que tengan una media de cero y una desviación estándar de uno, lo que permite que los datos se comparen y combinen de manera más efectiva en los modelos predictivos. Este apartado es fundamental ya que así se evita que las variables con escalas diferentes tengan un impacto desproporcionado en los análisis posteriores. De esta manera, los modelos pueden hacer comparaciones significativas entre las diferentes variables y producir predicciones más precisas. (Fernández-Delgado et al., 2014).

Para aplicar la estandarización a la base de datos, se utilizó la técnica *Z-score*. Esta técnica se basa en que, para cada variable numérica, se calcula su media y su desviación estándar. Luego, se resta la media a cada valor y se divide por la desviación estándar. Esto dará como resultado una nueva variable con una media de cero y una desviación estándar de uno.

Diagnostico_TDAH	Independiente	Idiomas	Atencion_Detalles_A	Atencion_Tiempo_A	Escuchar_A	Instrucciones_A	Organización_A
0	0.0	-0.709535	-0.694787	-1.254658	0.414320	-0.624958	-0.356257
1	0.0	0.296214	-0.694787	-0.429523	0.414320	-0.624958	-0.356257
2	0.0	-1.715285	-0.694787	0.395613	-0.403123	-1.485750	-1.132578
3	0.0	0.296214	0.915355	1.220748	-0.403123	-1.485750	-1.132578
4	0.0	0.296214	0.110284	1.220748	1.231764	1.096625	1.196385

5 rows x 29 columns

Tabla 16.- Tabla resultante al aplicar la técnica *Z-score* a las variables categóricas

### 5.2.3 Desbalanceo de la variable a predecir

Al analizar los datos, se observó que la variable a predecir, "Diagnostico\_TDAH", presentaba un desbalanceo en las clases. En concreto, se encontró que el 80% de los individuos en la base de datos no tenían diagnosticado el TDAH, mientras que solo el 20% restante lo tenían. Este desbalanceo puede llegar a generar problemas en la capacidad predictiva del modelo, ya que el algoritmo tenderá a clasificar la mayoría de las observaciones en la clase más frecuente.

Para poder solucionar este problema, se utilizó la técnica de *oversampling* mediante *SMOTE* (*Synthetic Minority Over-sampling Technique*). *SMOTE* es una técnica que consiste en generar nuevos datos sintéticos de la clase minoritaria, en base a patrones extraídos de los datos existentes. De esta manera, se logra aumentar la cantidad de datos de la clase minoritaria, disminuyendo así el desbalanceo de las clases. (Batista et al., 2004).

Para el estudio actual, se aplicó *SMOTE* para generar nuevos datos sintéticos de la clase minoritaria (individuos diagnosticados con TDAH) con el fin de que se igualara con el número de observaciones en ambas clases. Posteriormente, se volvió a realizar el entrenamiento del modelo con los datos balanceados y se evaluó su capacidad predictiva.

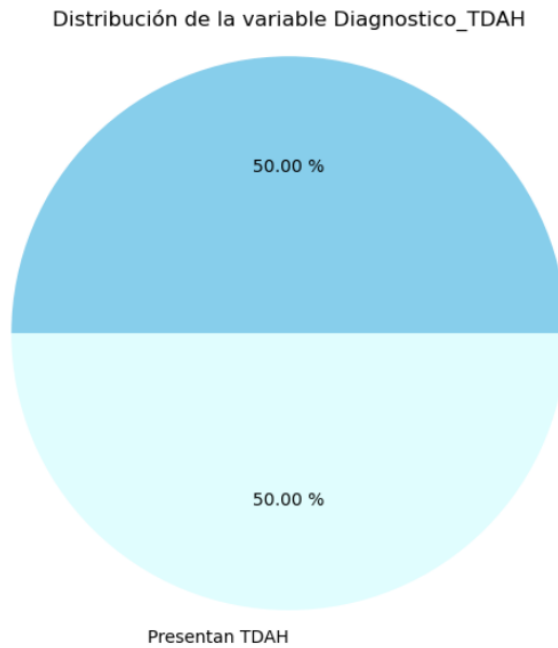


Imagen 24.- Distribución de la variable "Diagnostico\_TDAH" al aplicar la técnica SMOTE. Fuente: Propia.

### 5.3 Optimización de hiperparámetros

Un hiperparámetro es una variable predefinida que influye en la forma en que un modelo de aprendizaje automático aprende y produce predicciones. Los hiperparámetros se establecen manualmente antes de entrenar el modelo y no se modifican durante el entrenamiento, a diferencia de los parámetros del modelo, que se aprenden a partir de los datos de entrenamiento. La optimización de los hiperparámetros es importante para maximizar el rendimiento del modelo y evitar el sobreajuste. Al encontrar los mejores valores de los hiperparámetros, se asegura que el modelo tenga una precisión y generalización adecuadas en nuevos datos. (Li et al., 2018).

En el caso actual, esta optimización se ha llevado a cabo a través de la técnica *Grid SearchCV*. El *Grid SearchCV* es una técnica que se basa en encontrar los mejores hiperparámetros de un modelo a partir de una evaluación exhaustiva de una serie de valores predefinidos. Además, para evitar el sobreajuste y obtener una estimación más fiable del rendimiento del modelo, se aplicó al mismo tiempo la técnica de validación cruzada con 5 *folds*. Esta técnica será explicada y detallada en el apartado de evaluación y comparación de modelos (5.4).

La técnica de *Grid SearchCV* se aplicó tanto a los datos balanceados, es decir, los datos que se les había aplicado la técnica de *SMOTE*, como a los datos que no, apareciendo los resultados en las tablas 17 y 18 respectivamente:



Model	Best Parameters	Train Score	Test Score
svc	{'C': 1, 'kernel': 'rbf'}	88.51%	90.06%
dtc	{'max_depth': 10, 'min_samples_split': 2}	83.56%	90.99%
rfc	{'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}	87.91%	90.99%
mlp	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (10, 10)}	83.85%	89.13%

Tabla 17.- Tabla que muestra los resultados obtenidos al optimizar los hiperparámetros del conjunto de datos balanceado para los diferentes modelos predictivos estudiados. Fuente: Propia.

Model	Best Parameters	Train Score	Test Score
svc	{'C': 0.1, 'kernel': 'rbf'}	88.45%	88.46%
dtc	{'max_depth': 3, 'min_samples_split': 5}	85.68%	89.01%
rfc	{'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 50}	88.45%	88.46%
mlp	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (10, 10)}	87.90%	88.46%

Tabla 18.- Tabla que muestra los resultados obtenidos al optimizar los hiperparámetros del conjunto de datos balanceado para los diferentes modelos predictivos estudiados. Fuente: Propia.

Después de comparar los resultados de ambas tablas, se observa que los modelos ajustados con los datos balanceados obtienen una mejor capacidad de generalización en el conjunto de datos de prueba, lo cual indica que tiene una mayor capacidad de generalizar a datos nuevos y no vistos anteriormente. Por esta razón, se optó por seleccionar los modelos ajustados con la técnica *SMOTE* y los hiperparámetros optimizados a través de la técnica *GridSearchCV*.

## 5.4 Evaluación y comparación de los modelos predictivos

Sabiendo ya que modelos se han tenido en cuenta, es momento de hablar sobre los métodos de evaluación que se han usado para comparar los modelos y así poder elegir el que mejor se adapta a las características del estudio. Dentro del proceso de evaluación, existen varias métricas a tener en cuenta:

- **Accuracy:** Esta métrica se caracteriza por medir la proporción de casos clasificados correctamente por el modelo. Un valor alto de *accuracy* indica que el modelo es capaz de predecir correctamente un alto porcentaje de casos. Sin embargo, puede ser engañoso si el conjunto de datos está desbalanceado.
- **Precision:** Esta métrica mide la proporción de casos clasificados como positivos que realmente son positivos. Es decir, mide la capacidad del modelo para no clasificar como positivo a un caso negativo. Un valor alto de *precision* indica que el modelo es muy preciso en la clasificación de casos positivos.
- **Recall:** Esta métrica mide la proporción de casos positivos que son correctamente identificados por el modelo. Es decir, mide la capacidad del modelo para detectar casos

positivos. Un valor alto de *recall* indica que el modelo es capaz de detectar la mayoría de los casos positivos.

- **Sensitivity:** Esta métrica es una medida específica del rendimiento de los modelos de clasificación binaria. El objetivo es medir la proporción de verdaderos positivos que se identifican correctamente. Un valor alto de *sensitivity* indica que el modelo es capaz de detectar correctamente la mayoría de los casos positivos.
- **Specificity:** Esta métrica mide la proporción de verdaderos negativos que se identifican correctamente. Un valor alto de *specificity* indica que el modelo es capaz de detectar correctamente la mayoría de los casos negativos.

(Géron, 2019).

Todos estos valores se pueden calcular a partir de la matriz de confusión. La matriz de confusión es una tabla que muestra la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos del modelo. La diagonal principal de la matriz representa los casos clasificados correctamente, mientras que las otras dos celdas representan los casos clasificados incorrectamente. Como se puede ver en la imagen 25, a partir de esta matriz, se pueden calcular diferentes métricas de evaluación del modelo, como las métricas *sensitivity* y *specificity*.

		Predicted Class		
		No	Yes	
Observed Class	No	TN	FP	Model Performance
	Yes	FN	TP	

TN	True Negative	Accuracy	$= (TN+TP)/(TN+FP+FN+TP)$
FP	False Positive	Precision	$= TP/(FP+TP)$
FN	False Negative	Sensitivity	$= TP/(TP+FN)$
TP	True Positive	Specificity	$= TN/(TN+FP)$

Imagen 25.- Imagen explicativa de la matriz de confusión. Fuente: (Confusion Matrix | Everything About Data Science, n.d.).

En el caso del estudio actual, las métricas en las que se va a basar la comparación de modelos serán *sensitivity* y *specificity*. La métrica *sensitivity* permite saber la capacidad que tiene el modelo para identificar correctamente a los pacientes con TDAH, es decir, la proporción de pacientes con TDAH que son clasificados correctamente como positivos por el modelo. Por otro lado, la métrica *specificity* se refiere a la capacidad del modelo para identificar correctamente a los pacientes que no tienen TDAH, es decir, la proporción de pacientes sin TDAH que son clasificados correctamente como negativos por el modelo.

Estas medidas son importantes en el contexto del diagnóstico del TDAH, ya que se busca evitar tanto falsos positivos como falsos negativos. Un falso positivo ocurre cuando un paciente sin TDAH es clasificado como positivo por el modelo, lo que podría llevar a un diagnóstico erróneo y tratamientos innecesarios. Por otro lado, un falso negativo ocurre cuando un paciente con TDAH es clasificado como negativo, lo que podría llevar a la falta de tratamiento adecuado y a la persistencia de los síntomas del TDAH.

Además, otra técnica de validación que se usó fue la validación cruzada con *5 folds*. Esta técnica se utilizó para prevenir el sobreajuste en el proceso de optimización de hiperparámetros y para



evaluar el rendimiento del modelo de manera más precisa y prevenir el sobreajuste a los datos de entrenamiento.

En el caso de este estudio, se utiliza la técnica de validación cruzada *k-fold*, donde los datos se dividieron en *k* subconjuntos. Luego, se realizó *k* iteraciones del entrenamiento y evaluación del modelo para que en cada iteración se utilizase uno de los subconjuntos como conjunto de prueba y el resto como conjunto de entrenamiento. De esta forma, cada subconjunto es utilizado una vez como conjunto de prueba y *k-1* veces como conjunto de entrenamiento.

Para calcular las predicciones de cada modelo y optimizar sus hiperparámetros se utiliza una validación cruzada de 5 *folds*, es decir, se divide el conjunto de datos en 5 subconjuntos y se realizan 5 iteraciones. En cada iteración, se ajusta el modelo con los datos de entrenamiento y se evalúa el rendimiento en los datos de prueba. Luego, se calcula el promedio de los resultados de las 5 iteraciones como el rendimiento general del modelo.

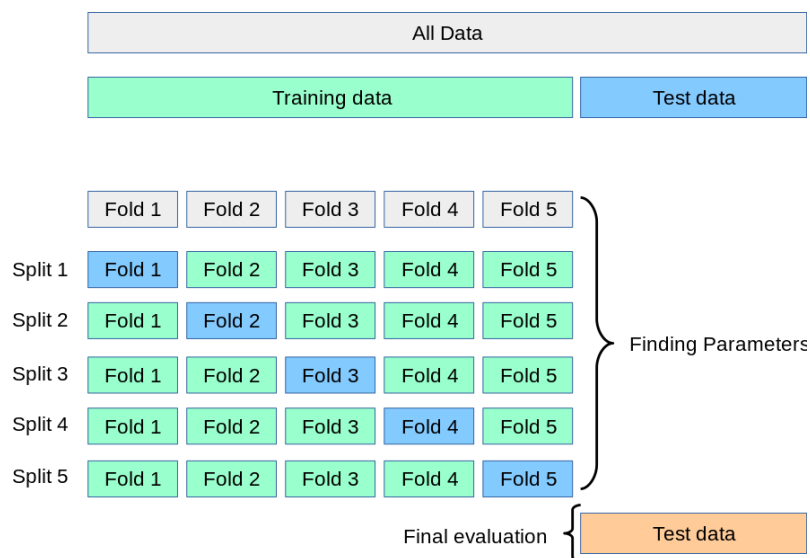


Imagen 26.- Imagen explicativa del funcionamiento de la validación cruzada con 5 folds. Fuente: (Cómo Realizar La Validación Cruzada y Cruce Seccional de Datos - Blog US.NUMERICA.MX, n.d.)

Otra de las métricas que se ha tenido en cuenta para evaluar los distintos modelos es la curva ROC. La curva ROC (*Receiver Operating Characteristic*) es una medida de la capacidad de un modelo para distinguir entre clases en un problema de clasificación binaria. En este caso, representa la relación entre la tasa de verdaderos positivos (*sensitivity*) y la tasa de falsos positivos (*Specificity*) a través de diferentes umbrales de probabilidad. La curva ROC se representa gráficamente trazando la tasa de verdaderos positivos en el eje Y y la tasa de falsos positivos en el eje X. La línea diagonal desde (0,0) hasta (1,1) representa la tasa de verdaderos positivos y la tasa de falsos positivos de un modelo aleatorio.(Fawcett, 2006).



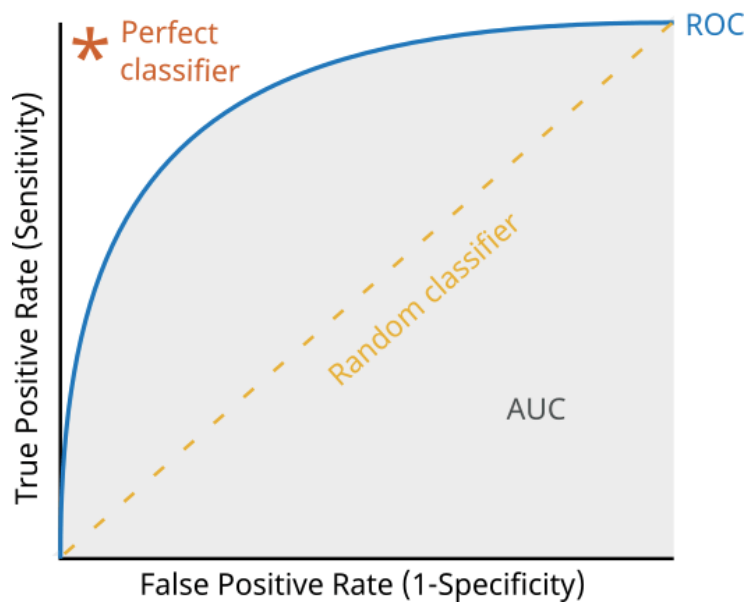


Imagen 27.- Ejemplo de Curva ROC. Fuente: (Compare Deep Learning Models Using ROC Curves - MATLAB & Simulink, n.d.).

Un modelo con una curva ROC que se aproxima al borde superior izquierdo del gráfico se considera un modelo con buena capacidad de clasificación. La medida de rendimiento de un modelo a través de la curva ROC es el área bajo la curva (AUC), que puede interpretarse como la probabilidad de que el modelo clasifique correctamente una instancia aleatoria de la clase positiva por encima de una instancia aleatoria de la clase negativa. Un modelo con un AUC de 1 es un modelo perfecto, mientras que un modelo con un AUC de 0.5 es tan bueno como un modelo aleatorio. (Fawcett, 2006).

#### 5.4.1 Análisis de los resultados obtenidos

A continuación, se procede a presentar los resultados obtenidos al aplicar las métricas explicadas en el capítulo anterior. En las tablas 19 y 20 se muestran los valores obtenidos para todos los algoritmos seleccionados para el estudio. Los resultados obtenidos se muestran para ambos conjuntos de datos: el original y el balanceado. De esta manera, se podrá determinar cuál de los modelos es el más efectivo para aplicarlo en el estudio de este trabajo final de grado.

	SMOTE	Accuracy	Precisión	Recall	Sensitivity	Specificity	AUC
Máquinas de soporte vectorial		86.60%	79.20%	99.40%	99.40%	73.90%	86.60%
Árboles de clasificación		81.10%	75.30%	92.50%	92.50%	69.60%	81.70%
Random Forest		83.90%	77.40%	95.70%	95.70%	72.00%	83.50%
Redes neuronales		81.70%	73.60%	98.80%	98.80%	64.60%	81.70%

Tabla 19.- Resultados de las métricas al evaluar los modelos con un conjunto de datos balanceados. Fuente: Propia.

NO SMOTE	Accuracy	Precisión	Recall	Sensitivity	Specificity
Máquinas de soporte vectorial	88.50%	0.00%	0.00%	0.00%	100.00%
Árboles de clasificación	71.40%	10.30%	19.00%	19.00%	78.30%
Random Forest	88.50%	0.00%	0.00%	0.00%	100.00%
Redes neuronales	88.50%	0.00%	0.00%	0.00%	100.00%

Tabla 20.- Resultados de las métricas al evaluar los modelos con un conjunto de datos no balanceados. Fuente: Propia.

Al comparar los valores de las diferentes métricas obtenidas en los datos no balanceados con los datos balanceados mediante *SMOTE*, se puede observar una clara mejora en el rendimiento de los modelos tras aplicar esta técnica de balanceo. La mayoría de los modelos muestran una sensibilidad relativamente baja en los datos no equilibrados, lo que les dificulta la identificación de la clase minoritaria. Sin embargo, en los datos equilibrados, la sensibilidad de todos los modelos ha aumentado significativamente, lo que indica que pueden detectar con mayor precisión la clase minoritaria. Además, se observa un aumento en la métrica *specificity* de algunos modelos, como los árboles de clasificación, lo que sugiere que son capaces de identificar la clase mayoritaria con mayor precisión.

Por otro lado, las métricas *sensitivity* y *specificity* se distribuyeron de forma más uniforme cuando se utiliza el conjunto de datos equilibrado de *SMOTE*. El modelo SVM tiene buenos resultados en la predicción de casos positivos y negativos, con un AUC de 0,866 y una *sensitivity* y *specificity* del 99,4% y el 73,9%, respectivamente. Por lo que se puede concluir que para el estudio actual es necesario aplicar la técnica *SMOTE*, ya que los datos balanceados presentan un mejor rendimiento en la detección de ambas clases.

## 5.5 Modelo final

Después de evaluar los resultados de los diferentes modelos predictivos utilizando las métricas de *sensitivity*, *specificity* y AUC, se decidió seleccionar el modelo de máquinas de soporte vectorial (SVM) para su implementación final.

En primer lugar, se observa que la implementación del método de balanceo de datos *SMOTE* ha mejorado significativamente los resultados de las métricas *sensitivity* y *specificity* en todos los modelos, incluyendo el SVM, que ya mostraba buenos resultados sin balanceo. Si bien algunos modelos como Árboles de Decisión y *Random Forest* tuvieron valores significativos en las métricas, el SVM destaca tanto en *sensitivity* como en AUC, que son de gran importancia para el estudio actual. Además, al evaluar los modelos sin balancear, el SVM es el único modelo que logra mantener una alta *specificity* (0.739), lo que significa que puede predecir correctamente los casos negativos, que son de especial interés para poder evitar diagnósticos erróneos. En cuanto a la curva ROC, el SVM obtiene un valor de AUC de 0.866, lo que indica que tiene una buena capacidad para distinguir entre casos positivos y negativos. Por todos estos motivos, se decidió decantarse por utilizar el modelo SVM.

Sabiendo cual es el modelo, se procede a explicar e interpretar los resultados obtenidos de éste. El modelo SVM aplicado al estudio es definido con los siguientes parámetros:

- $C = 1$ .
- $Kernel = 'rbf'$ .
- $Probability = True$ .

Estos parámetros determinan que se está utilizando una función de *kernel* radial para transformar los datos y un parámetro de coste  $C$  igual a 1 para penalizar los errores de clasificación. La función de *kernel* radial es útil para datos no linealmente separables, ya que permite proyectar los datos en un espacio de mayor dimensión donde se puedan separar más fácilmente. El parámetro de coste  $C$  controla el equilibrio entre la precisión y la generalización del modelo. (Cortes & Vapnik, 1995).

Además, se habilitó la opción de probabilidad para que el modelo pudiera proporcionar una estimación de la probabilidad de pertenencia a cada clase, en lugar de solo una etiqueta de clase. Esto permitió graficar la curva ROC del modelo y evaluar su desempeño de manera más completa.

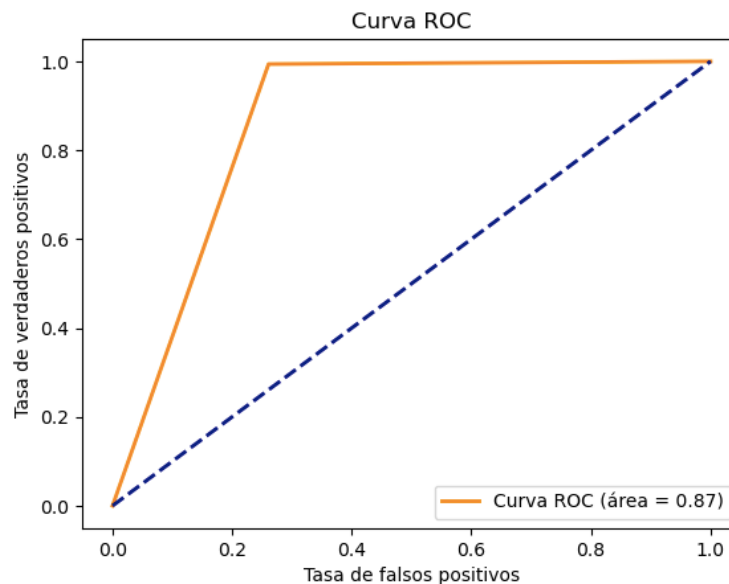


Imagen 28.- Curva ROC del modelo final. Fuente: Propia.

Para evaluar la importancia de cada característica en el modelo SVM seleccionado se ha utilizado la técnica de "*permutation importance*", la cual es una herramienta útil para medir la importancia de cada característica en un modelo predictivo. Esta técnica consiste en permutar aleatoriamente los valores de una característica en el conjunto de datos y medir el impacto que esto tiene en la precisión del modelo. Si la precisión disminuye significativamente después de permutar una característica, entonces se considera que esa característica es importante para el modelo.

Una vez entrenado el modelo, se obtiene el resultado mostrado en la imagen 29, donde destacan las características "Distraccion\_A", "Interrumpir\_H", "Mover\_H", "Atencion\_Tiempo\_A" y "Escuchar\_A" con unas puntuaciones de importancia de 0.02081, 0.01739, 0.00839, 0.00745 y 0.00621, respectivamente.

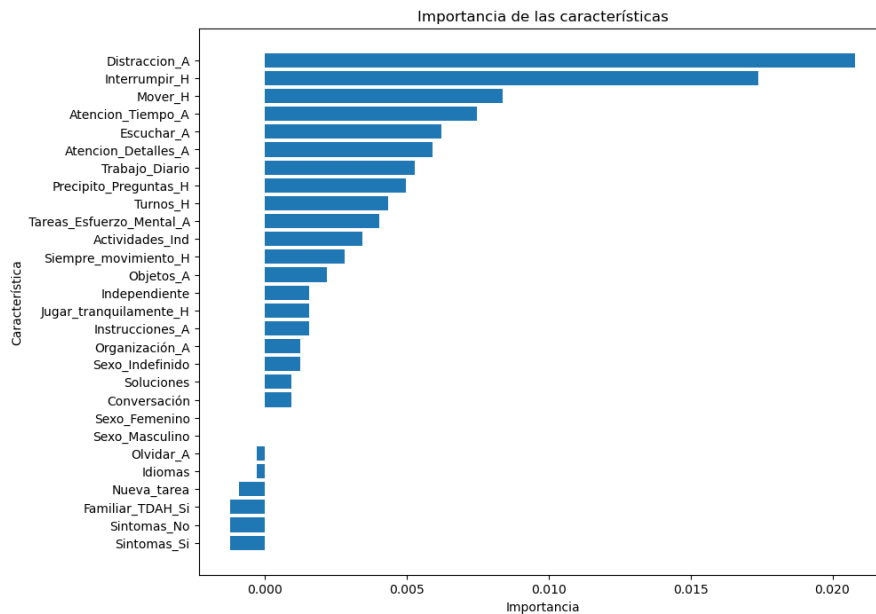


Imagen 29.- Distribución de la importancia de las variables para el modelo final con todas las variables. Fuente: Propia.

Como se puede observar, "Olvidar\_A", "Idiomas", "Nueva\_tarea", "Familiar\_TDAH\_Si" y "Síntomas\_No" tienen una importancia negativa, lo que indica que su inclusión en el modelo puede disminuir la precisión de éste. Una vez eliminadas, se volvió a entrenar el modelo y se obtuvieron los mismos resultados en cuanto a las métricas de *sensitivity* y *specificity*. Por esta razón, se optó por no eliminar estas características y mantenerlas ya que los resultados finales no varían.

Finalmente, hay que añadir que los valores más importantes que muestra el modelo corresponden a los valores que el DSM-V estandariza como síntomas relacionados con el TDAH. Asimismo, las preguntas añadidas personalmente debido a la experiencia de haber sufrido el TDAH también tienen una importancia relevante, por lo que esto justifica haberlas incluido en el modelo.

# 6. CONCLUSIONES DEL TRABAJO

---

Llegados a este punto del proyecto, es momento de concluirlo y analizar si se han podido cumplir todos los objetivos marcados. Además, también se propondrán futuras líneas de continuación de este trabajo final de grado y como el grado en Ciencia de Datos ha aportado conocimientos para desarrollar la solución final del estudio.

## 6.1 Conclusiones del análisis

Habiendo finalizado el estudio, es momento de realizar las conclusiones sobre los resultados obtenidos y de analizar si se han cumplidos los diferentes objetivos que se marcaron al inicio del proyecto. Como se ha comentado en los primeros apartados de la memoria, el objetivo principal siempre ha sido el poder desarrollar un modelo de aprendizaje automático capaz de detectar posibles casos de TDAH en niños y adolescentes. Teniendo en cuenta el resultado final, en el que se ha propuesto el uso del modelo predictivo SVM como posible método para el diagnóstico del TDAH, se puede afirmar que se ha cumplido totalmente.

El marco teórico sobre el TDAH ha permitido comprender mejor la realidad de este trastorno y entender la importancia de explorar nuevas alternativas para su tratamiento utilizando tecnologías innovadoras, como es el caso del estudio actual mediante la aplicación de *machine learning*. A través de esta revisión e investigación, se ha descubierto que existen proyectos y estudios que utilizan dichas técnicas. Sin embargo, aún se necesita seguir indagando y proponiendo nuevas formas para poder detectar tempranamente el TDAH. Es necesario continuar investigando en este campo para desarrollar nuevas herramientas y estrategias que permitan una detección temprana del TDAH, lo que ayudaría a mejorar el pronóstico de los pacientes y a prevenir posibles complicaciones a largo plazo.

Uno de los objetivos principales del proyecto era el de generar una base de datos para entrenar el modelo. Para ello, se diseñó un cuestionario en *Google Forms*, lo que permitió una mayor difusión y fácil acceso a los participantes. Sin embargo, a pesar de esta estrategia, no se logró obtener el número de muestras esperado. Dado que el tiempo para la realización del TFG era limitado, se decidió continuar con el proyecto utilizando el número de muestras conseguidas, aunque se esperaba un mayor número de participantes. Con una base de datos más amplia, se podría conseguir una muestra más representativa de la incidencia del TDAH en niños y adolescentes en España.

A nivel práctico, se ha realizado una exhaustiva limpieza de los datos gracias al uso de diferentes técnicas estadísticas a través del lenguaje de programación *Python*. Gracias a ello se ha logrado el objetivo de tener los datos limpios y preparados para entrenar a los diferentes modelos.

Con relación a los modelos, se ha conseguido obtener resultados prometedores y se han identificado los criterios específicos más relevantes para el diagnóstico del TDAH, que coinciden



con los establecidos en el DSM-V y con los síntomas personales añadidos. A través de las conclusiones, se ha podido determinar que los criterios específicos proporcionados por el DSM-V para el diagnóstico del TDAH tienen diferentes grados de relevancia, y se han identificado aquellos que son más importantes y aquellos que tienen menor peso en el diagnóstico. Además, al haber introducido nuevos criterios basados en mi experiencia personal, se han propuesto posibles criterios adicionales que podrían considerarse para su inclusión en el DSM-V, debido al peso significativo que han demostrado tener en el modelo. Estos criterios adicionales podrían ser útiles para mejorar la precisión del diagnóstico del TDAH, como evaluar las dificultades que un individuo puede tener para realizar tareas o actividades cotidianas, medir el grado de independencia de los individuos o determinar si tienen dificultades para iniciar o mantener una conversación. Este punto era el más importante a conseguir ya que engloba todos los demás y le da un punto de valor extra al obtener un resultado funcional.

Finalmente, cabe destacar que, como toda nueva tecnología, es necesario investigarla y nunca sustituir los métodos tradicionales, sino implementarlos a la vez para obtener el mejor resultado posible. La tecnología no debe reemplazar, debe aportar.

## 6.2 Trabajos futuros

Este apartado sirve para aportar posibles futuras mejoras e ideas de negocio que pueden originarse a partir del resultado final del estudio. Como se ha comentado en la memoria, los modelos de aprendizaje automático son tan buenos como los datos con los que se entrenan, por lo que se podría explorar la posibilidad de desarrollar un modelo de aprendizaje automático con un conjunto de muestras más amplio, tal y como se ha comentado en las conclusiones. Esto permitiría tener un conjunto de muestras que incluya registros de todas las comunidades autónomas para así poder analizar patrones de prevalencia del TDAH por región. Además, se podría incluir diferentes fuentes de datos, como registros médicos, de comportamiento y de rendimiento académico. Esto permitiría crear un perfil más completo y preciso de cada paciente y así mejorar la eficacia del modelo.

Una posible idea de negocio sería crear una plataforma digital sobre el TDAH. Esta plataforma tendría dos secciones, una para todo el público que serviría como fuente de información sobre el trastorno, y otra en la que solo se podría acceder si después de contestar el cuestionario saliese que el encuestado tiene posibilidades de padecer TDAH. En esta última sección, podrían tener acceso tanto el personal sanitario como los pacientes y que sirviese como medio de seguimiento de la enfermedad, permitiendo monitorizar los síntomas, incluyendo aquellos que son más personales y específicos para cada paciente. Además, se podrían generar cuestionarios periódicos con los síntomas más relevantes, y en base a los resultados se podrían introducir nuevas variables para mejorar continuamente el modelo. En conclusión, esta plataforma permitiría un seguimiento más personalizado y efectivo del TDAH, mejorando así la calidad de vida de los pacientes y proporcionando información valiosa para la investigación y el tratamiento de la enfermedad.

### 6.3 Relación del trabajo desarrollado con los estudios cursados

En este apartado, se pretende hacer un ejercicio de retro inspección sobre como el haber cursado el grado en Ciencia de Datos ha ayudado a conseguir el resultado final de este estudio. A lo largo del proyecto se utilizaron métodos de análisis exploratorio de datos para encontrar patrones y aspectos significativos en la base de datos creada a través del cuestionario. Estas técnicas fueron aprendidas en la asignatura de "Análisis exploratorio de los datos". Además, se utilizaron varios modelos predictivos, como los árboles de decisión o redes neuronales, los cuáles fueron estudiados en las asignaturas de Modelos Descriptivos y Predictivos. La evaluación y optimización de los modelos también ha sido una parte fundamental del trabajo, la cual fue dada en la asignatura de "Evaluación, Despliegue y Monitorización de Modelos". Por otra parte, también se utilizaron técnicas de visualización para presentar los resultados de manera clara y concisa, lo que está relacionado con la asignatura de "Visualización". Cabe decir que, aunque algunas asignaturas, como "Análisis exploratorio de datos" y "Modelos descriptivos y predictivos I", están más directamente relacionadas con las tecnologías utilizadas en este proyecto, también se aplicaron conocimientos adquiridos en otras asignaturas. Por ejemplo, se emplearon conceptos de metodología de proyectos y gestión de tareas de la asignatura "Proyecto", y se utilizaron los conceptos fundamentales de programación aprendidos en asignaturas como "Algorítmica" y "Programación".

Además, para llevar a cabo este proyecto, ha sido necesario aplicar una serie de competencias transversales aprendidas a lo largo de la carrera, como la capacidad de planificar y gestionar el tiempo de forma eficiente, el ser capaz de redactar un estudio de forma técnica, estar dispuesto a seguir aprendiendo continuamente o saber comunicarse de manera efectiva. Por lo tanto, queda claro que gracias a los conocimientos que el grado en Ciencia de Datos aporta a sus estudiantes, ha sido posible desarrollar este TFG.



## 7. BIBLIOGRAFÍA

---

- Alpaydin, E. (2010). *Introduction to Machine Learning* (2nd ed.). MIT Press.
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5)* (5th ed.).
- An Inquiry Into the Nature and Origin of Mental Derangement: Comprehending a ... - Sir Alexander Crichton - Google Libros.* (n.d.). Retrieved April 28, 2023, from [https://books.google.se/books?id=xHVJAAAAYAAJ&pg=PP9&hl=es&source=gbs\\_selected\\_pages&cad=3#v=onepage&q&f=false](https://books.google.se/books?id=xHVJAAAAYAAJ&pg=PP9&hl=es&source=gbs_selected_pages&cad=3#v=onepage&q&f=false)
- Barkley, R. A. (2014). Attention-deficit/hyperactivity disorder. *The Lancet*, 383(9922), 24–35.
- Barkley, R. A., & Murphy, K. R. (2010). *Attention-deficit hyperactivity disorder: A clinical workbook* (3rd ed.). Guilford Press .
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Biederman, J. (2005). Attention-deficit/hyperactivity disorder: A selective overview. *Biological Psychiatry*, 57(11), 1215–1220.
- Biederman, J., & Faraone, S. V. (2005). Attention-deficit hyperactivity disorder. *The Lancet*, 366(9481), 237–248.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brinkman, H., Niemeyer, J., & Van der Stoep, J. (2019). A machine learning approach to the diagnosis and treatment of ADHD. *Journal of Medical Systems*, 43(6), 1–9.
- Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 17(1), 113–126.
- Clasificación de redes neuronales artificiales - Diego Calvo.* (n.d.). Retrieved May 3, 2023, from <https://www.diegocalvo.es/clasificacion-de-redes-neuronales-artificiales/>
- Cómo Realizar la Validación Cruzada y Cruce Seccional de Datos - Blog US.NUMERICA.MX.* (n.d.). Retrieved May 5, 2023, from <https://us.numerica.mx/carreras-universitarias/como-realizar-la-validacion-cruzada-y-cruce-seccional-de-datos/>
- Compare Deep Learning Models Using ROC Curves - MATLAB & Simulink.* (n.d.). Retrieved May 7, 2023, from <https://www.mathworks.com/help/deeplearning/ug/compare-deep-learning-models-using-ROC-curves.html>
- Confusion Matrix | Everything About Data Science.* (n.d.). Retrieved May 5, 2023, from <http://scaryscientist.blogspot.com/2016/03/confusion-matrix.html>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.



- Cortese, S. (2018). The neurobiology and genetics of Attention-Deficit/Hyperactivity Disorder (ADHD): What every clinician should know. *European Journal of Paediatric Neurology*, 22(2), 215–228.
- DSM-5: Trastorno específico del aprendizaje.* (n.d.). Retrieved April 28, 2023, from <https://www.fundacioncadah.org/web/articulo/dsm-5-trastorno-especifico-del-aprendizaje.html>
- DuPaul, G. J., & Stoner, G. (2014). *ADHD in the schools: Assessment and intervention strategies* (3ª edición). Guilford Publications.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
- Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133–3181.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer .
- Haz una Infografía - Acceder.* (n.d.). Retrieved April 30, 2023, from <https://infograph.venngage.com/signin>
- Kumar, A., & Rai, P. (2021). Performance analysis of various machine learning algorithms on a bank marketing dataset using OneHotEncoder. *Journal of Ambient Intelligence and Humanized Computing*, 12(2), 1631–1641.
- Ley Orgánica de Protección de Datos - LOPDGDD 3/2018 | Grupo Atico34.* (n.d.). Retrieved May 7, 2023, from <https://protecciondatos-lopd.com/empresas/nueva-ley-proteccion-datos-2018/>
- Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., & Talwalkar, A. (2018). A System for Massively Parallel Hyperparameter Tuning. *ArXiv*.
- Máquinas Vectores de Soporte Clasificación – Teoría - Aprende IA.* (n.d.). Retrieved May 3, 2023, from <https://aprendeia.com/maquinas-vectores-de-soporte-clasificacion-teoria/>
- National Institute of Mental Health. (2016). Attention Deficit Hyperactivity Disorder (ADHD): The Basics. *National Institute of Mental Health*.
- Neuroimagen en el diagnóstico del TDAH - Neuropsicólogo Ulises Espino Rodríguez.* (n.d.). Retrieved April 28, 2023, from <https://ulises-espino.jimdofree.com/2013/08/28/estudios-de-imagenolog%C3%ADa-cerebral-que-apoyan-el-diagn%C3%B3stico-del-tdah/>
- Neurotransmisores: qué son, tipos y descripción de los más conocidos | Psyciencia.* (n.d.). Retrieved April 28, 2023, from <https://www.psyciencia.com/neurotransmisores-que-son-tipos-y-descripcion-de-los-mas-conocidos/>



- Ng, A. (2017). Machine learning for medical diagnosis: history, state of the art and future challenges. *IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, 247–252.
- Qué es el TDAH - Feaadah*. (n.d.). Retrieved April 28, 2023, from <https://www.feaadah.org/que-es-el-tdah/>
- ¿Qué es un diagrama de árbol de decisión? | Lucidchart*. (n.d.). Retrieved May 7, 2023, from <https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision>
- Random Forests Definition | DeepAI*. (n.d.). Retrieved May 3, 2023, from <https://deepai.org/machine-learning-glossary-and-terms/random-forest>
- RGPD - Resumen general del Reglamento | Sage España*. (n.d.). Retrieved May 7, 2023, from <https://www.sage.com/es-es/rgpd/>
- Rubia, K. (2018). *Neuroimaging in child and adolescent psychiatric disorders*. Cambridge University Press.
- Rutter, M., & Taylor, E. (2002). *Child and adolescent psychiatry: modern approaches*. Blackwell Science.
- Sánchez-Muñoz, M. A., Castellanos-Domínguez, G., & Olivares-Méndez, M. A. (2020). El uso de cuestionarios autoadministrados en la detección del TDAH. *Actas Españolas de Psiquiatría*, 48(6), 235–242.
- Treatment of ADHD | CDC*. (n.d.). Retrieved April 28, 2023, from <https://www.cdc.gov/ncbddd/adhd/treatment.html>

# 8. ANEXOS

---

## 8.1 Cuestionario

16/5/23, 21:22

Test de autoevaluación TDAH

### Test de autoevaluación TDAH

¡Hola! Mi nombre es Alejandra Sánchez, estudiante de último año de Ciencia de Datos en la Universidad Politécnica de Valencia. Actualmente, me encuentro realizando mi Trabajo Final de Grado, el cual consiste en investigar posibles nuevos patrones de conducta que ayuden a diagnosticar lo antes posible a personas que sufren TDAH. Este formulario es completamente anónimo y es el primer paso para recabar los datos necesarios para mi investigación. Es por ello que es muy importante que sea respondido siempre por la misma persona (por ejemplo, en el caso de niñas o niños pequeños que sean ellos quienes responden el formulario y no los padres, madres o personas legales a su cargo). ¡Muchas gracias por vuestra participación!

\* Indica que la pregunta es obligatoria

---

#### Preguntas generales

1. **Indica tu sexo:** \*

*Marca solo un óvalo.*

- Masculino
- Femenino
- Indefinido



2. **Indica en qué Comunidad Autónoma resides: \***

*Marca solo un óvalo.*

- Andalucía
- Aragón
- Islas Baleares
- Canarias
- Cantabria
- Castilla-La Mancha
- Castilla y León
- Cataluña
- Comunidad de Madrid
- Comunidad Foral de Navarra
- Comunidad Valenciana
- Extremadura
- Galicia
- País Vasco
- Principado de Asturias
- Región de Murcia
- La Rioja

3. **Indica tu nivel de educación actual: \***

*Marca solo un óvalo.*

- Primaria
- Secundaria
- Preparatoria o bachillerato
- Educación superior

4. **Indica a qué centro educativo asististe o has asistido en tu etapa escolar (Primaria):** \*

*Marca solo un óvalo.*

- Público
- Privado
- Concertado

5. **Soy una persona que:** \*

*Marca solo un óvalo.*

- Ha sido diagnosticada con Déficit de Atención
- No ha sido diagnosticada con Déficit de Atención, pero creo que puedo presentar algunos síntomas
- No ha sido diagnosticada con Déficit de Atención y considero que no presento ningún síntoma

6. **Indica cómo consideras que es tu relación familiar:** \*

*Marca solo un óvalo.*

- Muy buena
- Buena
- Mala
- Muy mala
- Ni buena, ni mala



16/5/23, 21:22

Test de autoevaluación TDAH

7. **Indica si sabes de algún familiar directo que presente déficit de atención: \***

*Marca solo un óvalo.*

- Sí  
 No  
 Tal vez

8. **Indica si eres hija/o único: \***

*Marca solo un óvalo.*

- Sí *Salta a la pregunta 10*  
 No *Salta a la pregunta 9*

**Pregunta familiar**

9. **Indica el número de hermanos/as que tienes: \***

*Marca solo un óvalo.*

- 1  
 2  
 3 o más

**Hábitos**

A continuación se mostrarán una serie de situaciones en las cuales tienes que elegir aquella opción con la que te sientas más identificado/a.

10. **Soy una persona que en mi tiempo libre disfruto de series, películas o libros: \***

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

11. **Soy una persona que disfruta haciendo deporte, me sirve para desconectar y no me supone mucho sacrificio: \***

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

12. **Me considero una persona activa en mi día a día: \***

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo



16/5/23, 21:22

Test de autoevaluación TDAH

13. **Me considero una persona empática. Me pongo en la piel de los demás para sentir de verdad lo que la otra persona está experimentando, sobre todo, si está pasando un mal momento:** \*

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

14. **Me considero una persona independiente, no dependo de otros para hacer algo:** \*

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

15. **Me considero una persona con facilidad para aprender idiomas:** \*

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo



16. **Me considero una persona constante, soy responsable y me esfuerzo por lograr mis metas:** \*

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

17. **Me considero una persona que en el caso de tener alguna duda pregunto o preguntaba a los profesores sin ningún tipo de problema:** \*

*Marca solo un óvalo.*

- Totalmente en desacuerdo
- En desacuerdo
- Ni de acuerdo ni en desacuerdo
- De acuerdo
- Totalmente de acuerdo

18. **El tipo de asignatura en la que más suelo o solía destacar es:** \*

*Marca solo un óvalo.*

- Ciencias
- Letras
- Arte
- Historia
- Ninguna de las anteriores



19. **Le dedico diariamente a las tareas escolares:** \*

Marca solo un óvalo.

- Menos de 30 minutos
- Entre 1 y 2 horas
- Entre 2 y 4 horas
- Más de 4 horas

**Síntomas**

En las siguientes afirmaciones se propondrán varias situaciones cotidianas donde tienes que elegir cómo de acuerdo está con ellas. La escala es la siguiente:

- 1 : Totalmente en desacuerdo.
- 2 : En desacuerdo.
- 3 : Ni de acuerdo ni en desacuerdo.
- 4 : De acuerdo.
- 5: Totalmente de acuerdo

20. **No presto atención a detalles o cometo errores por descuido:** \*

Marca solo un óvalo.

\_\_\_\_\_

1

\_\_\_\_\_

2

\_\_\_\_\_

3

\_\_\_\_\_

4

\_\_\_\_\_

5

\_\_\_\_\_

\_\_\_\_\_

21. **No mantengo la atención durante largos períodos de tiempo: \***

Marca    solo un óvalo.

—

1

2

3

4

5

—

22. **No escucho cuando me hablan: \***

Marca    solo un óvalo.

—

1

2

3

4

5

—

23. **No sigo instrucciones, no finalizo tareas:** \*

Marca solo un óvalo.

—

1

2

3

4

5

—

24. **Encuentro dificultades para organizarme:** \*

Marca solo un óvalo.

—

1

2

3

4

5

—

25. **Evito tareas de esfuerzo mental: \***

Marca    solo un óvalo.

—

1

2

3

4

5

—

26. **Pierdo objetos con mucha facilidad: \***

Marca    solo un óvalo.

—

1

2

3

4

5

—

27. **Me distraigo con estímulos externos: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

28. **Soy olvidadizo/a en las actividades diarias: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

29. **Muevo manos y pies en situaciones en las que debería estar tranquilo/a:** \*

Marca *solo* un óvalo.

1

2

3

4

5

16/5/23, 21:22

Test de autoevaluación TDAH

30. **Me levanto constantemente en situaciones en las que debería estar sentado/a:**

\*

Marca solo un óvalo.

—

1

2

3

4

5

—



31. **Corro y salto en situaciones inadecuadas:** \*

Marca    solo un óvalo.

—

1

2

3

4

5

—

32. **Tengo dificultad para jugar tranquilamente:** \*

Marca    solo un óvalo.

—

1

2

3

4

5

—

16/5/23, 21:22

Test de autoevaluación TDAH

33. **Tengo la necesidad de estar siempre en movimiento: \***

Marca solo un óvalo.

1

2

3

4

5

34. **Suelo hablar en exceso: \***

Marca solo un óvalo.

1

2

3

4

5

35. **Me precipito al responder preguntas: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

36. **Tengo dificultad para respetar los turnos: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

16/5/23, 21:22

Test de autoevaluación TDAH

37. **Interrumpo a los demás en sus actividades:** \*

Marca solo un óvalo.

1

2

3

4

5

38. **Tengo dificultades para encontrar soluciones a los problemas de la vida diaria:** \*

Marca solo un óvalo.

1

2

3

4

5

16/5/23, 21:22

Test de autoevaluación TDAH

39. **Tengo dificultades para aprender una nueva tarea, como por ejemplo, aprender cómo llegar a un nuevo lugar:** \*

Marca solo un óvalo.

—

1

2

3

4

5

—

40. **Tengo dificultades para comenzar y mantener una conversación: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

41. **Tengo dificultades para relacionarme con personas que no conozco: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

42. **Tengo dificultades para mantener una amistad: \***

Marca solo un óvalo.

—

1

2

3

4

5

—

43. **Tengo dificultades para llevarme bien con personas cercanas a mi: \***

Marca solo un óvalo.

—

1

2

3

4

5

—



44. **Tengo dificultades para hacer nuevos amigos:** \*

Marca solo un óvalo.

—

1

2

3

4

5

—

16/5/23, 21:22

Test de autoevaluación TDAH

45. **Tengo dificultades para llevar a cabo mi trabajo diario o las actividades escolares diarias:** \*

Marca solo un óvalo.

—  
1  —  
2  —  
3  —  
4  —  
5  —  
—

46. **Tengo facilidad para posponer las tareas diarias: \***

Marca *solo* un óvalo.

—

1

2

3

4

5

—

16/5/23, 21:22

Test de autoevaluación TDAH

47. **Tengo dificultades para participar en actividades en grupo: \***

Marca solo un óvalo.

1

2

3

4

5

**48. Tengo dificultades para participar en actividades individuales: \***

Marca solo un óvalo.

1

2

3

4

5

**49. Respecto las afirmaciones anteriores, ¿has marcado algún 4 o un 5? \***

Marca solo un óvalo.

Sí Salta a la pregunta 50

No

**Aparición de los síntomas****50. ¿Estos síntomas han estado presentes antes de los 12 años? \***

Marca solo un óvalo.

Sí

No



## 8.2 Librerías utilizadas en Python

```
In [1]: # Tratamiento de datos y análisis estadístico
# -----
import numpy as np           # Manipulación de arrays y operaciones matemáticas
import pandas as pd         # Manipulación y análisis de datos tabulares
import seaborn as sns       # Visualización de datos estadísticos
import scipy.stats as stats  # Funciones estadísticas
from scipy.stats import chi2_contingency # Prueba de independencia de Chi-cuadrado

# Preprocesado
# -----
from sklearn.model_selection import train_test_split # División de datos en conjuntos de entrenamiento y prueba
from sklearn.preprocessing import MinMaxScaler, StandardScaler # Escalado de características

# Optimización de parámetros y evaluación
# -----
from sklearn.model_selection import GridSearchCV, KFold, cross_val_predict # Optimización de parámetros
from sklearn.metrics import (
    f1_score, make_scorer, confusion_matrix, accuracy_score,
    precision_score, recall_score, roc_curve, auc # Métricas de evaluación del modelo
)

# Modelado
# -----
from sklearn.svm import SVC # Máquina de vectores de soporte (SVM)
from sklearn.tree import DecisionTreeClassifier # Árbol de decisión
from sklearn.ensemble import RandomForestClassifier # Bosque aleatorio
from sklearn.neural_network import MLPClassifier # Red neuronal multicapa
from sklearn.linear_model import LogisticRegression # Regresión logística
from sklearn.model_selection import cross_val_score # Validación cruzada del modelo
from imblearn.over_sampling import SMOTE # Técnica de sobre-muestreo para abordar el desequilibrio de clases
```

Imagen 30.- Diferentes librerías utilizadas en Python. Fuente: Propia.

### 8.3 Curva ROC de los modelos analizados

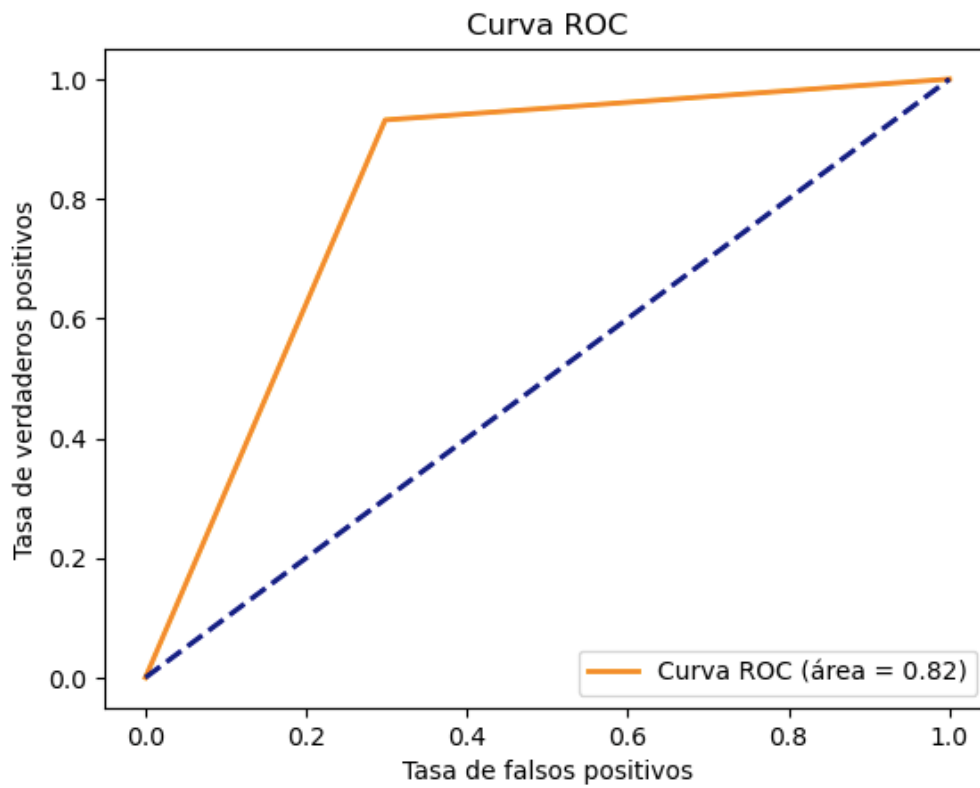


Imagen 31.- Curva ROC del modelo de árbol de clasificación. Fuente: Propia.

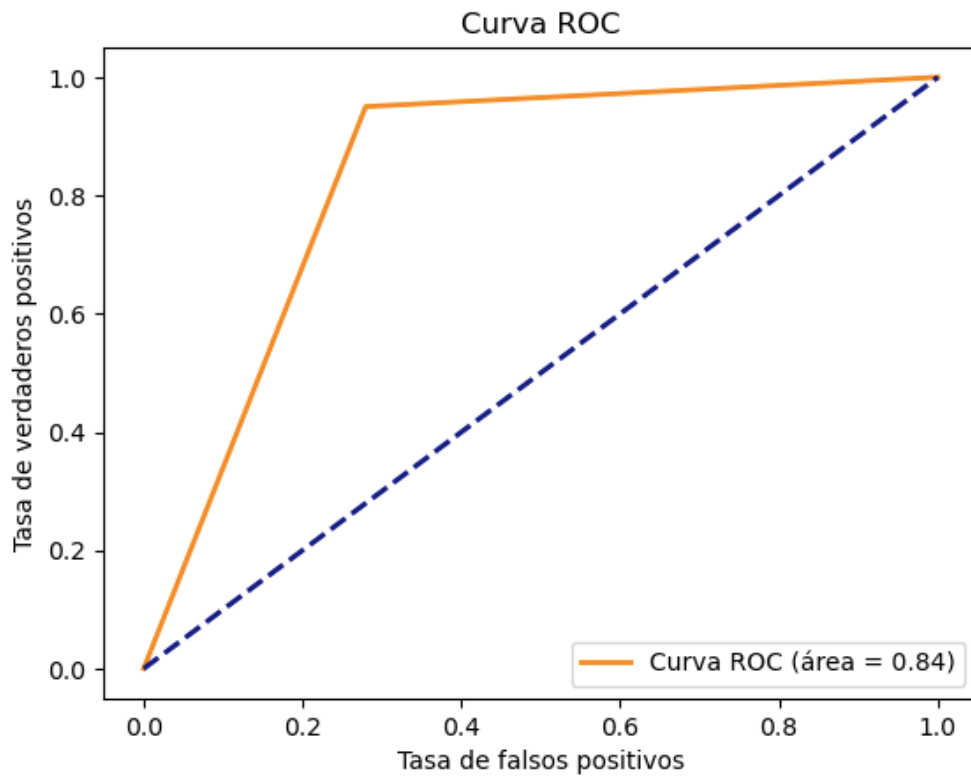


Imagen 32.- Curva ROC del modelo Random Forest. Fuente: Propia.

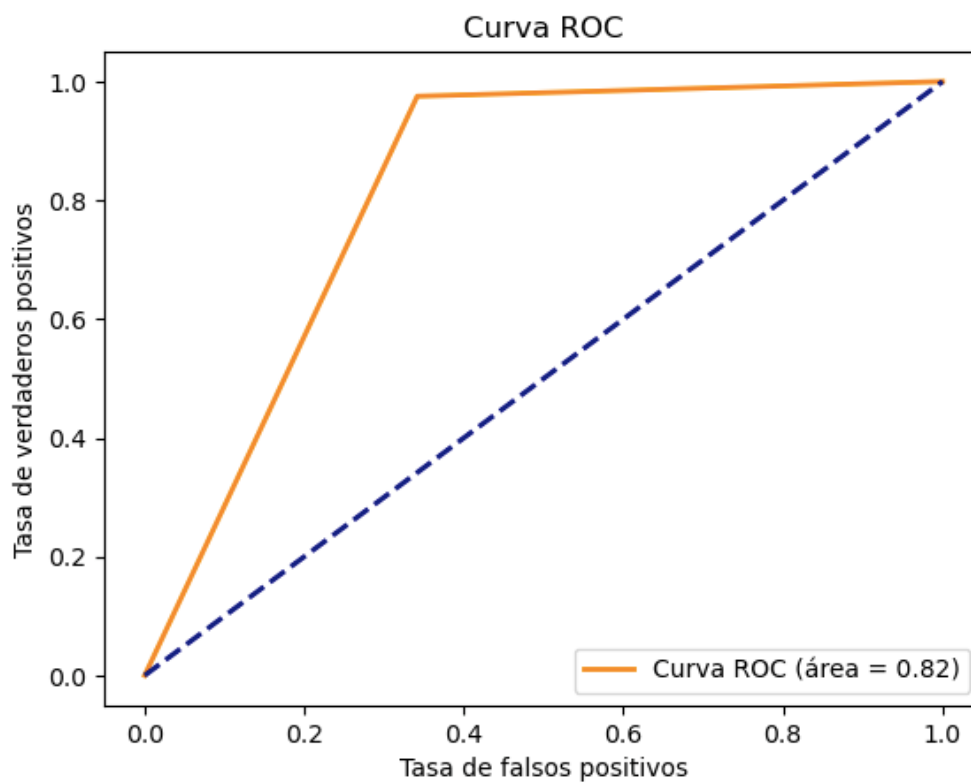


Imagen 33.- Curva ROC del modelo de Redes Neuronales. Fuente: Propia.



## 8.4 Relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS)

Los objetivos de Desarrollo Sostenible (ODS) son una serie de objetivos marcados por las Naciones Unidas en 2015, con el fin de poner fin a la pobreza, proteger el planeta y garantizar que para el 2030 todas las personas disfruten de paz y prosperidad. Por lo que respecta al presente trabajo final de grado, de los 17 objetivos marcados, se han identificado que el estudio contribuye en los siguientes:

- ODS 3: Salud y Bienestar. Debido a que el objetivo principal del trabajo es el de conseguir una temprana detección del TDAH tanto en niños como adolescentes, esto puede conseguir una mejora en la calidad de vida de los afectados notoria.
- ODS 4: Educación de calidad. Durante el desarrollo del proyecto se ha podido comprobar como el TDAH tiene un enorme impacto en el rendimiento académico. No solo por las propias dificultades que pueda tener el afectado, sino porque las técnicas de enseñanza convencionales pueden no ser suficientes para según que niveles de TDAH. Es por ello, que, si se detecta lo antes posible, se puede preparar al personal docente para que adapte las metodologías de enseñanza.
- ODS 10: Reducción de las desigualdades. Al ser un trastorno que afecta a todos los ámbitos de la vida cotidiana, pueden crearse desigualdades a raíz de ello. Con una temprana detección y un tratamiento personalizado, se podría llegar a evitar estas situaciones.
- ODS 17: Alianzas para lograr objetivos. Gracias al haber colaborado con distintas asociaciones del TDAH, se ha podido generar una base de datos para crear y entrenar el modelo predictivo.

