Revista Iberoamericana de Automática e Informática Industrial

Revista Iberoamericana de Automática e Informática Industrial 21 (2024) 205-217

www.revista-riai.org

Aprendizaje automático en el diagnóstico médico. Un caso de estudio en la identificación del Trastorno del Espectro Autista a partir del comportamiento ocular

Roberto Chávez-Trujillo[a,*], Rosa M. Aguilar[b], José Luis González-Mora[c]

[a]*Escuela de Doctorado y Estudios de Posgrado, Universidad de La Laguna, Avda. Astrofísico Francisco Sánchez, s/n. 38271, La Laguna, España*
[b]*Departamento de Ingeniería Informática y de Sistemas. Universidad de La Laguna, Camino San Francisco de Paula, s/n. 38271, La Laguna, España*
[c]*Departamento de Ciencias Médicas Básicas, Facultad de Ciencias de La Salud. Universidad de La Laguna, C/ Sta. María Soledad, s/n. 38200, La Laguna, España*

To cite this article: Chávez-Trujillo, R., Aguilar, R.M., González-Mora, José Luis. 2024. Machine learning in medical diagnosis. A case study in the identification of Autism Spectrum Disorder from ocular behaviour. Revista Iberoamericana de Automática e Informática Industrial 21, 205-217. https://doi.org/10.4995/riai.2024.20484

Resumen

A pesar de los avances recientes, el diagnóstico del autismo sigue siendo un desafío complejo debido a la necesidad de recursos médicos especializados, tiempo y materiales. Esto a menudo resulta en diagnósticos tardíos, incluso en la edad adulta, dificultando las intervenciones efectivas. Por otro lado, el campo de la inteligencia artificial y el aprendizaje automático ha experimentado un notable progreso. Estas técnicas han abierto nuevas oportunidades entre otras muchas áreas, en el diagnóstico médico, incluyendo el Trastorno del Espectro Autista (TEA). El objetivo principal de este artículo es ofrecer una visión general de la aplicabilidad de las técnicas de aprendizaje automático en el diagnóstico médico, a través de un caso de uso específico en el TEA. Empleando datos de seguimiento ocular, se ha desarrollado un modelo de clasificación basado en el algoritmo *XGBoost*, que logra una sensibilidad del 82 % y una especificidad del 74 % al clasificar muestras individuales. Además, al combinar este modelo con un algoritmo de votación por mayoría, se obtienen unos muy destacados resultados de clasificación en el conjunto de pruebas.

Palabras clave: Análisis de bio-señales, Análisis de datos, Inteligencia artificial, Aprendizaje automático.

Machine learning in medical diagnosis. A case study in the identification of Autism Spectrum Disorder from ocular behaviour.

Abstract

Despite recent advances, autism diagnosis remains a complex challenge due to the need for specialized medical resources, time, and materials. This often leads to late diagnoses, even in adulthood, hindering effective interventions. On the other hand, the field of artificial intelligence and machine learning has seen remarkable progress. These techniques have opened up new opportunities in various areas, including medical diagnosis and Autism Spectrum Disorder (ASD). The primary objective of this article is to provide a general overview of the applicability of machine learning techniques in medical diagnosis, using a specific case of ASD as an example. A classification model based on the XGBoost algorithm has been developed, achieving a sensitivity of 82 % and a specificity of 74 % when classifying individual samples. Furthermore, by combining this model with a majority voting algorithm, highly noteworthy classification results are obtained in the test set.

Keywords: Bio-signals analysis, Data analysis, Artificial intelligence, Machine learning.

1. Introducción

Según la definición actualmente aceptada en los manuales de diagnóstico internacionalmente reconocidos (CIE-11, DSM-5), el Trastorno del Espectro Autista (TEA) se relaciona con un desarrollo cerebral atípico, generalmente evidente en los comportamientos sociales de las personas que lo padecen (American Psychiatric Association, 2013). Por lo general, algunos sig-

*Autor para correspondencia: rchavezt@ull.edu.es

nos del trastorno suelen manifestarse alrededor de los dos o tres años de edad, aunque en otras ocasiones pueden pasar desapercibidos. Esta patología se caracteriza por deficiencias significativas en la interacción social y la comunicación, así como por mostrar intereses restringidos y comportamientos estereotipados (Rosen et al., 2021). Estas características limitan el desarrollo natural de un individuo, especialmente en las áreas de la comunicación y el comportamiento social (Preeti et al., 2017), lo que impacta en su capacidad para percibir y comprender los pensamientos, intenciones y emociones de las personas con las que se relacionan.

Hasta el momento, no se conoce con certeza cuál es la causa de este trastorno, pero se especula con distintos orígenes. La mayoría de estudios indican que posiblemente la causa más predominante sea la de origen genético (Mendelsohn and Schaefer, 2008), aunque aún no se conocen con exactitud cuáles son los genes involucrados o si se debe a interacciones entre múltiples de ellos (Abrahams and Geschwind, 2008). Otras investigaciones relacionan el aumento reciente de casos de autismo con una amplia variedad de causas externas, como contaminantes ambientales (D. Nevison Cynthia, 2014) o circunstancias parentales como la edad y el sobrepeso (Xu et al., 2013; Libbey et al., 2005). Aunque en algunos casos se puede observar que la tendencia al alza de estos contaminantes puede ser compatible con el aumento de la incidencia, no se ha llegado a una conclusión definitiva sobre su causalidad.

Sumado a esto, es importante destacar que este desafío se vuelve aún más crucial debido al aumento significativo en la incidencia del trastorno durante el tiempo en que se tienen registros (Maenner et al., 2020). Desde la década de 1980, ha habido un drástico aumento en la incidencia del TEA, en parte debido a una mayor conciencia sobre el trastorno y modificaciones en los criterios de diagnóstico. Sin embargo, no todo el incremento puede atribuirse a estos factores, y la proporción exacta de aumento genuino aún no se conoce con certeza. Según datos publicados por los Centros para el Control y la Prevención de Enfermedades (CDC), los casos diagnosticados no han dejado de aumentar, pasando, por ejemplo, de 1 de cada 150 niños en el año 2000 a 1 de cada 54 en 2016, lo que representa un aumento del 154 % (Centers for Disease Control and Prevention, CDC, 2020).

Por lo tanto, dada esta perspectiva y la falta de medios de prevención eficaces y tratamientos suficientemente efectivos, es de vital importancia diagnosticar el TEA a una edad temprana (Chawarska et al., 2013), ya que se ha demostrado que, con una pronta intervención y un entorno educativo adecuado, se pueden lograr mejoras significativas, especialmente en estos primeros años cuando el cerebro aún está en pleno desarrollo.

1.1. TEA y Comportamiento ocular

Vista la complejidad del diagnóstico, la carencia de herramientas simples y universales y el aumento en la incidencia, se propone explorar la aplicabilidad y utilidad de métodos basados en el aprendizaje automático para la clasificación y análisis de datos en el contexto de esta patología. Los datos utilizados provienen de un dispositivo de seguimiento ocular, que registra el comportamiento visual de los sujetos con TEA frente a los controles.

En la presente investigación, nos centraremos en el comportamiento visual de sujetos con TEA, ya que es uno de los mejores indicadores que puede revelar la presencia de este trastorno de manera temprana, sencilla y efectiva. Esta alteración en la mirada ha sido confirmada, incluso en niños de 16 meses de edad, y además sirve como indicador de la gravedad del autismo y de las futuras capacidades sociales y verbales del individuo (Hannah Furfaro, Spectrum, 2019).

Por esta razón, la investigación sobre el procesamiento facial en el TEA ha atraído mucha atención en los últimos años, especialmente considerando que la comunicación y la interacción interpersonal se basan en gran medida en la interpretación de las señales visuales y no verbales de aquellos con los que interactuamos (Tanaka and Sung, 2013). Es conocido el hecho de que las personas con TEA muestran una atención reducida hacia los rostros humanos, especialmente en el área de los ojos. Según (Tanaka and Sung, 2013), este comportamiento se asocia con una estrategia adaptativa para evitar la amenaza social y la incomodidad causada por el contacto visual directo. De hecho, se ha demostrado que este comportamiento es identificable en niños de tan solo un año de edad (Zwaigenbaum et al., 2004), e incluso antes (Jones and Klin, 2013).

1.2. Aprendizaje Automático

Teniendo en consideración el tipo y la cantidad de datos con los que se debe trabajar, resulta justificable emplear técnicas de inteligencia artificial, específicamente del ámbito del aprendizaje automático, para llevar a cabo un análisis efectivo de los mismos. El aprendizaje automático es un subcampo de la inteligencia artificial que se centra en la investigación y construcción de sistemas capaces de aprender a partir de los datos, en contraposición al seguimiento de instrucciones explícitas predefinidas. Estos sistemas tienen la capacidad de "aprender" de manera automática una función que mapea desde los datos de entrada hasta la salida deseada, basándose en ejemplos, y pueden aplicarse a tareas complejas para las cuales no es factible diseñar algoritmos explícitos (Bishop, 2019).

En los últimos años, estas técnicas han experimentado un desarrollo notable y han demostrado su utilidad en diversos campos, tales como la visión artificial, el procesamiento del lenguaje natural, la robótica o la medicina, entre otros. Aunque en el ámbito del aprendizaje automático existen algoritmos "clásicos" que han sido ampliamente estudiados, como los árboles de decisión, las máquinas de soporte vectorial y el método de los k-vecinos más cercanos, en las últimas décadas ha surgido un creciente interés en el aprendizaje profundo, especialmente en las redes neuronales. Estas poderosas técnicas han revolucionado la forma en que abordamos problemas complejos relacionados con el procesamiento de datos y el análisis de patrones.

Las redes neuronales artificiales, son un tipo de modelo de aprendizaje automático inspirado en la anatomía y el funcionamiento del cerebro humano. Por regla general, y como uno de sus aspectos distintivos radica en su estructura, compuestas por múltiples capas de neuronas interconectadas, que procesan y transforman la información de entrada para producir una salida. Cada neurona artificial realiza una serie de simples operaciones matemáticas en las que pondera con determinados pesos sus entradas, y transmite su resultado a las neuronas de la capa siguiente, tras hacer pasar el resultado por una cierta función

no lineal. Lo que hace que estas redes sean tan poderosas es su capacidad para aprender automáticamente a partir de los datos, ajustando los pesos de las conexiones entre neuronas a medida que se les presenta un conjunto de ejemplos de entrada y sus correspondientes salidas deseadas. Se puede entender a estos modelos como un "aproximador universal" de funciones (Schmidhuber, 2015; Goodfellow et al., 2016).

En el panorama actual del aprendizaje automático, las redes neuronales han recibido una atención considerable debido a su capacidad para abordar una amplia gama de problemas y su destreza en tareas que involucran datos no estructurados, como visión por computadora y procesamiento del lenguaje natural (NLP) (LeCun et al., 2015). Estas redes profundas han demostrado ser excelentes herramientas en la resolución de problemas complejos, lo que ha llevado a su predominancia en diversas áreas. Sin embargo, cuando se trata de problemas que implican datos tabulares, como en el caso del problema que nos ocupa, la elección del algoritmo adecuado es crucial. A pesar del éxito de las redes neuronales en otros dominios, existen razones sólidas para considerar el uso de "XGBoost", también conocido como *Gradient Boosted Trees*, como una alternativa viable.

XGBoost es ampliamente reconocido en la comunidad científica por su excelente rendimiento y su capacidad para lidiar directamente con datos tabulares, considerándose como el "state of the art" para este tipo de problemas. Este algoritmo destaca por su eficacia en la captura de relaciones no lineales y patrones complejos en los datos tabulares, lo que resulta especialmente relevante en problemas donde las interacciones entre las características pueden ser sutiles y difíciles de modelar. Además, XGBoost ofrece ventajas en términos de eficiencia computacional y tiempo de entrenamiento, lo que lo convierte en una opción valiosa en entornos donde los recursos pueden ser limitados.

El algoritmo XGBoost, según su implementación descrita por (Chen and Guestrin, 2016), está basado en *ensembles* (combinaciones de modelos) de un tipo de árboles de decisión denominados "CART" (*Classification and Regression Trees*). A diferencia de los árboles de decisión convencionales, estos asignan un valor real a cada "hoja" del modelo, lo que permite abordar tanto problemas de regresión como de clasificación, además de facilitar la interpretabilidad del modelo. La optimización de los parámetros, en este caso, la estructura del árbol, se lleva a cabo mediante el algoritmo de "descenso de gradiente", de manera aditiva, agregando un nuevo árbol en cada paso, como forma de abordar la complejidad del problema (Friedman, 2001).

Los árboles de decisión, en su concepción más simple, son estructuras que dividen el espacio de entrada en regiones homogéneas según ciertas reglas. Sin embargo, un solo árbol suele ser demasiado simple para capturar la complejidad de los datos reales, por lo que, en casos como XGBoost, se recurre a métodos de agrupamiento, que combinan varios árboles para obtener un modelo más robusto y preciso. En esta estrategia, la idea principal radica en que al agrupar varios modelos "débiles", se puede obtener un mejor rendimiento que con cualquiera de ellos por separado. Dentro de las diversas técnicas de combinación existentes, XGBoost emplea el *boosting*, que ha demostrado ser de las más efectivas. Este método se basa en la construcción secuencial de "modelos débiles", donde cada uno de ellos se centra en corregir los errores del anterior, lo que permite capturar relaciones complejas en los datos y aumentar la precisión final.

1.3. Aprendizaje Automático y TEA

Atendiendo a los múltiples beneficios que proporcionan estas técnicas para el análisis de grandes volúmenes de datos, no resulta inusual que se les encuentre aplicabilidad en campos de estudio cada vez más diversos, siendo uno de los más prometedores, el diagnóstico médico.

Un reciente metanálisis (Wei et al., 2023), resalta la utilidad y potencial de uso del aprendizaje automático en el diagnóstico del Trastorno a partir de datos de seguimiento ocular, aunque identifica ciertas debilidades comunes a gran parte de los estudios revisados. Por un lado, se encuentra el problema del reducido tamaño muestral, como ya se ha introducido con anterioridad y, por otro lado, se aprecia la extendida falta de un conjunto individual de testeo para el modelo de clasificación.

En este contexto, se debe mencionar el estudio de (Liu et al., 2015), en el que se aplicaron técnicas de aprendizaje automático sobre datos de la dinámica ocular para descubrir patrones de movimiento, indicativos de la presencia del TEA. Para ello emplearon únicamente imágenes de rostros humanos y herramientas de agrupamiento como *k-means* para identificar las áreas que presentaban un mayor interés visual. Con esta metodología, generaron histogramas de fijación visual, que usados como base para entrenar un modelo SVM, les permitió alcanzar una precisión del 80 %.

Por otro lado, en (Jiang and Zhao, 2017), los autores también emplearon los patrones de movimiento entre ambos grupos como indicador de la presencia del trastorno. Sin embargo, la principal innovación del estudio consistió en el uso de técnicas de aprendizaje profundo, y un algoritmo de selección de imágenes. Más concretamente, definieron una serie de métricas (duración y amplitud de las fijaciones, distancia al centro de la pantalla, etc.), que podrían caracterizar el comportamiento visual de los individuos. Posteriormente, mediante un método de puntuación pudieron escoger las imágenes que maximizaban la diferencia entre ambos grupos, de modo que el entrenamiento del algoritmo se pudiera llevar a cabo con las escenas más relevantes.

Finalmente, emplearon una red neuronal para la extracción de un vector de características de cada una de las imágenes que sirvió de entrada al algoritmo de clasificación (SVM), logrando una precisión final del 90 %.

Estos estudios, junto con el metanálisis señalado, subrayan la importancia de un enfoque metodológico adecuado, al mismo tiempo que sustentan la viabilidad del uso de estas técnicas para la detección de una patología tan compleja como el TEA. Nuestro trabajo busca contribuir a la mejora en el conocimiento y detección de este trastorno por medio de herramientas simples a la par que potentes, como los algoritmos de *ensembles* y una clasificación individualizada.

2. Métodos

Como se ha mencionado en la Introducción, el objetivo de la presente investigación es servir como ejemplo de caso de uso

de técnicas de aprendizaje automático en el diagnóstico médico, específicamente para el Trastorno del espectro autista. Todo ello, como paso previo a un estudio más completo y pormenorizado con el objetivo final de lograr una herramienta de diagnóstico confiable. Definiendo aún más, se presentará el desarrollo de un algoritmo de aprendizaje automático funcional que, basándose exclusivamente en los datos obtenidos por un software de seguimiento ocular, sea capaz de clasificar a una serie de individuos en los grupos "con TEA" o "sin TEA", con un alto nivel de fiabilidad.

2.1. Estructura de datos

Inicialmente, para este estudio disponemos de datos de la dinámica ocular de 16 jóvenes participantes, número que aumentará a medida que avance la investigación.

La muestra estudiada consistió en 6 sujetos sanos sin historial de enfermedades clínicas o psiquiátricas (controles), con una distribución por sexos de 3 mujeres y 3 hombres de edades comprendidas entre 20 y 23 años (media= 21.5 ± 1.2) y 6 sujetos con TEA (todos varones) de edades comprendidas entre 5 y 24 años (media= 14.8 ± 7.9).

Sin embargo, existen 4 sujetos adicionales cuya categorización es incierta. Estos individuos, etiquetados como "Dudosos", presentan características que no permiten una clasificación clara ni como "Controles"ni como casos de TEA, y podrían incluso corresponder a otras patologías con sintomatología similar. Debido a esta ambigüedad diagnóstica, y sin pérdida en la capacidad de generalización, se ha decidido no incluir a estos sujetos en el conjunto de datos empleado para el entrenamiento de los modelos, a fin de mantener la claridad y precisión en la clasificación de los grupos de estudio.

A modo de resumen, los 16 participantes se encuentran distribuidos en los siguientes grupos:

- Seis individuos diagnosticados con TEA, etiquetados: Casos.
- Seis individuos sin patología, etiquetados: Controles.
- Cuatro individuos sin un diagnóstico claro, que no han podido ser clasificados por los profesionales médicos en ninguna de las dos categorías, etiquetados: Dudosos.

A continuación, se describe el procedimiento seguido para la adquisición de datos:

Cada participante es acomodado en una habitación preparada para llevar a cabo el experimento, minimizando todos los estímulos externos. Seguidamente, se les muestra a través de una pantalla una sucesión de once presentaciones (tamaño de 960x720 px), cada una con dos escenas (imágenes) de variada tipología, como caras, personas, paisajes, objetos, etc., específicamente seleccionadas por el personal médico experimentado, con el objetivo de hacer detectables las diferencias en la respuesta ocular entre ambos grupos. Adicionalmente, entre cada escena se ha configurado una pantalla de calibración (*escena de fijación*) con el fin de corregir la desviación en la mirada y establecer una línea base para ajustar otras variables. Durante el tiempo que dura la misma (2 segundos), solo es visible una marca en forma de cruz en el centro de la pantalla, a la cual se debe dirigir la mirada para continuar con el experimento.

Paralelamente, mientras los sujetos examinan estas imágenes, se recopila información visual (coordenadas del punto de fijación, tipo de movimiento ocular, duración de los movimientos, etc.) mediante un dispositivo de seguimiento ocular de la marca Tobii, modelo TX-300, con una resolución de grabación de 1024x768 px y una frecuencia de muestreo de 300 Hz, por lo tanto, obteniendo una nueva muestra cada 3-4 ms. Cada una de las escenas es mostrada durante 4 segundos.

La Figura 1 muestra el aspecto general del archivo y los tipos de datos proporcionados. En promedio, cada archivo (uno por participante) suele tener un tamaño en el rango de 44,000 registros de datos (filas) y 50 características (columnas). Por lo tanto, es evidente que identificar un trastorno como el TEA a partir de datos sobre la dinámica ocular puede no ser una tarea trivial, teniendo en cuenta la enorme cantidad de datos generados por el dispositivo de seguimiento, las múltiples variables existentes y las relaciones entre ellas. En definitiva, se hace esencial confiar en algoritmos modernos para el procesamiento y análisis de datos, destacando las técnicas incluidas en el campo del aprendizaje automático.

GazePointLeftY (ADCSmm)	GazePointRightX (ADCSmm)	GazePointRightY (ADCSmm)	StrictAverageGazePointX (ADCSmm)	StrictAverageGazePointY (ADCSmm)	EyePosLeftX (ADCSmm)	EyePosLeftY (ADCSmm)	EyePosLeftZ (ADCSmm)	EyePosRightX (ADCSmm)	EyePosRightY (ADCSmm)	EyePosRightZ (ADCSmm)	DistanceLeft	DistanceRight	PupilLeft	PupilRight	TEA
165.1	397.59	158.37	393.59	161.73	249.79	212.33	609.82	310.25	211.09	608.83	646.09	646.09	2.98	2.91	1
166.79	398.06	155.01	394.1	160.9	249.79	212.35	609.88	310.26	211.09	608.84	646.16	646.16	2.96	2.94	1
166	397.23	152.68	393.62	159.34	249.79	212.36	609.9	310.25	211.09	608.83	646.18	646.18	2.97	2.94	1
166.82	399.35	159.46	394.3	163.14	249.79	212.36	609.89	310.25	211.09	608.82	646.17	646.17	2.96	2.88	1
170.71	397.93	154.81	394.42	162.76	249.79	212.36	609.9	310.25	211.09	608.84	646.18	646.18	2.92	2.94	1
161.69	397.95	154.13	393.82	157.91	249.79	212.36	609.9	310.25	211.08	608.83	646.18	646.18	3	2.93	1
171.19	397.51	158.61	394.08	164.9	249.79	212.36	609.91	310.25	211.08	608.82	646.2	646.2	2.92	2.89	1
166.7	397.27	153.69	393.96	160.19	249.79	212.36	609.92	310.25	211.08	608.82	646.2	646.2	2.94	2.93	1
164.46	396.74	153.87	392.48	159.16	249.79	212.36	609.92	310.25	211.08	608.82	646.2	646.2	2.97	2.93	1
164.71	396.97	152.63	393.93	158.67	249.79	212.36	609.91	310.25	211.08	608.82	646.19	646.19	2.95	2.92	1
170.59	397.04	153.15	393.74	161.87	249.79	212.34	609.85	310.25	211.09	608.83	646.13	646.13	2.9	2.92	1
159.87	398.85	153.75	393.88	156.81	249.8	212.4	609.94	310.26	211.15	608.92	646.23	646.23	2.98	2.92	1
171.72	398.84	161.2	394.74	166.46	249.82	212.42	609.89	310.26	211.21	608.96	646.2	646.2	2.93	2.88	1
177.48	399.83	172.79	394.83	175.14	249.84	212.5	609.96	310.25	211.27	609	646.29	646.29	2.93	2.86	1
182.22	400.96	175.54	396.01	178.88	249.85	212.56	609.99	310.22	211.35	609.06	646.34	646.34	2.94	2.9	1
188.2	400.27	180.31	395.01	184.26	249.86	212.65	610.07	310.18	211.35	608.94	646.45	646.45	2.93	2.9	1
191.6	398.93	187.13	393.31	189.36	249.87	212.73	610.16	310.17	211.41	609	646.57	646.57	2.92	2.9	1
197.48	396.29	186.18	391.92	191.83	249.87	212.76	610.18	310.15	211.45	609.01	646.59	646.59	2.9	2.91	1

Figura 1: Muestra parcial de datos obtenidos por el dispositivo de seguimiento ocular.

De las 50 características disponibles, muchas de ellas no son relevantes para el análisis, como las marcas de tiempo o aquellas que muestran una alta correlación, ya que representan medidas individuales para cada ojo por separado. En el primer estudio, se decidió seleccionar solo 9 de ellas (las más relevantes desde el punto de vista médico), las cuáles son: *SceneName*, *GazeEventType*, *GazeEventDuration*, *FixationPointX*, *FixationPointY*, *GazePointIndex*, *StrictAverageGazePointX*, *StrictAverageGazePointY* y *TEA*. El significado de cada característica se presenta a continuación:

- SceneName - Indica a cuál de las escenas, (incluyendo también la escena de fijación), corresponden los datos registrados.
- GazeEventType - Tipo de evento de movimiento ocular clasificado por la configuración del filtro de fijación. Puede tomar los valores Fijación, Sacada o Sin clasificar.
- GazeEventDuration - Duración de cada movimiento ocular registrado, [ms].
- FixationPointX, FixationPointY - Coordenadas horizontales y verticales del punto de fijación sobre la escena, [píxeles].
- GazePointIndex - Representa el orden en el que la muestra fue adquirida por el software Tobii Studio. El índice es un número autoincremental que comienza con 1 (primera muestra de la mirada).

- StrictAverageGazePointX, StrictAverageGazePointY - Coordenadas horizontales y verticales del punto de la mirada promediado para ambos ojos sobre la pantalla, [mm].
- TEA - Variable objetivo. Esta es una variable categórica añadida manualmente al conjunto de datos que representa la 'clase' a la que pertenece cada observación. Solo toma los valores '0' o '1', (Control, TEA, respectivamente).
- id - Identificador del individuo que participa en el experimento. Esta variable no se utiliza directamente en el algoritmo de clasificación, solo se emplea para formar los conjuntos de entrenamiento, prueba y validación, como se detalla en la sección 2.2.

Además de las variables medidas por el propio dispositivo de seguimiento ocular, hemos generado otras variables derivadas de éstas, dada su potencial utilidad en el proceso de clasificación.

- r - Módulo resultante de la conversión a coordenadas polares de las variables de posición StrictAverageGazePointX (ADCSmm) y StrictAverageGazePointY (ADCSmm), [mm].
- theta - Ángulo resultante de la conversión a coordenadas polares de las variables de posición StrictAverageGazePointX (ADCSmm) y StrictAverageGazePointY (ADCSmm), [rad].
- A1...C4 - Se clasifican los puntos de las variables FixationPointX (MCSpx) y FixationPointY (MCSpx) en base a una rejilla de 256 x 256 px superpuesta sobre la imagen. Se genera una nueva columna para cada una de las rejillas, con el valor asignado '1' si el punto específico está dentro de esta cuadrícula y '0' en caso contrario, consultar la Figura 3.
- DurACC - Variable en la que se almacena la suma acumulada de GazeEventDuration, durante cada escena. Se pone a 0 en cada escena y cambio de individuo, [ms].

2.2. *Preprocesado, Ingeniería de características*

Para el entrenamiento del algoritmo de clasificación, hemos optado por el enfoque clásico basado en la división en un conjunto de entrenamiento y validación en una proporción del 70-30 % respectivamente, además de un tercer conjunto utilizado solo para el testeo final del algoritmo (Figura 2). Aunque la partición óptima no es universal, y depende en gran medida del tipo de problema y la estructura de los datos, numerosas investigaciones se refieren a la división del 70-30 % como uno de las mejores en términos de rendimiento (Nguyen et al., 2021).

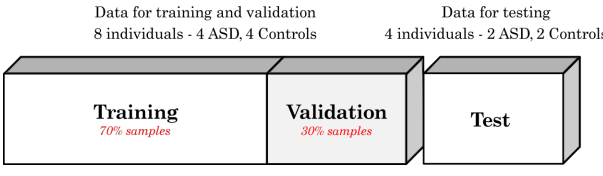


Figura 2: Estructura de los datos y particiones empleadas para el entrenamiento, validación y testeo del algoritmo.

Además, contrariamente a la tendencia de uso generalizada, y como se deduce del estudio realizado en (Vabalas et al., 2019), el simple método de división entre entrenamiento y prueba bien ejecutado, proporciona resultados más robustos que los obtenidos por otros métodos como la validación cruzada K-fold, al menos en casos donde las muestras disponibles son limitadas, como es el caso en este estudio.

Teniendo en cuenta que en esta investigación trabajamos con muestras de diferentes individuos, la división se ha llevado a cabo teniendo en cuenta esta particularidad, es decir, tratando siempre de mantener el conjunto de datos equilibrado en cuanto al número de muestras de individuos con TEA y controles (Tabla 1). En particular, se asignaron ocho individuos (cuatro de cada clase) al conjunto de entrenamiento y validación, y cuatro (dos con TEA y dos controles) al conjunto de pruebas (Figura 2).

Continuando con la estructura de los datos, es importante destacar algunos aspectos clave. En primer lugar, disponemos de un conjunto de datos equilibrado entre las dos clases: TEA y controles (Tabla 1). Este equilibrio es significativo y deseable, ya que, como se ha estudiado ampliamente (López et al., 2013), una distribución desigual puede dar lugar a resultados de clasificación inesperados, por medio de la obtención de un modelo sesgado hacia la clase mayoritaria. Por otro lado, como en cualquier problema en el ámbito de la ciencia de datos, es necesario realizar una "limpieza" y preparación previa de los datos para tratar con aspectos como: las muestras faltantes, valores anómalos y la conversión de tipos con métodos como *Label Encoding* o *One Hot Encoding*. En este caso, al analizar ambos conjuntos, vemos que para ciertas variables hay una cantidad significativa de datos *NaN* o vacíos (Tabla 2), especialmente en *FixationPointX* y *FixationPointY* (un 36 % de datos faltantes del total) y en *StrictAverageGazePointX* y *StrictAverageGazePointY* (un 10 % de datos faltantes).

Tabla 1: Distribución de las muestras por clase y conjunto de datos. La clase 0 corresponde a los individuos de control y la clase 1 a los diagnosticados de TEA. Ambos conjuntos de datos están perfectamente equilibrados entre las clases.

	Training and Validation	Testing
Clase 0	161.929	80.966
Clase 1	161.975	80.965
Total	323.904	161.931

Tabla 2: Número de datos NaN por variable y conjunto de datos

	Training and Validation NaN	Test NaN
SceneName	0	0
GazeEventType	0	0
GazeEventDuration	0	0
FixationPointX	117.379	40.311
FixationPointY	117.379	40.311
GazePointIndex	45	0
StrictAveragePointX	34.940	9.558
StrictAveragePointY	34.940	9.558
TEA	0	0
id	0	0

Para llevar a cabo el tratamiento de estas variables, es útil conocer cómo se han generado. Como se mencionó en la anterior sección 2.1, las variables FixationPoint muestran las coordenadas en píxeles solo de los movimientos oculares registrados como *fijaciones*, sin considerar los eventos *sacadas* y las

no-clasificadas, por lo que este alto nivel de datos vacíos es razonable. Para las variables de StrictAverageGazePoint, se ha decidido eliminar todas las muestras con datos faltantes, ya que se ha encontrado que en la mayoría de estos casos (92 %), tampoco hay datos disponibles para el resto de variables que indican ubicación de la mirada. Estos casos, son ejemplo de una medida para el aseguramiento de la calidad de los datos implementada en el propio software de adquisición. Dado que el sistema empleado para realizar el seguimiento ocular se basa en la detección de la pupila y en las reflexiones que la luz infrarroja emitida provoca sobre ella, cuando el software no es capaz de detectar esta información con un determinado umbral de certeza, directamente, se descartan estas muestras.

Posteriormente, continuando con la preparación de los datos, se procede a la corrección del tipo de datos y codificación de las variables no numéricas. Dado que no es posible conocer de antemano cuál será el mejor método, se han creado dos conjuntos de datos, uno con una codificación numérica simple (Tabla 3) y el otro utilizando la técnica de 'One-Hot Encoding', cada uno con sus ventajas e inconvenientes (Seger, 2018).

Tabla 3: Variables codificadas numéricamente.

SceneName		GazeEventType		id_train		id_test	
fijación	0	Fixation	0	03p	1	07p	9
scene1	1	Saccade	1	04p	2	08p	10
scene2	2	Unclassified	2	05p	3	05c	11
scene3	3			06p	4	09c	12
...	...			01c	5		
scene20	20			02c	6		
scene21	21			03c	7		
scene22	22			04c	8		

En esencia, mediante la codificación one-hot se suele mejorar la precisión de la clasificación, pero a costa de obtener conjuntos de datos con representaciones de alta dimensionalidad (se generan tantas nuevas variables como valores posibles de la variable original), lo que aumenta el costo en tiempo de entrenamiento y cómputo. Por otro lado, el uso de una codificación numérica simple es notablemente más rápido, pero puede llevar a "malentendidos" por parte del algoritmo que empeoran el resultado. Puede suceder que se asigne un mayor peso en la clasificación a aquellas etiquetas con un valor más alto, en nuestro caso, esto puede dar como resultado que el algoritmo otorgue más importancia a las últimas escenas.

Como se introdujo previamente, se han generado nuevas variables a partir de los datos disponibles, con la intención de facilitar la clasificación de las muestras. En este caso, siguiendo la idea presentada en (Cristino et al., 2010), hemos procedido a generar una cuadrícula ficticia sobre las imágenes, de manera que podamos obtener medidas derivadas como, qué áreas reciben atención durante más tiempo o, dónde podemos encontrar más fijaciones. Específicamente, la subdivisión elegida consiste en doce cuadros (4x3) sobre las imágenes de 1024x768 píxeles, como se puede ver en la siguiente Figura de ejemplo (Figura 3).

El propósito principal de esta idea es la implementación de un procedimiento para la comparación de secuencias de fijaciones y estudiar su posible uso como otra variable de entrada para los algoritmos de aprendizaje automático que ayude a mejorar la predicción.

El enfoque empleado es la comparación de secuencias de fijaciones, sacadas, etc., tratando los datos como una serie temporal y teniendo en cuenta aspectos que no son visibles en las imágenes estáticas, como la secuencialidad de los datos, es decir, la trayectoria y el tiempo que cada individuo pasa en las diferentes áreas. Existe una amplia variedad de métodos que se pueden utilizar para analizar la similitud de dos conjuntos/series de datos, como DTW (*Dynamic Time Warping*), u otros basados en *cosine similarity* (Cassisi et al., 2012; Goshtasby, 2012).

En nuestro caso hemos optado por otra técnica. Esta se basa simplemente en la conversión de las fijaciones en secuencias de movimientos oculares. A modo de ejemplo, si contamos con tres zonas etiquetadas como A, B y C con cierto tiempo de fijación en cada una de ellas de 80, 120 y 60 ms respectivamente, teniendo en cuenta la dimensión temporal y estableciendo una ventana de 20 ms, obtendríamos en orden cronológico "AAAA" en la primera área de interés (ROI), "BBBBBB" en la segunda y "CCC" en la tercera, dando como resultado una secuencia final "AAAABBBBBBCCC".

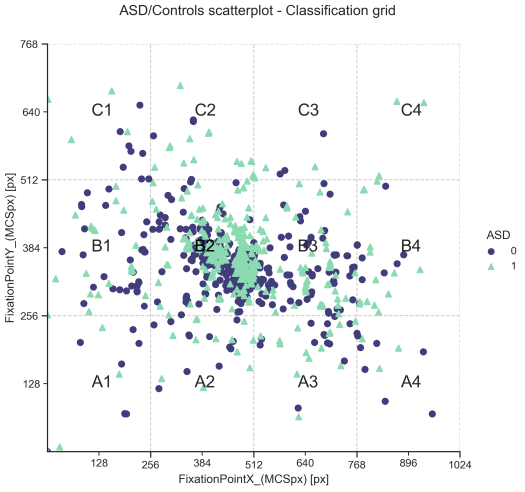


Figura 3: Ejemplo de cuadrícula sobre el diagrama de dispersión que muestra las fijaciones para la escena de calibración inicial para ambas clases, TEA y Controles (1,0) respectivamente.

Para llevar a cabo esta tarea, contamos con regiones de interés específicas (ROIs). En nuestro caso, estas ROIs se presentan como una cuadrícula etiquetada. Hemos optado por llevar a cabo nuestra investigación de manera "a ciegas", lo que implica que desconocemos la imagen real que se encuentra detrás de cada una de las escenas. Esto nos impide establecer las ROIs de forma manual, pero evitando al mismo tiempo hacer suposiciones previas sobre las imágenes que pueden influir en la clasificación. Puede observarse que la secuencia se genera con los nombres de las áreas donde se encuentra la fijación, y si se desea tener en cuenta también la dimensión temporal, solo es necesario repetir esa etiqueta de manera proporcional al tiempo que dure la fijación.

Una vez obtenida la secuencia, es posible aplicar técnicas de análisis en busca de similitudes, que van desde los algoritmos básicos de comparación y medición de distancia (Goshtasby, 2012), hasta el uso de aprendizaje automático/redes neuronales. La opción seleccionada va en línea con la idea presentada en (Cristino et al., 2010). De modo que, emplearemos algoritmos de alineación de secuencias (Day, 2010), que son ampliamente utilizados en bioinformática, como el algoritmo

Needleman-Wunsch (búsqueda de alineación global) (Needleman and Wunsch, 1970), que es utilizado comúnmente en el estudio de proteínas y ácidos nucleicos.

Dentro del proceso de ingeniería de características, procedemos a generar otra variable, *DurACC*, que se obtiene a partir de la suma acumulada de las duraciones de cada evento registrado por el dispositivo de seguimiento ocular (variable *GazeEventDuration*). Esta variable es igualmente fundamental, ya que suple la falta de una variable temporal en el conjunto de datos, abriendo la puerta al estudio de otros parámetros clave, como las velocidades y aceleraciones de los movimientos oculares, así como otros estudios dinámicos que se detallarán en las secciones siguientes.

2.3. Escalado y selección de características

Dentro del flujo de procesamiento de datos, como paso previo a la implementación de los modelos de clasificación, se ha realizado la eliminación de valores extremos utilizando el valor del rango intercuartílico, IQR, como una de las técnicas más eficientes en el caso de variables distribuidas normalmente (Kaltenbach, 2011). En consecuencia, definimos el IQR como la diferencia entre el tercer y el primer cuartil, $Q_3 - Q_1$, de modo que los datos ubicados fuera de k veces el rango intercuartílico son descartados. Por lo tanto, una muestra dada se considera un valor atípico si su valor es menor que $Q_1 - k \cdot IQR$ o mayor que $Q_3 + k \cdot IQR$.

En este caso particular, al establecer el valor de "k" en 3, se reconocen como *outliers* aproximadamente 12,000 muestras en el conjunto de entrenamiento (Tabla 4). Hemos podido detectar que, en casi todas las variables, las muestras clasificadas como outliers provienen de individuos con Trastorno del Espectro Autista (estos individuos contribuyen en torno al 70 % - 95 % del total, dependiendo del valor de k), y ocurren principalmente en la escena de calibración, denominada "fijación". Es importante señalar que con un valor de $k = 3$, estamos identificando los outliers "extremos". Si bien, es más común utilizar un valor de 1.5 veces el Rango Intercuartílico (IQR) ya que, asumiendo una distribución normal estándar, obtendríamos los datos contenidos en el rango de -2.7σ a 2.7σ, lo que corresponde al 99.65 % de las muestras. Sin embargo, teniendo en cuenta el tamaño del conjunto de datos y sus particularidades, perderíamos muchas muestras y representatividad, dado que, como hemos observado, la mayoría de los outliers corresponden a la clase TEA.

Tabla 4: Número de valores atípicos por variable para diferentes valores de k, de 3 a 1,5 veces IQR.

Número de outliers por variable			
Train Set, (k=3)		*Train Set, (k=1.5)*	
r	8	r	205
theta	2.405	theta	18.702
GazeEventDuration	5.069	GazeEventDuration	16.509
FixationPointX	0	FixationPointX	0
FixationPointY	0	FixationPointY	0
StrcitAverageGazePointX	2	StrcitAverageGazePointX	100
StrcitAverageGazePointY	7.631	StrcitAverageGazePointY	27.882

Por otro lado, se lleva a cabo el escalado de características mediante los dos principales métodos existentes, normalización y estandarización. Aspecto importante, considerando la existencia de variables en diferentes escalas y magnitudes, lo cual puede afectar al rendimiento de algunos algoritmos de clasificación (Geller, 2019). Se sabe que existen ciertos algoritmos que se ven muy afectados por la escala de los datos (Raschka, 2014), mientras que otros son inmunes a ella. Más específicamente, se pueden diferenciar en dos clases, típicamente teniendo, por un lado, aquellos basados en árboles de decisión o derivados de ellos, a los que podemos llamar robustos a la escala, y por otro lado, aquellos basados en distancias (KNN o SVM) y los que emplean algoritmos de descenso de gradiente (redes neuronales o regresión logística) como optimizadores, que, debido a su propio cálculo interno, se ven afectados por este factor.

- Normalización - Técnica también conocida como escalado Min-Max en la cual los datos se reescalan para abarcar un rango entre los valores 0 y 1, de manera que el dataset se modifica según sus valores extremos (1). Este método solo cambia la escala de los datos y, por lo tanto, no afecta a la forma de la distribución.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

- Estandarización - Este método se implementa restando la media poblacional de cada muestra y dividiéndola por la desviación estándar (2). Esto resultará en una distribución con media = 0 y varianza = 1. Es importante notar que, al aplicar esta operación, la forma de la distribución de los datos se ve afectada. Además, a diferencia de la normalización, no se establecen límites en el rango de los datos.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i) \tag{3}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} \tag{4}$$

Teniendo en cuenta el aumento del número de variables como resultado del proceso de ingeniería de características, se consideró conveniente emplear técnicas de selección de características con el fin de emplear sólo aquellas que optimicen la clasificación de la clase objetivo, (TEA), contribuyendo además, a reducir el tiempo de entrenamiento requerido. Para ello, se han empleado diferentes métodos, entre *ANOVA F-value*, *Mutual Information*, o *RFE* (*Recursive Feature Elimination*). En la práctica, se obtuvieron resultados similares con todos los métodos, siendo las tres mejores características StrictAverageGazePointY, DurACC y group, aunque es importante señalar que finalmente, como se detalla en la siguiente sección, esta combinación de variables no fue empleada ya que produjo un deficiente resultado en el conjunto de prueba debido a un alto sobreajuste a los datos de entrenamiento y una pobre generalización del algoritmo.

2.4. Modelos de aprendizaje automático y metodología

Dado que a priori no es posible conocer qué combinación de variables dará los mejores resultados y los métodos de selección de características tampoco han resultado del todo útiles, hemos realizado una evaluación de múltiples variaciones de características y escenas, evaluadas con los algoritmos de aprendizaje automático más comunes: Regresión Lineal (LR), Análisis Discriminante Lineal (LDA), k-Vecinos más Cercanos (KNN), Árboles de Clasificación y Regresión (CART), Naive Bayes (NB), *Random Forest* (RF), Máquina de Vectores de Soporte (SVM) y Extreme Gradient Boosting (XGB).

Es importante señalar que en este estudio se ha seguido un enfoque de análisis "guiado por los datos". Es decir, aunque se dispone de datos provenientes de 22 escenas (y su correspondiente calibración), se desconoce cual es la tipología o la imagen real que se encuentra detrás de cada una de ellas. Este enfoque fue elegido para evitar suposiciones previas que pudieran introducir sesgos en la generación del modelo. En lugar de seleccionar escenas basándonos en suposiciones sobre su relevancia potencial, como por ejemplo, asumir que imágenes con interacciones sociales serían más informativas, optamos por un análisis imparcial. Después de realizar numerosas pruebas entrenando sobre varias escenas simultáneamente y no obtener resultados concluyentes, decidimos enfocarnos en llevar a cabo el entrenamiento empleando únicamente aquellas escenas individuales que, tras un análisis exhaustivo, mostraron ser más eficaces para la diferenciación buscada.

En esta primera fase del estudio, hemos optado por un enfoque basado en modelos clásicos de aprendizaje automático debido a su relativa simplicidad y mayor velocidad de entrenamiento. En total, se han probado 17 combinaciones de variables en cada una de las 23 escenas con ocho algoritmos diferentes. Esto llevó a la ejecución de más de 3,000 modelos de clasificación, de los cuales algunas de las mejores combinaciones se muestran en la Tabla 5.

Se puede observar que, de entre los 3000 modelos evaluados, una de las combinaciones más efectivas incluye las variables FixationPointX, FixationPointY y GazeEventType en la escena 6, utilizando el algoritmo eXtreme Gradient Boosting, XGB. Esta configuración logra una precisión de clasificación del 76,56 % en el conjunto de pruebas (90,51 % en la validación). Actualmente, XGBoost es uno de los algoritmos que brinda mejores resultados para datos estructurados de pequeño y mediano tamaño. Por tanto, se ha decidido seguir desarrollando el modelo con el algoritmo XGB, incluyendo la optimización de sus hiperparámetros.

3. Resultados y discusión

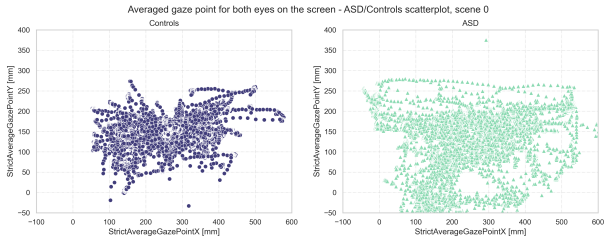
En la presente investigación, además del modelo de Aprendizaje Automático para la clasificación de individuos, deseamos llevar a cabo un análisis más detallado de la naturaleza del comportamiento visual de las personas, en este caso, jóvenes, con TEA. Se han examinado aspectos como la dispersión de la mirada, la preferencia por ciertas regiones específicas o las trayectorias que siguen durante el escaneo visual de las escenas.

Toda esta información sirve como un valioso respaldo durante la definición del problema de Aprendizaje Automático para obtener una mejor comprensión del conjunto de datos, identificar relaciones y patrones en los mismos y, claramente, ayudar a encontrar las mejores estrategias sobre cómo abordarlo. Además, permite realizar una comparación de la información obtenida con el conocimiento actualmente aceptado sobre el TEA. Por lo tanto, a continuación, se discutirán algunos de estos resultados gráficos.

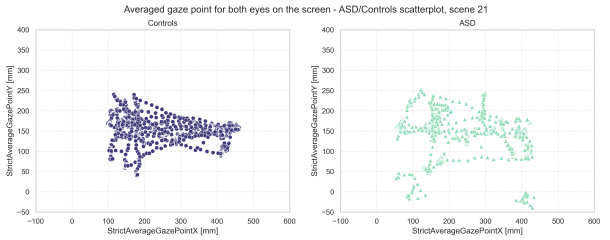
3.1. Gráficos

3.1.1. Mapas de calor y gráficos de dispersión

En primer lugar, podemos obtener una idea del tipo de conjunto de datos con el que estamos tratando al visualizar algunas de sus variables características. En este caso, se presenta un diagrama de dispersión utilizando las variables *StrictAverageGazePointX-Y*, que muestra la ubicación de la mirada de los individuos, separados según la condición de TEA/control, para cada una de las escenas. Con estos gráficos, es posible visualizar de manera rápida y sencilla las diferencias que existen en la forma de mirar de los individuos dependiendo de cada escena (Figura 4). Cabe destacar que la diferencia no es la misma en todas las escenas estudiadas; algunas muestran un comportamiento visual aparentemente similar entre ambos grupos, mientras que en otras se observa una diferencia notable. También es importante señalar que no todas las imágenes son de la misma tipología; algunas muestran elementos sociales como rostros o personas, mientras que otras pueden ser objetos simples o paisajes.



(a)



(b)

Figura 4: Representación gráfica de las coordenadas de la ubicación de la mirada sobre la pantalla de los individuos de control y los casos de TEA para una escena determinada. a) Escena de calibración. b) Escena 21.

Lo que se puede observar en varias escenas diferentes es una dispersión apreciablemente mayor en la mirada de los individuos con TEA, especialmente notable en la escena de calibración 0 (Figura 4(a)), considerando que durante esta etapa, el participante debería enfocar su mirada solo en un punto en el centro de la pantalla. Esto es coherente con múltiples investigaciones que se han llevado a cabo, como en (Schmitt et al., 2014). Según Schmitt et al., se evidencian movimientos oculares anormales durante las *sacadas*, que mostraron "precisión

Tabla 5: Muestra de las mejores combinaciones de variables, modelo y escena. (Precisión sobre el conjunto de test).

Variables	Best Scene	Best Model	Val. Acc.
StrictAverageGazePointX, StrictAverageGazePointY	10	NB	75,04
StrictAverageGazePointX, StrictAverageGazePointY, DurACC	6/18	LDA/LR	72,89
StrictAverageGazePointX, StrictAverageGazePointY, rejilla	6	LR	73,16
FixationPointX, FixationPointY, GazeEventType	6	XGB	76,56

reducida, alta variabilidad entre ensayos, velocidad máxima reducida y duración prolongada" (Schmitt et al., 2014).

La siguiente imagen (Figura 5) muestra el histograma del número de fijaciones ocurridas en cada área (como se define en la Figura 3), para cada escena y clase. Además, podemos observar cómo en algunas escenas existen diferencias notables entre los individuos con TEA y los controles. Por ejemplo, en la Figura 5, es evidente cómo parece que el área principal de interés de los sujetos con TEA se desplaza hacia la derecha.

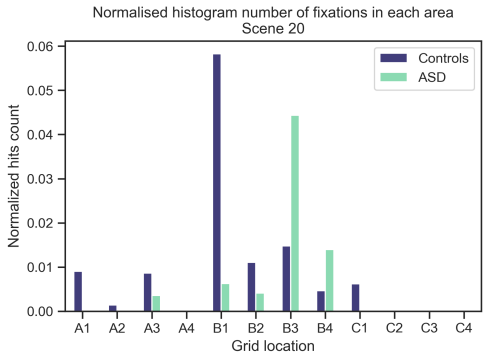


Figura 5: Histograma normalizado representando el número de fijaciones registradas en cada área de la rejilla según clase y escena. Escena 20

Lo discutido anteriormente se puede apreciar de otra manera, mediante un mapa de calor. En este, se muestra la ubicación de las fijaciones de los individuos en las diferentes escenas, resaltando las áreas en las que existe una mayor insistencia. La Figura 6 muestra la misma escena que se puede ver en la Figura anterior, y también se puede observar el efecto detectado. Para esta escena en particular, los individuos con TEA dirigen su atención a una zona diferente de la imagen.

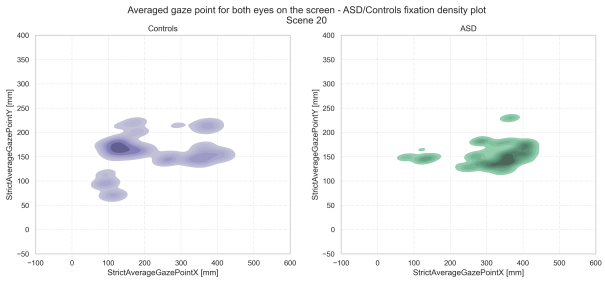
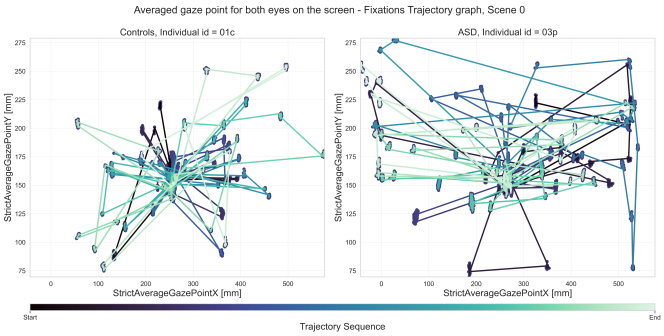


Figura 6: Diagrama de densidad de fijación para la escena 4 del punto de atención sobre la pantalla, promediado para ambos ojos. Izquierda: controles, derecha: TEA.

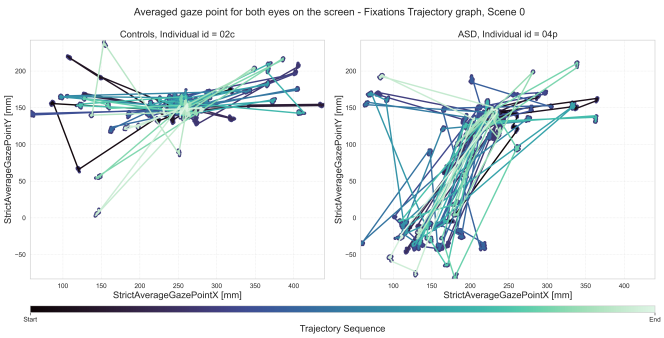
Como ya se ha comentado, en esta primera fase de la investigación y con el fin de evitar hacer ciertas suposiciones sobre los datos, no conocemos las imágenes que se ocultan detrás de cada escena, pero es evidente que algunas de ellas deben tener una tipología que hace detectable un comportamiento diferente entre ambas clases. Se sabe, por ejemplo, que las personas con autismo suelen preferir visualizar objetos inanimados en lugar de otras personas interactuando. Según investigaciones recientes, durante la presentación de escenas cotidianas de personas conversando, los niños con desarrollo típico tienden a dirigir su atención a la misma área el 80 por ciento del tiempo, centrándose generalmente en rasgos clave de un rostro, como el área de los ojos, la nariz y la boca, mientras que aquellos diagnosticados con TEA miran en otra dirección, especialmente a partes no esenciales de la escena (Hughes and Spectrum, 2008).

3.1.2. *Gráficos de trayectoria*

Otra forma de visualización explorada consiste en la representación de las trayectorias seguidas por la mirada de los participantes durante el escaneo visual de las imágenes. En la Figura 7, se han representado las variables StrictAverageGazePointX/Y, en una única escena y un par específico de individuos.



(a)



(b)

Figura 7: Gráfico de trayectoria. Evolución de las variables *StrictAverageGazePointX* e Y en función del tiempo. La escala de color indica la ruta temporal de la mirada, comenzando desde el valor más oscuro hasta el más claro. En ambas subfiguras se muestra un individuo control en la parte izquierda de la misma, y un TEA en la parte derecha. Se puede observar una clara diferencia entre los individuos, con las personas diagnosticadas con TEA mostrando movimientos más erráticos y dispersos.

Nuevamente, lo ya descrito se puede ver aún con más claridad que en gráficos anteriores. La Figura 7 muestra el comportamiento visual durante la escena 0 (calibración) de distintos pares de individuos. En ambos casos, se muestra un individuo control en la parte izquierda de la imagen, frente a un sujeto TEA en la parte derecha de la misma. Se puede apreciar claramente la mayor dispersión en las trayectorias de la mirada de los pacientes con TEA, mostrando una apariencia errática, comportamiento especialmente notorio en la primera imagen (Figura 7(a)). Mientras que en la segunda, (Figura 7(b)) el paciente parece alternar entre dos puntos diferentes de la misma. No debe olvidarse que durante la etapa de calibración, la pantalla se mantiene oscura, sin ningún otro estímulo que una cruz situada en el centro de esta, y a la que los participantes fijar su mirada. Estas marcadas diferencias en el comportamiento visual durante un momento de enfoque fijo pueden resultar cruciales para una mejora en la comprensión del TEA.

Obtener estos gráficos para cada uno de los individuos y escenas abre otra posible vía a explorar, como por ejemplo la aplicación de la teoría de grafos para analizar la conectividad y patrones de mirada en detalle. Este enfoque está en estudio y se deja como trabajo futuro, ya que podría proporcionar información valiosa sobre cómo se relacionan los elementos visuales y cómo evoluciona la atención visual en distintas situaciones.

Todos los gráficos presentados anteriormente muestran que, al menos desde el punto de vista del comportamiento visual, existen diferencias detectables entre los dos grupos de individuos, incluso a simple vista, lo que refuerza la idea de que es posible generar un método de aprendizaje automático para la clasificación de esta patología empleando este tipo de datos.

3.2. Resultados detallados

Como ya se introdujo previamente, después de realizar múltiples pruebas que abarcaron diferentes combinaciones de escenas y variables, se ha alcanzado un resultado óptimo utilizando el algoritmo XGBoost, empleando únicamente tres variables y una única escena, con una precisión en el conjunto de pruebas del 76.56 %.

Respecto a los resultados, siguiendo la terminología comúnmente aceptada, nos referimos al *recall* de la clase positiva como sensibilidad y al de la clase negativa como especificidad y, teniendo en cuenta la definición de la matriz de confusión, en la que se representan las muestras de clases reales y predichas (Figura 8), podemos definir las métricas utilizadas de la siguiente manera:

		Predicho	
		Positivo	Negativo
Real	Positivo	TP	FN
	Negativo	FP	TN

Figura 8: Matriz de confusión. TP y TN, verdaderos positivos y negativos, respectivamente. Del mismo modo, FP y FN hacen referencia a falsos positivos y falsos negativos.

- Precisión. Indica el número de verdaderos positivos entre todos los predichos como tales. Basándonos en los valores de la matriz de confusión, se define de la siguiente manera:

$$\text{Precisión} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

- *Recall*. Refleja la capacidad del algoritmo para detectar todas las instancias positivas (o negativas).
  - Sensibilidad. Representa la tasa de verdaderos positivos. También conocida como tasa de detección o recall de la clase verdadera, indica la probabilidad de ser clasificado como positivo entre todos los casos verdaderamente positivos. Definida como:

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

  - Especificidad. Representa la tasa de verdaderos negativos. Complementaria a la anterior, también se denomina recall de la clase negativa. Indica la capacidad de clasificar a los individuos verdaderamente negativos como negativos. Definida como:

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

- Puntuación F1. Se calcula como la media armónica ponderada entre la precisión y la sensibilidad, siendo el mejor valor posible 1 y el peor, 0.

$$F_1 = 2\,\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

En la siguiente tabla, se muestra un resumen de las métricas de clasificación seleccionadas (Tabla 6), tanto para el conjunto de validación como para el conjunto de pruebas.

Tabla 6: Métricas del informe de clasificación (conjunto de validación y prueba).

Métricas de clasificación en el conjunto de validación

	Precisión	Recall	f1-Score
Control	0.96	0.83	0.89
TEA	0.83	0.96	0.89

Métricas de clasificación en el conjunto de test

	Precisión	Recall	f1-Score
Control	0.80	0.74	0.77
TEA	0.77	0.82	0.79

Como es esperable, se puede observar que los resultados son algo peores en el conjunto de test (individuos completamente nuevos y algo de sobreajuste en el modelo), aunque se puede apreciar cómo el algoritmo se desempeña bastante bien detectando la clase TEA. Se puede notar que la sensibilidad de esta clase (como se definió previamente) alcanza un porcentaje del 82 %, es decir, de todas las muestras pertenecientes a individuos con autismo, se detectan correctamente el 82 % de ellas. Este buen resultado se logra a expensas de la precisión para la clase de TEA, de modo que, de todas las muestras que el algoritmo clasifica como positivas, solo el 77 % son realmente positivas.

Del mismo modo, para la clase negativa, ocurre de manera simétrica. En este caso, se logra una mayor precisión a costa de sacrificar la especificidad. Sin embargo, como quedará más claro en la matriz de confusión (Figura 9), el mayor objetivo

de una herramienta como esta sería la detección de la mayor cantidad de individuos con autismo, aunque para ello aumente ligeramente el error de tipo I.

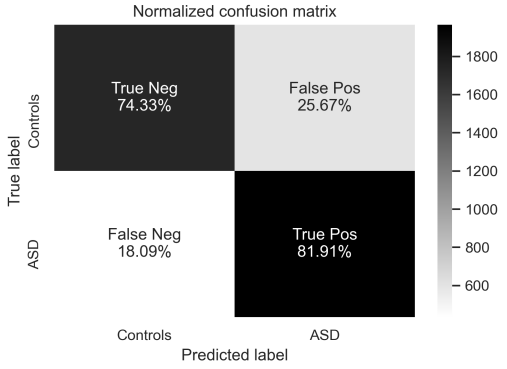


Figura 9: Matriz de confusión normalizada por filas (Etiqueta real).

La matriz de confusión anterior (Figura 9) está calculada sobre las muestras del conjunto de pruebas y normalizada por filas, de acuerdo con el valor real. Como se puede observar, el algoritmo se desempeña adecuadamente en el diagnóstico de muestras positivas (TEA), como lo muestra el porcentaje de verdaderos positivos (81.91 %), mientras que la especificidad disminuye al 74.33 %. En cuanto a las muestras mal clasificadas, tenemos que la cantidad de errores de tipo II es menor que los de tipo I, es decir, de las muestras realmente positivas (TEA), son diagnosticadas incorrectamente un menor porcentaje que de las muestras negativas (controles). Este es un aspecto generalmente preferible en el ámbito médico, en el cual el costo de un falso negativo es generalmente mayor que el de un falso positivo.

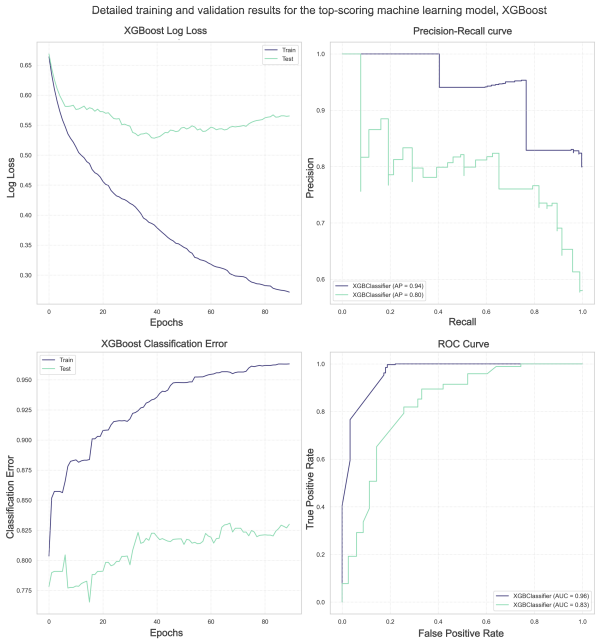


Figura 10: Múltiples gráficos de entrenamiento. Columna izquierda: gráficos de entrenamiento del algoritmo (*logarithmic loss* y error de clasificación), columna derecha: curva ROC y relación precisión-recall.

En la Figura 10, se presentan unos gráficos que representan el proceso de entrenamiento y evaluación del algoritmo. En la columna izquierda, se muestran gráficos relacionados con el entrenamiento del modelo, como la Función de Costo y el Error de Clasificación en función de la iteración. Se utilizó la técnica de *Early Stopping* para evitar el sobreajuste, pero aún persiste en cierta cantidad, como se puede ver en las curvas.

En la columna derecha, se encuentran las curvas de Precisión vs. Sensibilidad y la curva ROC. La primera muestra un hecho conocido y ya comentado, a medida que la sensibilidad aumenta, la precisión del modelo disminuye debido a la aparición de más falsos positivos. Por otro lado, la curva ROC muestra un adecuado equilibrio entre verdaderos positivos y falsos positivos.

Es relevante destacar que, debido al diseño del sistema se han clasificado las muestras individuales del rastreador ocular, siendo el objetivo final la clasificación de la patología a nivel de individuo. Por eso, implementamos un algoritmo de votación por mayoría que asigna un sujeto a una clase (0 o 1, control o TEA) si el número de sus muestras clasificadas supera un cierto porcentaje, k del total de muestras del individuo. Más formalmente definido:

$$\sum_{i=0}^{n} \hat{y}^0_{i,j} > k n_j \Rightarrow \hat{\mathrm{Y}}_j = \mathbf{0} \tag{9}$$

$$\sum_{i=0}^{n} \hat{y}^1_{i,j} > k n_j \Rightarrow \hat{\mathrm{Y}}_j = \mathbf{1} \tag{10}$$

dónde:

$\hat{y}^0_{i,j}$ = Cada una de las i-ésimas muestras de un individuo j clasificadas por el algoritmo. El superíndice indica la clase a la cual ha sido asignada, (0 o 1).
k = Constante. Factor que define el umbral de clasificación final como porcentaje del total de muestras. Valor seleccionado 0.7
n_j = Número total de muestras existentes para un individuo determinado j.
$\hat{\mathrm{Y}}_j$ = Clase final a la que se ha asignado el individuo j.

Como ejemplo detallado, con los datos del individuo con identificador "08p" (TEA): $n_j = 1199$, $\hat{y}^0_j = 106$, $\hat{y}^1_j = 1093$, y fijando un umbral, k, del 70 %. De acuerdo con (9) y (10), dado que más del 70 % de las muestras se han clasificado como clase 1 (en este caso, el porcentaje es aún mayor, 91.16 %), el individuo pertenecerá a la clase 1 (TEA). Es una idea simple, pero que funciona perfectamente con todos los sujetos del conjunto de test (Tabla 7), clasificándolos correctamente. Basándonos en este resultado, se puede decir que obtenemos una precisión en la clasificación de los individuos del 100 %, cada uno con un nivel de certeza diferente (pero siempre por encima del 70 %).

Tabla 7: Resultados de clasificación detallado de los individuos del conjunto de pruebas.

Individuo id	Muestras Positivas	Muestras Negativas	Muestras Totales	Clase Verdadera	Clasificación
07p	872	328	1200	1	TEA, 72.7 % muestras Clase 1
08p	1093	106	1199	1	TEA, 91.2 % muestras Clase 1
05c	321	859	1180	0	Control, 72.8 % muestras Clase 0
09c	279	878	1157	0	Control, 75.9 % muestras Clase 0

Como se mencionó en la Introducción, estudios previos como (Wang et al., 2015), (Liu et al., 2015) y (Jiang and Zhao, 2017) han empleado con éxito dispositivos de seguimiento ocular y aprendizaje automático. Sin embargo, si bien, se obtienen buenos resultados, se basan en modelos generalmente más complejos y multitud de imágenes. En contraste, nuestro enfoque se ha centrado en generar un modelo relativamente simple, aunque robusto, empleando datos fácilmente medibles.

Por otro lado, este trabajo se diferencia de la mayoría de los estudiados en el metanálisis de (Wei et al., 2023), en su rigurosa aplicación de conjuntos separados de entrenamiento, validación y test, lo cual es esencial para evitar el sobreajuste y garantizar la generalización de los modelos de machine learning. Esta ha sido una de las principales debilidades detectadas que, aunque dependiendo de las circunstancias no llega a suponer un fallo metodológico, presumiblemente esté provocando resultados excesivamente optimistas. Hecho que queda patente al observar la precisión media de los estudios reportados en (Wei et al., 2023) (84 % sin conjunto de test, 73 % añadiendo conjunto de test independiente).

Al implementar esta metodología, nuestro estudio no solo refuerza su fiabilidad, sino que también se alinea con las mejores prácticas actuales en la investigación de machine learning, abordando una de las principales limitaciones observadas en otros estudios sobre el diagnóstico del TEA.

Asimismo, es importante destacar los buenos resultados obtenidos con una única escena elevadamente discriminativa y un algoritmo potente como XGBoost. Se ha desarrollado un modelo de clasificación simple de mantener y rápido de entrenar y reentrenar cuando se tengan más individuos disponibles. El uso de este algoritmo, combinado con un sistema de votación por mayoría, ofrece una mejora en términos de precisión y robustez y permite una clasificación individualizada más detallada.

4. Conclusiones

En la presente investigación nos hemos enfocado en demostrar la aplicabilidad de un procedimiento de detección basado en aprendizaje automático sobre el Trastorno del Espectro Autista mediante datos del comportamiento ocular. En primer lugar, hemos destacado la diferencia existente en el procesamiento visual de estos sujetos, un factor conocido y ampliamente utilizado, decidiendo aprovechar esta característica para la clasificación.

Se han presentado los pasos más comúnmente seguidos en un problema de ciencia de datos, abarcando desde la adquisición, limpieza, preprocesado e inspección de los datos, hasta el entrenamiento y validación del modelo predictivo. Queda patente la gran capacidad discriminativa de un modelo no excesivamente complejo como XGBoost, que, suministrado con unas pocas variables relevantes y una única escena muy efectiva, es capaz de ofrecer grandes resultados como ha sido demostrado. A pesar de esto, todavía son necesarias más pruebas con nuevos individuos para corroborar con mayor precisión el nivel de confiabilidad del modelo antes de poder evaluar la viabilidad de este en un entorno real. Del mismo modo, queda como trabajo futuro, explorar otras opciones como la fusión de diversas escenas que arrojen buenos resultados y la aplicación de otras técnicas como teoría de grafos sobre los datos de trayectorias.

Referencias

Abrahams, B. S., Geschwind, D. H., may 2008. Advances in autism genetics: on the threshold of a new neurobiology. Nature Reviews Genetics 9 (5), 341–355.
DOI: 10.1038/nrg2346

American Psychiatric Association, May 2013. Diagnostic and Statistical Manual of Mental Disorders (5th ed.), 5th Edition. American Psychiatric Association.
DOI: 10.1176/appi.books.9780890425596

Bishop, C. M., 2019. Pattern recognition and machine learning. Information Science and Statistics. Springer Science+Business Media, LLC, New York, NY.

Cassisi, C., Montalto, P., Aliotta, M., Cannata, A., Pulvirenti, A., 2012. Similarity measures and dimensionality reduction techniques for time series data mining. In: Karahoca, A. (Ed.), Advances in Data Mining Knowledge Discovery and Applications. IntechOpen, Rijeka, Ch. 3.
DOI: 10.5772/49941

Centers for Disease Control and Prevention, CDC, Sep. 2020. Data & statistics on autism spectrum disorder. https://www.cdc.gov/ncbddd/autism/data.html.

Chawarska, K., Macari, S., Shic, F., Aug. 2013. Decreased Spontaneous Attention to Social Scenes in 6-Month-Old Infants Later Diagnosed with Autism Spectrum Disorders. Biological Psychiatry 74 (3), 195–203.
DOI: 10.1016/j.biopsych.2012.11.022

Chen, T., Guestrin, C., aug 2016. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
DOI: 10.48550/ARXIV.1603.02754

Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I. D., Aug. 2010. ScanMatch: A novel method for comparing fixation sequences. Behavior Research Methods 42 (3), 692–700.
DOI: 10.3758/brm.42.3.692

D. Nevison Cynthia, sep 2014. A comparison of temporal trends in United States autism prevalence to trends in suspected environmental factors. Environmental Health 13 (1).
DOI: 10.1186/1476-069x-13-73

Day, R., nov 2010. Examining the validity of the needleman-wunsch algorithm in identifying decision strategy with eye-movement data. Decision Support Systems 49 (4), 396–403.
URL: https://www.sciencedirect.com/science/article/pii/S0167923610000904
DOI: 10.1016/j.dss.2010.05.001

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (5), 1189 – 1232.
URL: https://doi.org/10.1214/aos/1013203451
DOI: 10.1214/aos/1013203451

Geller, S., Apr. 2019. Normalization vs Standardization: Quantitative analysis. https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf, accessed: 23-04-2021.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.

Goshtasby, A. A., 2012. Similarity and Dissimilarity Measures. Springer London, London, Ch. Chapter 2, pp. 7–66.
DOI: 10.1007/978-1-4471-2458-0_2

Hannah Furfaro, Spectrum, May 2019. Gaze patterns in toddlers may predict autism. https://www.spectrumnews.org/news/gaze-patterns-toddlers-may-predict-autism/, accessed: 23-03-2020.

Hughes, V., Spectrum, Sep. 2008. Eyes provide insight into autism's origins. https://www.spectrumnews.org/news/eyes-provide-insight-into-autisms-origins/, accesed: 2020-04-23.

Jiang, M., Zhao, Q., Oct. 2017. Learning visual attention to identify people with autism spectrum disorder. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 3287–3296.
DOI: 10.1109/iccv.2017.354

Jones, W., Klin, A., Nov. 2013. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. Nature 504 (7480), 427–431.
DOI: 10.1038/nature12715

Kaltenbach, H., 2011. A Concise Guide to Statistics. SpringerBriefs in Statistics. Springer Berlin Heidelberg.
URL: https://books.google.es/books?id=2L8xNcbRvYgC

LeCun, Y., Bengio, Y., Hinton, G., May 2015. Deep learning. Nature 521 (7553), 436–444.
DOI: 10.1038/nature14539

Libbey, J., Sweeten, T., McMahon, W., Fujinami, R., feb 2005. Autistic disorder and viral infections. Journal of NeuroVirology 11 (1), 1–10.
DOI: 10.1080/13550280590900553

Liu, W., Yu, X., Raj, B., Yi, L., Zou, X., Li, M., Sep. 2015. Efficient autism spectrum disorder prediction with eye movement: A machine learning framework. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 649–655.
DOI: 10.1109/acii.2015.7344638

López, V., Fernández, A., García, S., Palade, V., Herrera, F., nov 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113–141.
DOI: 10.1016/j.ins.2013.07.007

Maenner, M. J., Shaw, K. A., Jon Baio, e., mar 2020. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. MMWR. CDC Surveillance Summaries 69 (4), 1–12.
DOI: 10.15585/mmwr.ss6904a1

Mendelsohn, N. J., Schaefer, G. B., mar 2008. Genetic evaluation of autism. Seminars in Pediatric Neurology 15 (1), 27–31.
DOI: 10.1016/j.spen.2008.01.005

Needleman, S. B., Wunsch, C. D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48 (3), 443–453.
URL: https://www.sciencedirect.com/science/article/pii/0022283670900574
DOI: https://doi.org/10.1016/0022-2836(70)90057-4

Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., Pham, B. T., feb 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. Mathematical Problems in Engineering 2021, 1–15.
DOI: 10.1155/2021/4832864

Preeti, K., Srinath, S., Shekhar, P. S., Satish, C. G., Kommu, J. V. S., Feb. 2017. Lost time: Need for more awareness in early intervention of autism spectrum disorder. Asian Journal of Psychiatry 25, 13–15.
DOI: 10.1016/j.ajp.2016.07.021

Raschka, S., Jul. 2014. About feature scaling and normalization and the effect of standardization for machine learning algorithms. https://sebastianraschka.com/Articles/2014_about_feature_scaling.html, accessed: 23-04-2021.

Rosen, N. E., Lord, C., Volkmar, F. R., Feb. 2021. The diagnosis of autism: From kanner to dsm-iii to dsm-5 and beyond. Journal of Autism and Developmental Disorders 51 (12), 4253–4270.
DOI: 10.1007/s10803-021-04904-1

Schmidhuber, J., Jan. 2015. Deep learning in neural networks: An overview. Neural Networks 61, 85–117.
DOI: 10.1016/j.neunet.2014.09.003

Schmitt, L. M., Cook, E. H., Sweeney, J. A., Mosconi, M. W., 2014. Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunctions in cerebellum and brainstem. Molecular Autism 5 (1), 47.
DOI: 10.1186/2040-2392-5-47

Seger, C., 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. Ph.D. thesis, KTH, School of Electrical Engineering and Computer Science (EECS).

Tanaka, J. W., Sung, A., Oct. 2013. The "eye avoidance" hypothesis of autism face processing. Journal of Autism and Developmental Disorders 46 (5), 1538–1552.
DOI: 10.1007/s10803-013-1976-7

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A. J., nov 2019. Machine learning algorithm validation with a limited sample size. PLOS ONE 14 (11), e0224365.
DOI: 10.1371/journal.pone.0224365

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., Zhao, Q., Nov. 2015. Atypical Visual Saliency in Autism Spectrum Disorder Quantified through Model-Based Eye Tracking. Neuron 88 (3), 604–616.
DOI: 10.1016/j.neuron.2015.09.042

Wei, Q., Cao, H., Shi, Y., Xu, X., Li, T., Jan. 2023. Machine learning based on eye-tracking data to identify autism spectrum disorder: A systematic review and meta-analysis. Journal of Biomedical Informatics 137, 104254.
DOI: 10.1016/j.jbi.2022.104254

Xu, G., Jing, J., Bowers, K., Liu, B., Bao, W., sep 2013. Maternal diabetes and the risk of autism spectrum disorders in the offspring: A systematic review and meta-analysis. Journal of Autism and Developmental Disorders 44 (4), 766–775.
DOI: 10.1007/s10803-013-1928-2

Zwaigenbaum, L., Bryson, S., Rogers, T., Roberts, W., Brian, J., Szatmari, P., Jun. 2004. Behavioral manifestations of autism in the first year of life. International Journal of Developmental Neuroscience 23 (2-3), 143–152.
DOI: 10.1016/j.ijdevneu.2004.05.001