Ángel Parra Cuadrado

ANÁLISIS DE DATOS: INVESTIGACIÓN DE LA TRANSMISIBILIDAD DE LOS VIRUS RESPIRATORIOS EN LAS AULAS

TRABAJO DE FIN DE GRADO

dirigido por Edgar Batista de Frutos y Agustí Solanas Gómez

Grado de Ingeniería Informática

UNIVERSITAT ROVIRA I VIRGILI

Tarragona

2024

Resum

En el marc del projecte de recerca ‘Actua’, va sorgir la necessitat d'entendre com es propagava el virus de la COVID-19 per prevenir i mitigar el seu impacte. Atès que els nens passen la major part del seu temps a les aules de les escoles, hem decidit enfocar-nos en avaluar aquests espais, juntament amb les condicions ambientals i la densitat de l'aula, per poder millorar i minimitzar els riscos de contagi. Inicialment, es va realitzar una anàlisi de correlacions, a més de validar la rellevància estadística de les dades i determinar si són adequades per ser incloses en l'estudi. A partir de la base inicial obtinguda, es van aplicar tècniques d’anàlisi multivariant, començant amb una anàlisi de regressió lineal per obtenir una primera idea del que podríem trobar. Tot i que els resultats no van ser bons, el model suggeria la possible existència de relacions no lineals. Per estudiar aquestes relacions, es van emprar diversos models de Machine Learning i es van avaluar comparant les mètriques basades en la precisió de les prediccions. Finalment, s'han obtingut uns resultats preliminars interessants que serviran com a base per a futures investigacions, permetent una major comprensió dels paràmetres contextuals que influeixen en la transmissió de virus respiratoris.

Resumen

En el marco del proyecto de investigación ‘Actua’, surgió la necesidad de entender cómo se propagaba el virus de la COVID-19 para prevenir y mitigar su impacto. Dado que los niños pasan la mayor parte de su tiempo en las aulas de las escuelas, decidimos enfocarnos en evaluar estos espacios, junto con las condiciones ambientales y la densidad del aula, para poder mejorar y minimizar los riesgos de contagio. Inicialmente, se realizó un análisis de correlaciones, además de validar la relevancia estadística de los datos y determinar si eran adecuados para ser incluidos en el estudio. A partir de la base inicial obtenida, se aplicaron técnicas de análisis multivariante, comenzando con un análisis de regresión lineal para obtener una primera idea de lo que podríamos encontrar. Aunque los resultados no fueron buenos, el modelo sugería la posible existencia de relaciones no lineales. Para estudiar estas relaciones, se emplearon varios modelos de Machine Learning y se evaluaron comparando las métricas basadas en la precisión de las predicciones. Finalmente, se han obtenido unos resultados preliminares interesantes que servirán como base para futuras investigaciones, permitiendo una mayor comprensión de los parámetros contextuales que influyen en la transmisión de virus respiratorios.

Abstract

Within the framework of the ‘Actua’ research project, the need arose to understand how the virus spread to prevent and mitigate its impact on public health. Since children spend most of their time in school classrooms, we decided to focus on evaluating these spaces, along with environmental conditions and classroom density, to improve and minimize the risks of contagion. Initially, a correlation analysis was conducted, in addition to validating the statistical relevance of the data and determining whether they were suitable to be included in the study. Based on the initial findings, the results were used in multivariate analysis, beginning with a linear regression analysis to get a first idea of what we might find. Although the results were not promising, the model suggested the possibility of nonlinear relationships. To study these relationships, several Machine Learning models were employed, and they were evaluated by comparing metrics based on prediction accuracy. Finally, some interesting preliminary results have been obtained, which will serve as a solid foundation for future research, allowing for a better understanding of the contextual parameters that influence the transmission of respiratory viruses.

Índice

Índice de tablas

Índice de figuras

1 Introducción

La pandemia de COVID-19 destacó la necesidad de estudiar la transmisión de virus respiratorios y cómo podemos evitarla. Esta última pandemia no fue más que una de las muchas otras erupciones virales de las últimas décadas, como el SARS en 2003, la gripe A en 2010 y el MERS en 2012. Los virus respiratorios se propagan a través de aerosoles, pequeñas partículas que se encuentran suspendidas en el aire durante largos períodos y que se forman cuando hablamos, estornudamos o tosemos. La probabilidad de contagio mediante aerosoles, es decir inhalando dichas partículas emitidas por personas infectadas, es mayor en espacios cerrados, de ahí las medidas de ventilación que se aplicaron durante la pandemia. La aparición de estas condiciones junto con la peligrosidad de dichas enfermedades resalta la necesidad de estudiar los principales factores de riesgo de contraer dichas enfermedades, especialmente entre los grupos de población de riesgo, como los ancianos y los niños.

Para dar solución a ello, en 2021 se inició el proyecto ACTUA por parte de un grupo de investigadores de la Universitat Rovira i Virgili. Este proyecto tiene por objeto examinar cómo las condiciones contextuales de las aulas de escuelas de educación infantil y primaria afectan a la propagación de virus respiratorios. El aula de la escuela es un espacio confinado con individuos que pertenecen a diferentes hogares, por lo que hay una gran probabilidad de que aumenten las probabilidades de contagios. Por ello, se instalaron kits de sensores en las aulas para monitorizar variables como la temperatura, la humedad y la concentración de dióxido de carbono (CO_2), y se recopilaron datos epidemiológicos del estado de salud de dichas aulas, incluyendo estudiantes y profesorado. Con datos de dos cursos académicos (2022/23 y 2023/24), mi trabajo ha consistido en analizar las correlaciones entre las condiciones dentro de las aulas y el riesgo de transmisión de virus respiratorios, con el objetivo de interpretar esos resultados y, finalmente, crear un modelo de predicción efectivo.

1.1 Motivaciones

Inicialmente, buscaba un tema de Trabajo de Fin de Grado (TFG) el cual no hubiéramos abarcado lo suficiente durante el plan de estudios del grado y que, además, estuviera ganando protagonismo estos últimos años en el ámbito informático. Formar parte del proyecto ACTUA me ha dado la oportunidad de combinar dos temas muy relevantes actualmente: la ciencia de datos y la salud pública.

El trabajo práctico que se me ha permitido realizar sobre el análisis de datos con unos conjuntos de datos reales y actualizados me ha brindado la experiencia necesaria para comenzar a desarrollar habilidades prácticas que, posteriormente, podré aplicar en el campo del estudio. Además, la sensación de haber realizado alguna tarea o estudio que tenga un impacto directo en la salud pública es muy gratificante. No solo se me ha permitido crecer personal y profesionalmente gracias a este proyecto, sino que también he podido contribuir a buscar soluciones efectivas para localizar y disminuir la propagación de los virus respiratorios en las escuelas. El conjunto de todas estas experiencias ha enriquecido mi educación tanto a nivel académico como personal y me ha dado una perspectiva más realista sobre cómo la ciencia de datos puede ser aplicada para luchar contra los grandes problemas de la salud pública.

También me llena de satisfacción saber que el trabajo realizado contribuirá a proteger uno de los grupos más vulnerables de la población, los niños, en un entorno tan importante como una escuela. Básicamente, esta oportunidad de aprendizaje sobre una rama que inicialmente me generaba interés y la posibilidad de interacción social la que realmente me impulsa a querer integrarme en este proyecto y realizar un trabajo eficiente.

1.2 Objetivos

Mi tarea principal será proporcionar una investigación sobre los datos recopilados en las más de 120 aulas de educación infantil y primaria durante los dos últimos cursos académicos. Queremos obtener unas buenas conclusiones, a través de los estudios realizados, que nos faciliten la toma de decisiones para prevenir el contagio de virus respiratorios. Este proyecto es desafiante para mí porque tengo una base estadística básica y he trabajado relativamente poco con la programación aplicada a la estadística. Mi desafío principal será investigar y aprender de manera autodidacta para poder cumplir el objetivo a medida que trabajo en los diferentes estudios y voy comentando los resultados con el equipo de investigación. Mis objetivos a nivel profesional son los siguientes:

- Análisis de Datos:
  - Preprocesar los datos obtenidos y analizarlos para entender la situación actual de las diferentes aulas y así, poder analizar y proponer diferentes soluciones.
  - Realizar una investigación sobre qué metodologías y métricas me pueden ir bien con los conjuntos de datos a tratar.
  - Aplicar herramientas de visualización de datos para obtener de una manera eficaz un resumen de los análisis realizados y así, poder proporcionar unas conclusiones más completas.

- Machine Learning:
  - Investigar qué herramientas puedo utilizar en este proyecto y decidir si me pueden ser de utilidad con el conjunto de datos como, por ejemplo, modelos de predicción.
  - En caso de que el punto anterior se haya obtenido alguna herramienta, mi objetivo será aprender a usarla para explotarla lo máximo posible y sacar los mejores resultados posibles.

El objetivo personal de este trabajo de final de carrera es aprender, investigar y realizar estudios completos basados en los hallazgos descubiertos, para apoyar las decisiones con respecto a la prevención de la transmisión de los virus respiratorios en el entorno escolar. La experiencia es muy importante para mí ya que nunca he tratado con un proyecto relacionado con la ciencia de datos. Quiero agradecer al gran equipo de investigación por darme la oportunidad de trabajar con un conjunto de datos real y reciente. Eso me ha dado la oportunidad de tratar con un amplio conjunto de herramientas y algoritmos estadísticos, así que estoy agradecido y realmente motivado.

2 Marco Teórico

La pandemia de COVID-19 ha destacado la importancia de la adquisición de un mejor entendimiento sobre los mecanismos de transmisión de los virus respiratorios en una variedad de entornos, incluyendo, además de los hogares, escuelas y oficinas, que son lugares muy frecuentados. La transmisión de virus respiratorios no solo amenaza la salud individual, sino que también lleva a graves consecuencias para la salud pública, la economía y las interacciones sociales.

En un entorno escolar, los niños y los maestros pasan muchas horas en un mismo espacio cerrado, por lo que es importante reconocer y manejar los factores que permiten y facilitan la transmisión de los virus para proteger tanto a los estudiantes como a los maestros y para evitar la interrupción de la educación. Dado el cierre temporal de escuelas en todo el mundo debido a la pandemia de COVID-19 para ayudar a aplanar la curva, surgió la necesidad de trabajar en mecanismos efectivos de control y vigilancia de enfermedades en el entorno escolar. Este estudio pretende, con el uso de técnicas de análisis de datos, mejorar nuestra comprensión de los factores que contribuyen a la propagación de los virus en los entornos educativos. Una vez identificados los factores y sus interacciones, será posible diseñar e implementar intervenciones más efectivas para reducir la propagación de virus y mejorar la salud y el bienestar de estudiantes y profesores.

2.1 Proyecto ACTUA

El proyecto ACTUA es una iniciativa de investigación llevada a cabo por la Universitat Rovira i Virgili, cuyo objetivo principal es monitorizar y analizar la transmisión de virus respiratorios en entornos escolares mediante el uso de tecnologías avanzadas. Este proyecto ha implementado un sistema de monitorización contextual en más de un centenar de aulas de educación infantil y primaria en Cataluña, utilizando kits de sensores para recopilar datos ambientales y epidemiológicos [1].

El proyecto se centra en el desarrollo e instalación de un sistema de monitorización contextual inteligente y autónomo en más de un centenar de aulas de educación infantil y primaria en Cataluña con el fin de permitir el monitoreo activo de las variables ambientales y epidemiológicas y hacer un estudio inmediato de la relación entre las condiciones del aula y la propagación de enfermedades respiratorias. El proyecto propone investigar las condiciones que favorecen la propagación de los virus respiratorios dentro de las aulas y para lograrlo, han definido una serie de objetivos para conseguirlo:

1. Identificar si existe algún tipo de correlación entre variables ambientales y la probabilidad de propagación de virus.
2. Estudiar los papeles que juegan los aerosoles en la propagación de los virus respiratorios.
3. Crear un kit de monitorización (combinando sensores con redes de comunicaciones) que midan las variables ambientales de manera continua.
4. Crear un sistema de vigilancia contextual mediante la instalación de los kits de monitorización.
5. Crear un sistema de predicción y alerta, demostrando que es posible recoger datos que permiten actuar sobre el entorno para minimizar la propagación del virus.
6. Realizar un despliegue masivo en el número máximo de clases posibles para monitorizar y obtener una gran cantidad de datos para analizar.

El sistema tecnológico del Proyecto ACTUA se basa en una combinación de hardware y software diseñados específicamente para la recopilación, transmisión y análisis de datos. A continuación, se describen los componentes clave de la tecnología utilizada en este proyecto:

- Kit de sensores: Conjunto de sensores conectados a una Raspberry Pi ubicados en una caja cubierta de malla de aluminio de tamaño 20x20x10cm, diseñado para ser discreto y poder instalarlo de manera sencilla a la clase. La (Tabla 2**)** mostrará las variables que se reciben en los sensores.

- Redes de comunicación: Los kits de sensores están conectados a la red de la escuela para acceder a Internet, lo que permite la transmisión de datos al servidor central cada ocho horas. Esta conectividad es esencial para la monitorización en tiempo real y la gestión centralizada de los datos.

- Servidor ACTUA: Un servidor central recibe y almacena todos los datos recopilados por los kits de sensores. Este servidor está equipado con una base de datos que organiza y gestiona la información, facilitando su análisis posterior. Todos los datos se almacenan de manera segura y se implementan medidas de privacidad estrictas para garantizar la protección de la información sensible.

- Aplicación web: La aplicación web ha sido diseñada para ser intuitiva y fácil de usar, permitiendo a diferentes tipos de usuarios (administradores, analistas, directores y maestros) acceder a la información que necesitan. Desde visualizar los datos en tiempo casi real con una interfaz que muestra miniaturas de las aulas monitorizadas hasta poder descargar datos en formatos Excel, visualizar esos datos e incluso recibir alertas y notificaciones. Además, los maestros deberán completar unos cuestionarios de salud semanalmente para proporcionar datos sobre incidencias de síntomas respiratorios y ausencias relacionadas con enfermedades.

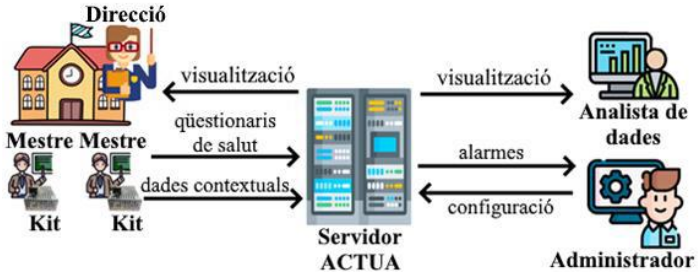


Figura 1. Estructura de la arquitectura del proyecto ACTUA

La (Figura 1), extraída del libro [2] creado por investigadores del proyecto, muestra el funcionamiento y el flujo de datos en la arquitectura tecnológica diseñada en el proyecto.

2.2 Técnicas de análisis multivariante

El análisis multivariante nos ha permitido estudiar y modelar las relaciones entre diferentes variables simultáneamente. Este tipo de análisis nos permitía descubrir patrones, correlaciones y dependencias entre múltiples dimensiones de datos el cual, nuestro objetivo principal era estudiar si las variables están interrelacionadas y cuanto pueden influir entre ellas. Nuestro objetivo principal en este proyecto era estudiar las variables que teníamos actualmente y ver si existe algún tipo de correlación con las incidencias relacionadas con un virus respiratorio.

2.2.1 Regresión Lineal Múltiple

La regresión lineal múltiple es una extensión de la regresión lineal, en la cual se tiene una variable dependiente a estudiar y múltiples variables independientes. El objetivo de la regresión múltiple era modelar la relación entre la variable dependiente y varias variables independientes. En este modelo es muy importante tratar la *heterocedasticidad*, la *multicolinealidad* y las diferentes métricas que tenemos para evaluar el modelo, asegurando que sea fiable y robusto.

2.2.2 Regresión Polinomial

La regresión polinomial es otro tipo de extensión de la regresión lineal en la que se modela la relación entre la variable dependiente y una o más variables independientes, pero en lugar de asumir una relación lineal, se asume que la relación puede ser representada por un polinomio de mayor grado. El objetivo de la regresión polinomial es capturar la relación no lineal entre las variables, permitiendo un ajuste más flexible a los datos cuando una línea recta no es suficiente. En este modelo, es importante tratar la *multicolinealidad*, que se puede generar de manera más sencilla debido a los términos polinomiales.

2.2.3 Test estadísticos

Durante el proyecto, se realizaron varios tests para confirmar si existen diferencias significativas entre varios grupos. Entre todos los tests, destacamos las siguientes técnicas:

- *T-Test*: test estadístico paramétrico que nos permitirá estudiar la diferencia entre los dos grupos es significativa. Se necesitan que los datos sean independientes, que sigan una distribución normal y una igualdad de varianza [3].

- *Wilcoxon*: test estadístico no paramétrico que tiene las mismas condiciones comentadas al t-test excepto que no deben seguir distribución normal.

- *ANOVA*: test estadístico que consiste en calcular la media de cada grupo y comparar la varianza de estas medias. Es muy útil para comparar más de tres grupos que, en este caso, solo tenemos dos.

2.3 Técnicas de Machine Learning

2.3.1 Decision Tree

Es un método de aprendizaje supervisado que hemos utilizado tanto para la clasificación como para la regresión. Este modelo, se representa la relación entre la variable dependiente y una o varias variables independientes mediante una estructura de árbol, donde los nodos internos representan pruebas y condiciones sobre las variables independientes, las ramas representan los resultados de estas pruebas y las hojas finales representan las predicciones. El objetivo principal es dividir el conjunto de datos en subconjuntos más pequeños que sean más homogéneos con respecto a la variable dependiente. En este modelo es importante considerar la profundidad del árbol, la selección de características y el manejo de *overfitting*, ya que podrían afectar a la generalización y precisión del modelo [4].

2.3.2 Random Forest

Es un método de aprendizaje supervisado que podremos utilizar tanto para tareas de clasificación como de regresión. Este modelo, se genera un conjunto de árboles de decisión a partir de subconjuntos aleatorios de las variables que tenemos y finalmente se realiza el promedio o se clasifican los datos para obtener un resultado final más estable y preciso. Hay que tener en cuenta los parámetros como el número de árboles y la profundidad máxima de estos ya que afectan directamente al rendimiento. Por último, este modelo tolera algo mejor los problemas de multicolinealidad y *overfitting* en comparación al modelo anterior, pero aún será importante estos términos para evaluar que el modelo sea confiable y eficaz [5].

2.3.3 Support Vector Machine

Este método de aprendizaje supervisado utilizado para ambos tipos de tareas, pero especialmente útil en tareas de clasificación. Se basa en encontrar un hiperplano que separe de la mejor forma posible las dos clases diferentes asegurando que las instancias de cada clase queden lo más lejos posible del hiperplano y consiguiendo así minimizar el error de predicción. SVM es especialmente fuerte en problemas de clasificación con dimensiones altas y cuando las clases no son linealmente separables, utilizando funciones kernel para transformar los datos en un espacio de mayor dimensión donde las clases puedan ser separadas linealmente. Los parámetros que hemos utilizado y seleccionado en función del problema que queremos abordar con nuestros datos son: la selección del kernel, el parámetro de regularización (C), y el margen al construir el modelo, ya que estos parámetros afectan directamente su capacidad para generalizar [6].

2.3.4 Multi-Layer Perceptron

Es un tipo de red neuronal que se utiliza para tareas de clasificación o para de regresión, según la tarea que necesitemos. Este tipo de red neuronal se basa en una red de nodos (neuronas) organizadas en diferentes capas. Estas capas están divididas en una capa de entrada, una o más capas ocultas y una última capa de salida. Cada neurona en una capa está conectada a las neuronas de las siguientes capas mediante pesos que se ajustan durante el proceso de entrenamiento para minimizar el error de predicción. Este modelo es especialmente útil en capturar relaciones no lineales y patrones complejos. Los parámetros que hemos configurado y seleccionado para mejorar el rendimiento de este modelo han sido el número de capas ocultas, la función de activación, el tamaño del lote y el número de iteraciones [7].

2.3.5 Modelo ensamblado

Es una técnica de aprendizaje supervisado que combina múltiples modelos de predicción para mejorar el rendimiento predictivo en comparación con un solo modelo. La idea principal de esta técnica es que, al combinar las predicciones de varios modelos, se pueden reducir los errores y las debilidades de cada uno, logrando un modelo final más robusto y preciso [8]. Existen varios enfoques para ensamblar modelos, y los más comunes son:

- *Bagging*: Combina diferentes modelos creados en diferentes partes de los conjuntos de datos original.

- *Boosting*: Los modelos se construyen de manera secuencial. Cada modelo intenta corregir los errores de su predecesor.

- *Stacking*: Combina múltiples modelos entrenados en el conjunto de datos original o en subconjuntos de este, y luego entrena un modelo final que aprende de los resultados combinados.

2.3.6 Búsqueda de los parámetros óptimos

Cada modelo tiene sus propias características y hay que definirlas antes de comenzar con el proceso de entrenamiento del modelo. Estos parámetros son importantes porque determinan como el modelo aprenderá de los datos. A partir de estos parámetros, cabe la posibilidad de aumentar la complejidad del modelo, el tiempo de computación, la tasa de aprendizaje, la capacidad de generalización, entre otros.

Nuestro objetivo principal ha sido encontrar la combinación más optima de los parámetros de los modelos. Para ello, hemos utilizado las herramientas **Grid Search** y **Random Search** para automatizar el proceso de búsqueda de los mejores parámetros para optimizar nuestros modelos, asegurando así, encontrar una combinación de características óptima. Estas dos herramientas utilizan un diccionario con varias características del modelo seleccionado y a partir de ahí, se realiza la búsqueda de los mejores parámetros de diferente manera. La técnica de **Grid Search** realiza una búsqueda exhaustiva en una cuadrícula predefinida por el diccionario mencionado, explorando todas las combinaciones posibles para identificar la configuración óptima dentro del espacio de búsqueda. Una de las principales desventajas que he observado ha sido el tiempo de computación, que puede ser considerablemente alto. Esto se debe a que el proceso implica la evaluación de un gran número de combinaciones, y cada configuración requiere su propio tiempo de entrenamiento. El **Random Search** en lugar de evaluar todas las combinaciones posibles, selecciona al azar un subconjunto de combinaciones del diccionario y realiza las pruebas. El tiempo de computación disminuye respecto a la otra herramienta, pero no garantiza que la combinación sea la más optima [9].

2.4 Métricas de evaluación

Una vez que tengamos el modelo creado y entrenado nos encontramos con la necesidad de evaluar si el modelo es bueno o no con respecto a la tarea que hemos definido inicialmente. Para definir si un modelo es bueno o no deberemos tener en cuenta que tipo de modelo estamos entrenando y cuál es nuestro objetivo. Para un modelo de clasificación, el

modelo será bueno si se asigna correctamente la clase predicha a un conjunto de características que no se han utilizado en el conjunto de entrenamiento. Para el modelo de regresión se evalúa la diferencia entre la predicción y el valor real, que es conocida como el error de la predicción y a partir de este error se pueden obtener otras métricas.

2.4.1 Modelos de Clasificación

Las métricas utilizadas se basan en la tasa de aciertos de las predicciones de las diferentes etiquetas. Estas son las métricas que hemos utilizado para seleccionar el modelo [10]:

- Matriz de Confusión: Para un modelo de clasificación binaria se indican cuatros tipos de predicciones que son: Verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. A partir de esas predicciones, podemos las siguientes métricas que nos permitirán evaluar el modelo y definir si es bueno para nuestra tarea o no.

- Accuracy: Es la métrica más completa que representa el porcentaje total de los valores correctamente clasificados, ya sean positivos como negativos.

$$Accuracy = \frac{(VP + VN)}{(VP + VN + FP + FN)}$$

- Precision: Similar a la métrica anterior pero el porcentaje representado muestra las clasificaciones positivas acertadas.

$$Precision = \frac{VP}{(VP + FP)}$$

- Recall: Proporción de verdaderos positivos entre todos los casos que realmente son positivos.

$$Recall = \frac{VP}{(VP + FN)}$$

- F1-Score: Esta métrica combina las métricas precisión y recall para obtener un valor mucho más objetivo. Es muy útil para nuestro caso, que tendremos datos desbalanceados.

$$F1\ Score = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

- ROC (*Receiver Operating Characteristic*): Es un gráfico que muestra el rendimiento de un modelo de clasificación binaria en diferentes umbrales de decisión. Se basa en una curva que traza la tasa de verdaderos positivos frente a la tasa de falsos positivos (TPR y FPR respectivamente) a medida que se varía el umbral de decisión del clasificador. Lo ideal es que esta curva se acerque lo máximo posible al eje Y ya que indicaría mayor capacidad de discriminación.

$$TPR = \frac{VP}{VP + FN}$$

$$FPP = \frac{FP}{FP + VN}$$

- AUC (*Area Under the Curve*): Es el área bajo la curva ROC y proporciona un único valor que resume la capacidad del modelo para discriminar entre clases positivas y negativas. Este valor se asume dentro del rango 0 y 1.

2.4.2 Modelos de Regresión

Las métricas para el modelo de regresión se basan en el error entre el valor real y la predicción, ese error se calcula realizando la diferencia. Por lo tanto, las métricas que hemos utilizado son las siguientes [11]:

- Error: Es la diferencia entre la predicción y el valor real.

$$Error = Resultado\ real - Resultado\ predicho$$

- MSE (Error Cuadrático Medio): Es el promedio de los errores al cuadrado entre los valores predichos por el modelo y los valores reales. Penaliza los errores grandes al elevar al cuadrado el error.

$$MSE = \frac{1}{N}\sum_{i=0}^{n}(Error)^2$$

- RMSE (Raíz Cuadrática Media del error): Simplemente es la raíz cuadrada del error medio, pero muestra el error en las mismas unidades que la variable dependiente.

$$\mathrm{R}MSE = \sqrt{MSE}$$

- MAE (Error Medio Absoluto): Es la media de los valores absolutos de los errores. Penaliza de manera lineal los errores, lo que hace más robusto a los outliers en comparación a las dos métricas anteriores.

$$MAE = \frac{1}{N}\sum_{i=0}^{n}|\ Error\ |$$

- R^2 - Coeficiente de Determinación: Compara la métrica MSE con la suma de los errores al cuadrado de un modelo base (predice la media de los valores reales). Esta métrica representa qué tan bien nuestro modelo explica la variabilidad de los datos, es decir, que porcentaje de los cambios en la variable dependiente puede ser explicado por las variables independientes.

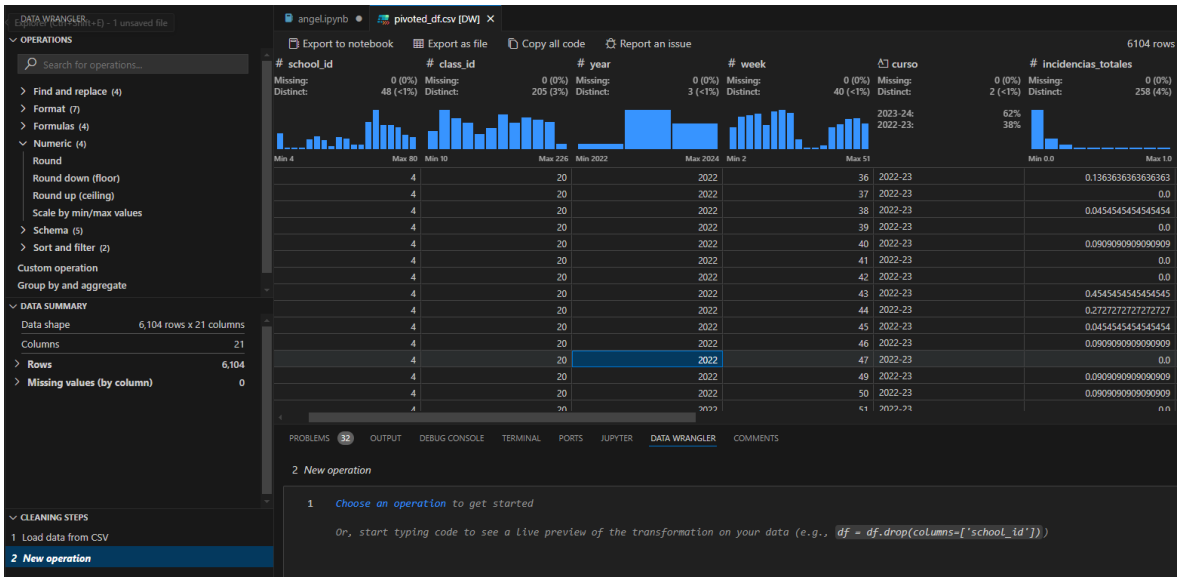
$$R^2 = 1 - \frac{MSE}{\sum_{i=0}^{n}(Valor\ real - Media\ de\ los\ valores\ reales)^2}$$

2.5 Tecnología usada

El entorno utilizado para crear nuestros programas y poder realizar los estudios, los procesamientos de datos y los análisis será **Jupyter Notebook**. Lo utilizaremos como extensión en el **Visual Studio Code** y esta extensión nos permitirá mejorar la estructura de nuestro código en diferentes celdas de contenido, como son: las celdas de códigos, las celdas markdown (lenguaje de marcado sencillo) y las celdas de resultados. Otra de las ventajas utilizadas sobre este entorno es la interactividad que tenemos en nuestro código, escribiendo y ejecutando fragmentos de código sin necesidad de ejecutar el código completo y guardando los resultados, variables o diferentes estructuras en la memoria de nuestro ordenador. Hemos seleccionado Python como el lenguaje de programación principal, aunque este entorno también soporta diferentes lenguajes de programación como son R, Julia y Scala, entre otros.

Sobre la visualización de datos utilizaremos la extensión **Data Wrangler** creada por Microsoft. Esta extensión está diseñada especialmente para la ciencia de datos y el aprendizaje automático. Nos permite visualizar las estructuras de datos, resumen de estos datos, nos permite crear código y ejecutar las funciones más frecuentes. Utilizaremos librerías como **Seaborn** y **Matplotlib** para generar gráficas de los resultados que obtenemos para poder facilitar la interpretación de estos.

Figura 2. Ejemplo de visualización de un dataframe en Data Wrangler



Para cumplir nuestro objetivo en este proyecto, utilizaremos la librería **Scikit-learn** [12], diseñada y recomendada para problemas de Data Science. Es una librería que nos ofrece un gran conjunto de herramientas como son los algoritmos de aprendizaje automático, algoritmos de clasificación, algoritmos de regresión, preprocesamiento de datos, evaluación y validación de modelos, entre muchos otros. Estas herramientas combinadas con una buena documentación que nos proporciona una explicación detallada sobre que los diferentes modelos, con sus ejemplos prácticos, sus métricas de evaluación, consejos sobre cómo mejorarlos y lo que es mejor, nos proporcionan información sobre estudios que se centran en la comparativa entre modelos en caso de que no nos decidamos en el modelo, facilitando muchísimo el uso de las herramientas proporcionadas y reduciendo el tiempo de aprendizaje y de programación sobre ellas.

3 Estudio previo

En este proyecto, el análisis de datos es una técnica fundamental para obtener información sobre los datos recibidos y partir de esta información, utilizar herramientas de visualización para ayudar a visualizar los resultados analizados y así, poder realizar conclusiones para la toma de decisiones. Disponemos de. tres tipos de *datasets* con los que trabajaremos:

- Datos relacionados con las distribuciones de las aulas.
- Datos sobre la salud
- Datos ambientales

Estos *datasets* comparten en común algunos campos que nos facilitarán el poder realizar búsquedas cruzadas entre diferentes los diferentes conjuntos. Estos campos son: el nombre del colegio, el código identificador del colegio, el nombre del aula y el identificador de esta aula.

3.1 *Dataset* de distribuciones de las aulas

Este *dataset* es esencial para estudiar las condiciones y características de las diferentes aulas que están colaborando con el proyecto. Con un total de 31 columnas para representar, entre otros, sobre la distribución, la orientación y la ubicación del aula. Los campos más relevantes son:

1. **school_code_gencat**: Id generado por la generalitat de Catalunya.
2. **nature**: Tipo de colegio (público o privado).
3. **city**
4. **comarca**
5. **provincia**
6. **latitude**: Latitud geográfica de la ubicación del colegio.
7. **longitude**: Longitud geográfica de la ubicación del colegio.
8. **num_children**: Capacidad de niños de la clase.
9. **class_lenght**
10. **class_width**
11. **class_height**
12. **ac**: Presencia de aire acondicionado en el aula.
13. **heating**: Presencia de calefacción en el aula.
14. **height_kit**: Altura del kit.
15. **floor**: Piso en el que se encuentra el aula dentro del colegio.
16. **orientation**: Orientación geográfica del aula (norte, sur, este, oeste).
17. **observations**: Observaciones adicionales o comentarios sobre el aula.

Además, la aplicación tiene en consideración la ubicación de ciertos elementos utilizando una estructura cuadriculada y dividiéndola en 9 segmentos. La estructura utilizada queda de la siguiente manera:

Tabla 1. Distribución de las diferentes ubicaciones del aula

	2			
1	7	8	9	3
	4	5	6	
	1	2	3	
	0			

La distribución mostrada en la (Tabla 1) se utiliza para ubicar ciertos elementos que podremos tener en cuenta a la hora de tomar decisiones. La puerta principal de la clase se ubicará en la pared '0' y la clase se dividirá en 9 segmentos y en 4 paredes (0-3). Estos elementos que utilizan estas coordenadas son:

- **loc_door**
- **loc_window**
- **loc_teacher**: Localización del profesor dentro del aula.
- **loc_kit**: Localización del kit.
- **wall_blackboard**: Localización de la pizarra.

3.2 *Dataset* sobre la salud

Este *dataset* nos proporciona información sobre las ausencias de los alumnos de las aulas reportando así, varios tipos de ausencias/incidencias. Estos reportes se generan de manera semanal y son los profesores de las aulas los encargados de completar un formulario indicando todo tipo de incidencias o ausencia que hayan podido detectar. Este factor humano externo se puede considerar un problema a lo largo del estudio porque se han detectado algunos reportes vacíos, dando a entender, que por algún motivo los profesores no han completado el formulario. Esos reportes constan de las siguientes variables relevantes:

- **num_childs_sympt**: alumnos que presentan algún síntoma de virus respiratorio.
- **num_childs_miss_ill**: alumnos ausentados por un virus respiratorio.
- **num_childs_miss_others**: alumnos ausentados por otros motivos.
- **num_childs_miss_unknown**: alumnos ausentados por motivos desconocidos.
- **num_childs_ill_twodays**: alumnos ausentados más de dos días por virus respiratorios.
- **use_childs_masks**: cantidad de alumnos que han utilizado mascarillas.
- **teacher_health:** estado de salud del profesor/a, dando las siguientes opciones:
    - **0:** No presenta síntomas o problemas respiratorios.
    - **1:** Sí presenta síntomas o problemas respiratorios.
    - **2:** Prefiere no contestar.

3.3 *Dataset* sobre datos ambientales

Este *dataset* nos proporcionará información sobre las variables ambientales que se reciben de las diferentes aulas. Para recibir estos datos, el equipo preparó un kit de sensores y herramientas que permiten realizar lecturas de los datos ambientales. Los sensores que se han utilizado son los siguientes:

Tabla 2. Características de los sensores utilizados en el kit

SENSOR	FABRICANTE	VARIABLE	UNIDAD	PRECISIÓN
ENVIRO+	Pimoroni	Temperatura	ºC	± 1ºC
		Humedad relativa	%	3%
		Presión atmosférica	hPa	± 1% hPa
		Luminosidad	Lux	Depende de la fuente
SCD30	Sensirion	Temperatura	ºC	± 0.4ºC
		Humedad relativa	%	± 3%
		CO_2	ppm	± 30ppm
PMS5003	Adafruit	Concentración de partículas (PM1, PM2.5 y PM10)	µg/m3	± 10% rango [100-500] µg/m3
LTR390	Adafruit	Luminosidad	Lux	≈ 25% depende de la fuente
		Radiación UV	UVI	± 1 UVI
MAX9814	Adafruit	Sonido	dB	± 1%
HUSKY-LENS	DFRobot	Movimiento	Índice numérico	Depende del entorno
IMÁN	RS Pro	Proximidad (apertura y cierre de puertas y ventanas)	Si / No	Errores en distancias cercanas a 1cm

Podemos observar en la (Tabla 2) que se reciben varias variables provenientes de múltiples sensores. Para manejar esto, hemos calculado la media de las variables repetidas (como la Temperatura), considerando la precisión de cada sensor. De esta manera, reducimos la cantidad de datos al estudiar el efecto de las diferentes variables. Los sensores realizan lecturas de manera frecuente generando una enorme cantidad de datos.

Finalmente, las variables a estudiar que obtenemos con los kits instalados en las distintas aulas son las siguientes:

- **CO_2**
- **TEMP**
- **HUM**
- **MOVEMENT**
- **NOISE**
- **INDEX_UV**
- **LLUM**
- **PM1**
- **PM25**
- **PM10**
- **PROX**
- **PRES**

3.4 Preparación del *dataset*

Teniendo en cuenta la gran cantidad de datos ambientales que tenemos en *dataset* final y que tenemos una cantidad de colegios considerables, hemos adoptado una estrategia para valorar como podemos estudiar e interpretar esta gran cantidad de datos. Cada sensor funciona de una manera independiente y hace una lectura cada cierto tiempo, así que, hemos agrupado todos los datos de manera diaria, calculando la media de cada sensor, y generando así un registro diario por cada sensor. Tras revisar esta estrategia, hemos detectado un problema y es que las medias se verían afectadas por los valores que corresponden a horarios no-lectivos. Para corroborar esta teoría hemos diferenciado dos grupos diarios: todos los datos que corresponden a un día y todos los datos, en horarios lectivos, que corresponden al mismo día. Comprobaremos que la diferencia entre estos dos grupos es realmente significativa, y en caso de serlo, centraremos toda nuestra atención en el grupo de horarios lectivos. Inicialmente, comprobamos que nuestros datos no siguen una distribución normal así que hemos decidido de utilizar la prueba *Wilcoxon* y, tras realizar la prueba obtenemos los siguientes resultados:

- `Wilcoxon statistic: 26400454498.0`
- `p-value: 0.0`

Tras agrupar todos los datos que tenemos de todos los colegios y separarlos en dos grupos diarios, podemos validar que hay una diferencia estadísticamente significativa entre los dos grupos. Finalmente, escogemos el grupo de los días que únicamente tienen en cuenta los datos en horarios lectivos para poder manejar, filtrar e interpretar los datos de una manera más sencilla.

Dado que tenemos varios conjuntos de datos que contienen una gran cantidad de variables disponibles, para poder empezar a estudiar los diferentes impactos, hemos decidido

simplificar varias variables para que nos sea más fácil analizar el impacto de ciertas variables con otras variables de diferentes conjuntos de datos. Entre tantas, hemos obtenido:

3.4.1 Incidencias totales

Los profesores reportan varios tipos de incidencias (explicados en el ***Dataset* sobre la salud**) el cual, nosotros nos centraremos en las incidencias respiratorias. Hemos creado una nueva variable numérica '**Incidencias_totales**' que se asignará a la semana anterior de la semana en la que se reportan las incidencias. Esta variable numérica se ha obtenido a través de varias variables de diferentes *dataset* a partir de la siguiente formula:

$$Incidencias_{totales} = \frac{\left(num\ childs_{sympt} + num\ childs_{miss\ ill} + teacher_{health}\right)}{(num_{childs} + 1\ )}$$

3.4.2 Volumen del aula y variable categórica del tamaño

Para los estudios que tengan relación con el tipo de variables de arquitectura, hemos creado dos nuevas variables categóricas que nos indicarán el tamaño de la clase y la densidad de alumnos que hay en la clase según el tamaño. Para obtener el tamaño de la clase hemos aplicado la siguiente fórmula:

$$volumen_{aula} = longitud * altura * amplitud$$

Una vez obtenida la variable volumen, hemos utilizado los percentiles para clasificar las aulas en diferentes categorías [Pequeña, Media-Pequeña, Media-Grande y Grande] para obtener más detalles en los datos dividiendo los datos en más grupos para poder capturar más variabilidad.

3.4.3 Densidad de alumnos y variable categórica de la densidad

Calcularemos la densidad de personas según el volumen y el número de personas que hay en la clase (número de alumnos + 1 profesor).

$$densidad_{aula} = \frac{(num_{childs} + 1)}{volumen_{aula}}$$

Siguiendo el mismo procedimiento utilizado en (Volumen del aula y variable categórica del tamaño), clasificaremos las densidades de las aulas en varios niveles [Bajo, Medio-Bajo, Medio-Alto, Alto].

3.4.4 Conjunto final

Tras la creación de estas nuevas variables y teniendo información relevante en diferentes *datasets*, se ha realizado una combinación, seleccionando las variables más importantes de cada conjunto y unificándolas en un único conjunto a estudiar. Esta unificación es posible gracias a la variable ***class_id***, que está presente en todos los conjuntos y permite identificar las variables correspondientes a cada clase. También se desglosó el formato de las variables ambientales. Originalmente, todas las variables ambientales estaban en una misma columna y para identificar el tipo de la variable que era con su valor numérico hacía falta otra columna llamada ***parameter_id***. El objetivo de esta transformación es tener

las diferentes variables ambientales identificadas como columnas para facilitar en tareas como son la comparación y los análisis.

Finalmente nos queda un único conjunto de datos a analizar con todos los datos relevantes que hemos considerado. Este *dataset* ofrece:

- **Todas las variables ambientales**
- **Incidencias Totales**
- **Densidad del aula y la categoría que la identifica**

4 Análisis de correlaciones

Tras la estrategia de filtrado y la agrupación de datos que hemos realizado, iniciamos este estudio con varios análisis de correlación. Nuestro objetivo ha consistido en buscar y analizar correlaciones significativas entre diferentes variables. Con este tipo de análisis, buscamos identificar qué variables pueden estar relacionadas entre sí, ya sean del mismo conjunto de datos o de otro.

4.1 Correlaciones variables ambiente-ambiente

Inicialmente, hemos seleccionado todas las variables ambientales de todos los colegios y realizaremos una primera búsqueda de correlaciones para visualizar que variables pueden tener relación entre ellas. Esta práctica nos ayuda a comprender como una variable ambiental puede influir en otra e incluso, verificar la consistencia de los datos ya que hay datos que deberían tener una fuerte correlación (como por ejemplo PM1, PM2.5 y PM10) y si estas correlaciones no apareciesen, deberíamos revisar los datos para localizar algún tipo de inconsistencia o problema.

Se han utilizado los dos tipos de correlaciones para poder combinar las correlaciones y obtener unos datos consistentes. La correlación de *Pearson* nos permite capturar la relación lineal y la fuerza de esta relación mientras que la correlación de *Spearman* nos indica la relación monótona entre las dos variables, dando la posibilidad de que esta relación no sea lineal, indicando así, que la variable puede cambiar con el cambio de otra variable, pero no necesariamente de forma lineal. Combinando los dos métodos, podemos entender mejor el comportamiento entre las variables que tenemos y tomar decisiones más consistentes basándonos en el tipo de relación que existen entre ellas.

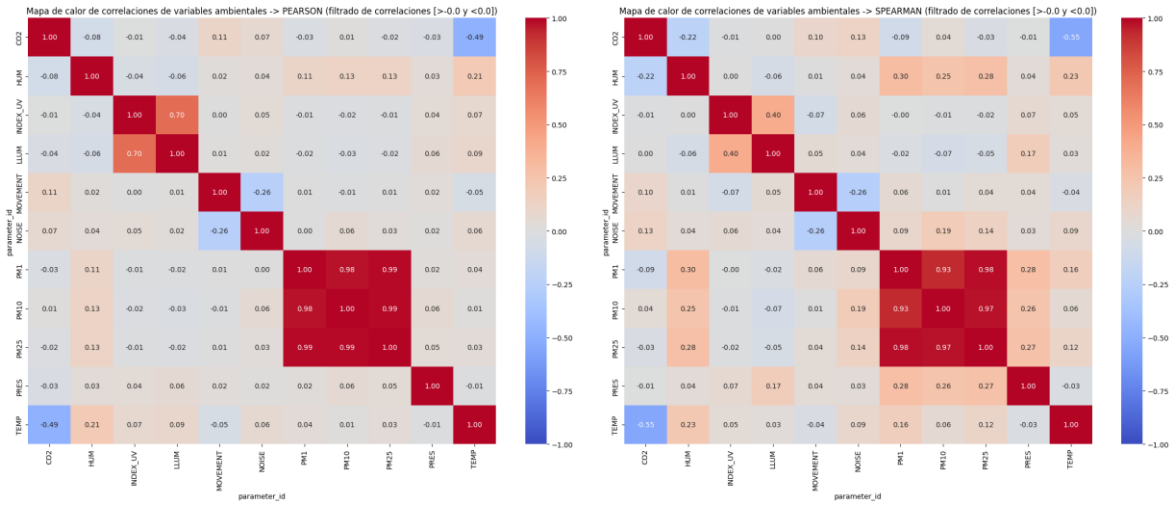
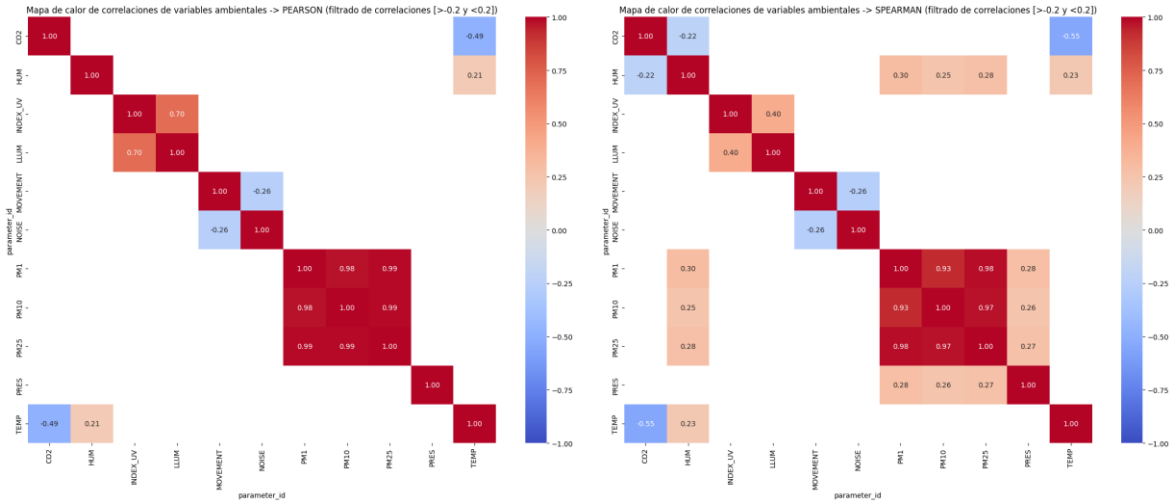


Figura 3. Mapa de calor de todas las correlaciones de las variables ambientales

En la (Figura 3) destacamos que los dos métodos utilizados se complementan bien, mostrando unos resultados muy similares. Las dos técnicas coinciden con la mayoría de las direcciones de las correlaciones, dando a entender, que el efecto entre las variables es claro.

Observamos que hay correlaciones interesantes para investigar y profundizar. El color rojo oscuro muestra una correlación positiva fuerte y el azul oscuro fuerte una correlación negativa fuerte. Con la gran cantidad de correlaciones que hemos obtenido, la interpretación se vuelve difícil, por lo que hemos realizado un filtro para visualizar las correlaciones más fuertes.

Figura 4. Mapa de calor de las correlaciones fuertes de las variables ambientales



Tal y como podemos observar en la (Figura 4), visualizaremos las correlaciones más fuertes y que en nuestro caso, nos interesa profundizar:

- Las relaciones PM1, PM 2.5 y PM10 tienen una alta correlación entre sí ya que estas variables miden diferentes tamaños de partículas del aire. Además, estas partículas tienen una relación moderada positiva con la presión atmosférica y la humedad (0.25 – 0.30) en el modelo de *Spearman*. Teniendo en cuenta esta correlación, podríamos eliminar dos categorías del tipo de partículas en un futuro para simplificar.

- El índice de ultravioleta (INDEX_UV) y la luminosidad tienen una moderada correlación positiva (0.70) en el modelo *Pearson* y una correlación débil positiva (0.40) en el modelo *Spearman*.

- Temperatura y humedad tienen una correlación positiva débil (0.21 / 0.23).

- El CO_2 tiene una correlación moderada negativa con la temperatura (0.49 / 0.55) y una correlación débil negativa (0.22) con la humedad del modelo *Spearman*.

- El ruido y el movimiento tienen una correlación débil negativo (0.26)

Tras estos resultados, hemos realizado la prueba T-Student para validar si la correlación entre las variables ambientales es significativamente diferente de 0. Un p-valor < 0.05 nos estaría sugiriendo rechazar la hipótesis nula que, en este caso, nos indica que la correlación obtenida es estadísticamente significativa y que no se debe al azar. Realizaremos una tabla para mostrar las correlaciones más significativas que hemos detallado:

Tabla 3. Correlaciones Pearson de variables ambientales significativas con su p-valor

VARIABLES	CORRELACIÓN PEARSON	P-VALOR
PM1 – PM10	0.98	0.0
PM1 – PM25	0.99	0.0
PM25 – PM10	0.99	0.0
INDEX_UV – LUM	0.70	0.0
TEMP – HUM	0.21	5.08e-42
CO_2 – TEMP	-0.49	4.89e-25
CO_2 – HUM	0.21	1.67e-07
NOISE – MOVEMENT	-0.26	1.28e-64

En la (Tabla 3) observamos las correlaciones obtenidas más fuertes teniendo un valor, igual o muy cercano a 0, y nos sugiere que las correlaciones obtenidas son fiables y no un error de muestreo.

Altos niveles de CO_2 sugieren una baja ventilación y temperatura, lo que puede aumentar el riesgo de transmisión de virus en ambientes cerrados. Una época del año con similares características sería el invierno, donde hay una menor temperatura, una menor ventilación a causa de estar todo cerrado y eso conllevaría a una mayor concentración de CO_2. Además, los niveles de humedad muy bajos pueden permitir que los virus permanezcan en el aire por más tiempo, mientras que niveles muy altos pueden incrementar la condensación y la transmisión por superficies. Otra prueba realizada ha sido dividir los datos en cursos y observar los resultados, sin embargo, no hemos obtenido unas correlaciones con diferencias significativas.

Este análisis de correlaciones nos ha proporcionado una base para entender cómo las diferentes variables ambientales interactúan entre sí y para más adelante, entender la propagación de los virus respiratorios. En nuestro caso, nos interesa estudiar la calidad del aire (variables como las partículas, CO_2, humedad) y la densidad de personas que puede haber en las clases (ruido y movimiento), que podrían considerarse uno de los factores más importantes en el impacto de las incidencias.

4.2 Correlaciones variables ambiente-salud

En este estudio realizado, nos hemos centrado en estudiar el impacto que tienen las variables ambientales respecto a las incidencias que reportan los profesores. Teniendo la variable que hemos calculado para obtener el número de incidencias que se ha generado en una semana, hemos analizado las variables ambientales de esa semana y expandido el estudio a todas las semanas del curso de todos los colegios.

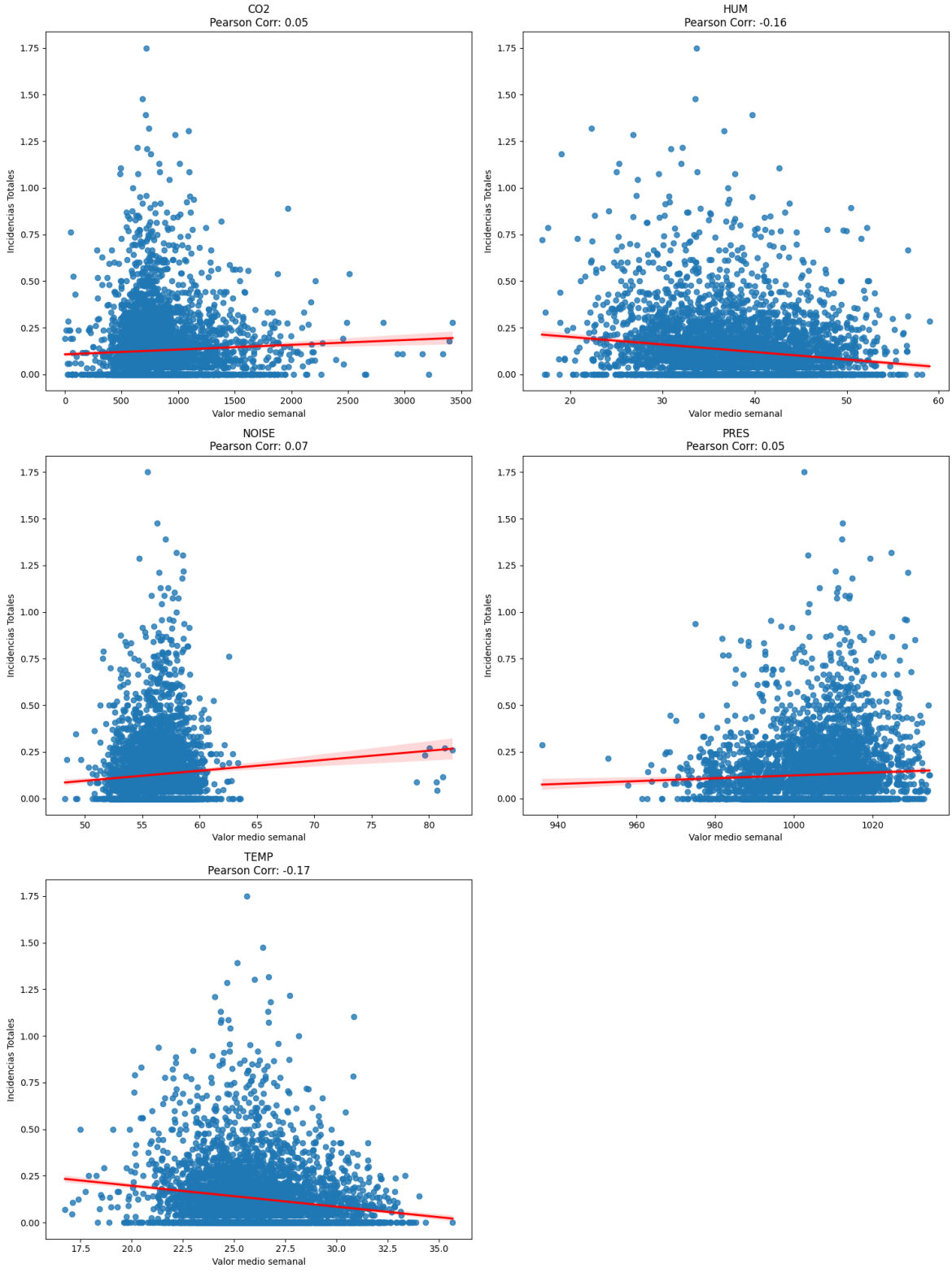
Tabla 4. Correlaciones Pearson / Spearman de las variables ambientales con las incidencias reportadas

VARIABLE	CORR SPEARMAN	CORR PEARSON	P-VALOR
CO_2	0.099	0.045	0.001
HUM	- 0.141	-0.155	4.05e-25
INDEX_UV	- 0.029	-0.004	0.925
LLUM	0.021	0.010	0.069
MOVEMENT	0.013	0.015	0.352
NOISE	0.08	0.06	8.99e-06
PM1	- 0.02	- 0.0124	0.212
PM25	-0.005	0.001	0.372
PM10	0.005	-0.003	0.452
PRES	0.029	0.046	0.001
TEMP	- 0.186	-0.16	2.58e-28

Tras varias pruebas, se ha obtenido finalmente los resultados que visualizamos en la (Tabla 4**)**, que nos proporciona las variables relevantes a la hora de estudiar las incidencias reportadas. Podemos observar que hay correlaciones extremadamente débiles y casi indicando que apenas tendrían impacto y, además, hay correlaciones que se pueden descartar, ya que aceptamos la hipótesis nula al tener un p-valor superior a 0.05. Aun así, nos vamos a centrar en las variables importantes de esta tabla que son el CO_2, la temperatura, la humedad, la presión y el ruido. Este conjunto de variables tiene un impacto en la variable de **Incidencias totales** de tal manera, que el p-valor obtenido no es suficiente para aceptar la hipótesis nula, indicando así, que la correlación obtenida es fiable y nos puede ser útil de cara a estudiarlas.

Los siguientes gráficos de dispersión muestran la distribución que siguen los valores recibidos semanales, detección de presencia de outliers e incluso algún tipo de patrón que nos permita identificar algún incremento o decremento sobre las incidencias.

Figura 5. Plot de las variables relevantes junto al impacto en las incidencias



En la (Figura 5), podemos observar que las variables tienen un intervalo medio (excepto la variable de humedad) donde podemos observar que hay un aumento en el número de incidencias reportadas. Para el CO_2 observamos que, si se aumenta considerablemente, el indicador de incidencias también aumenta, parecido al de la variable de la presión. Obtenemos un efecto contrario al de la temperatura, pero un intervalo de [25ºC–27ºC] donde

el número de incidencias aumenta hasta llegar a un número de reportes alto. Para la variable humedad cuesta más de detectar algún tipo de patrón ya que están todos los valores muy dispersos.

Se ha realizado también un desglose por cursos para ver si hay alguna diferencia en las correlaciones y en el p-valor indicando algún descarte en esta correlación. Para las correlaciones se han obtenido algún incremento/decremento respecto a la relación con el número de incidencias reportadas. Respecto al p-valor, las variables ambientales siguen teniendo unos valores similares excepto para el CO_2, que en el curso 2023-24, hemos obtenido un p-valor mayor a 0.05 aceptando la hipótesis nula de que la correlación puede haberse generado por error de muestreo o por azar. Aun así, trataremos esta variable como relevante a la hora de tomar decisiones con las incidencias ya que con el conjunto completo y con el curso 2022-23 hemos rechazado la hipótesis nula.

Como parte de la conclusión de este primer estudio analizando el impacto de las variables ambientales en el número de incidencias respiratorias reportadas semanalmente, hemos logrado extraer las variables ambientales que más influyen al momento de generar incidencias, y esto nos ha servido para hacer un estudio multivariante y varios tipos de modelo de predicción. Tras el desglose en cursos, hemos podido obtener una consistencia en las variables e incluso obtener una variable, CO_2, que podría generar alteraciones a la hora de predecir incidencias. Aun así, debemos tener en cuenta las variables con un p-valor alto ya que también puede ser útiles en un futuro a la hora de predecir o incluir un análisis con interacciones.

4.3 Correlaciones variables ambiente-arquitectura

Dado que las variables que hemos calculado anteriormente **Volumen del aula y la densidad del aula** tienen un valor fijado por curso/anual y los registros ambientales que tenemos están calculados semanalmente, hemos optado por agrupar los registros ambientales por curso. La decisión de tomar los valores ambientales por curso en vez de semanales serviría para evitar comparar los diferentes valores ambientales y provocar un desajuste de frecuencia temporal y así, obtener una consistencia en los datos debido a la estabilidad estadística.

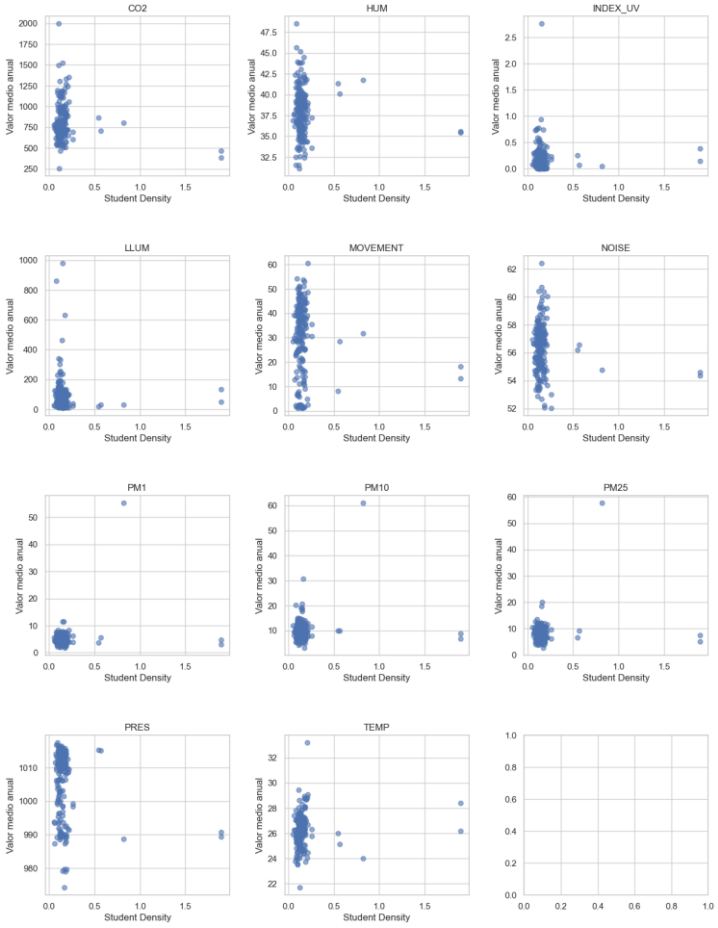
La cantidad de alumnos que puede haber en un aula, en función del tamaño de esta, es igual o más importante que el propio tamaño del aula. La concentración de personas en un espacio cerrado es un buen ejemplo para estudiar las variables ambientales. Dependiendo del número de personas, las variables ambientales pueden cambiar, facilitando la propagación de virus respiratorios.

Se ha realizado un estudio sobre el impacto de las variables ambientales sobre la densidad del aula. Inicialmente, mostraremos una tabla con las correlaciones obtenidas y su p-valor y seguidamente, mostraremos un plot con gráficos de dispersión sobre la distribución de las correlaciones de variables ambientales sobre la densidad de alumnos.

Tabla 5. Correlaciones de las variables ambientales - densidad del aula

VARIABLE	CORR SPEARMAN	CORR PEARSON	P-VALOR
CO_2	0,154	-0,137	0,055
HUM	-0,024	-0,044	0,542
TEMP	0,188	0,074	0,3
INDEX_UV	-0,149	0,006	0,937
LLUM	-0,1	-0,035	0,624
MOVEMENT	0,056	-0,097	0,177
NOISE	-0,019	-0,12	0,095
PM1	0,014	0,199	0,005
PM10	-0,001	0,148	0,039
PM25	-0,014	0,166	0,02
PRES	-0,056	-0,182	0,011

Figura 6. Plot de gráficos de dispersión de las variables ambientales – densidad del aula



Tal como se observa en la (Figura 6**),** podemos destacar una aglomeración de datos, además de que, se observan puntos demasiados dispersos que podrían considerarse outliers.

Tras investigar la causa de estos puntos dispersos, hemos validado que son datos correctos y representan aulas ‘especiales’ con una gran capacidad de personas. Se ha creado un filtro para analizar las aulas ‘normales’ donde únicamente se realizan actividades lectivas. Tras analizar los datos, hemos reducido la capacidad de número de niños a 35 ya que, un número mayor a ese umbral se podría considerar un aula especial, como son: aula de música, el comedor…

Figura 7. Plot de gráficos de dispersión de las variables ambientales – densidad del aula con el filtrado de aulas especiales

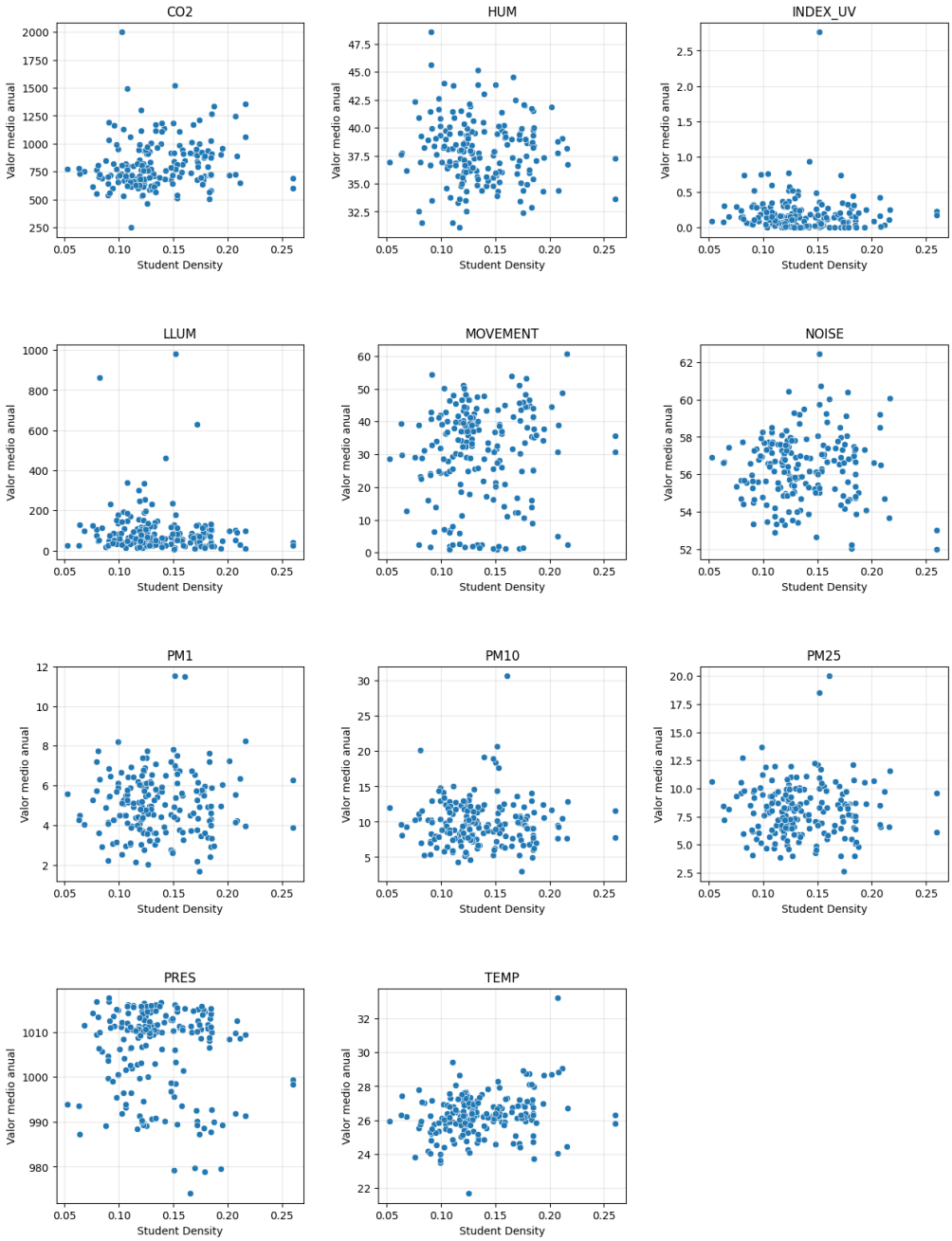


Tabla 6. Correlaciones filtradas de las variables ambientales sobre la densidad del aula

VARIABLE	CORR SPEARMAN	CORR PEARSON	P-VALOR
CO_2	0,187	0,141	0,051
HUM	-0,04	-0,059	0,414
TEMP	0,215	0,234	0,001
INDEX_UV	-0,166	-0,064	0,377
LLUM	-0,081	-0,077	0,287
MOVEMENT	0,097	0,076	0,298
NOISE	0,017	-0,02	0,78
PM1	0,02	0,033	0,651
PM10	-0,009	0,001	0,993
PM25	-0,015	-0,011	0,876
PRES	-0,042	-0,107	0,142

En el caso del conjunto de gráficos (Figura 7) podemos observar con mayor detalle los puntos diversos. Observamos que para las variables de temperatura y CO_2 hay una ligera tendencia positiva, indicándonos que cuanto mayor densidad de alumnos hay en un aula mayor es el número de estas variables. Para la presión observamos una ligera tendencia negativa y para el resto de las variables apenas se notan algún tipo de tendencia.

En la (Tabla 6) observamos que la mayoría de las variables ambientales coinciden con las correlaciones obtenidas por diferentes métodos, mostrando una pequeña variabilidad en el valor de las correlaciones. Además, hemos obtenido un p-valor alto para la mayoría de las variables, superando así el umbral y rechazando que tengan correlación con la densidad. Observamos que la variable temperatura tiene una correlación significativa con la densidad, sin embargo, la variable de CO_2 se mantiene al límite del umbral. Podríamos considerar que esta variable tiene una correlación significativa.

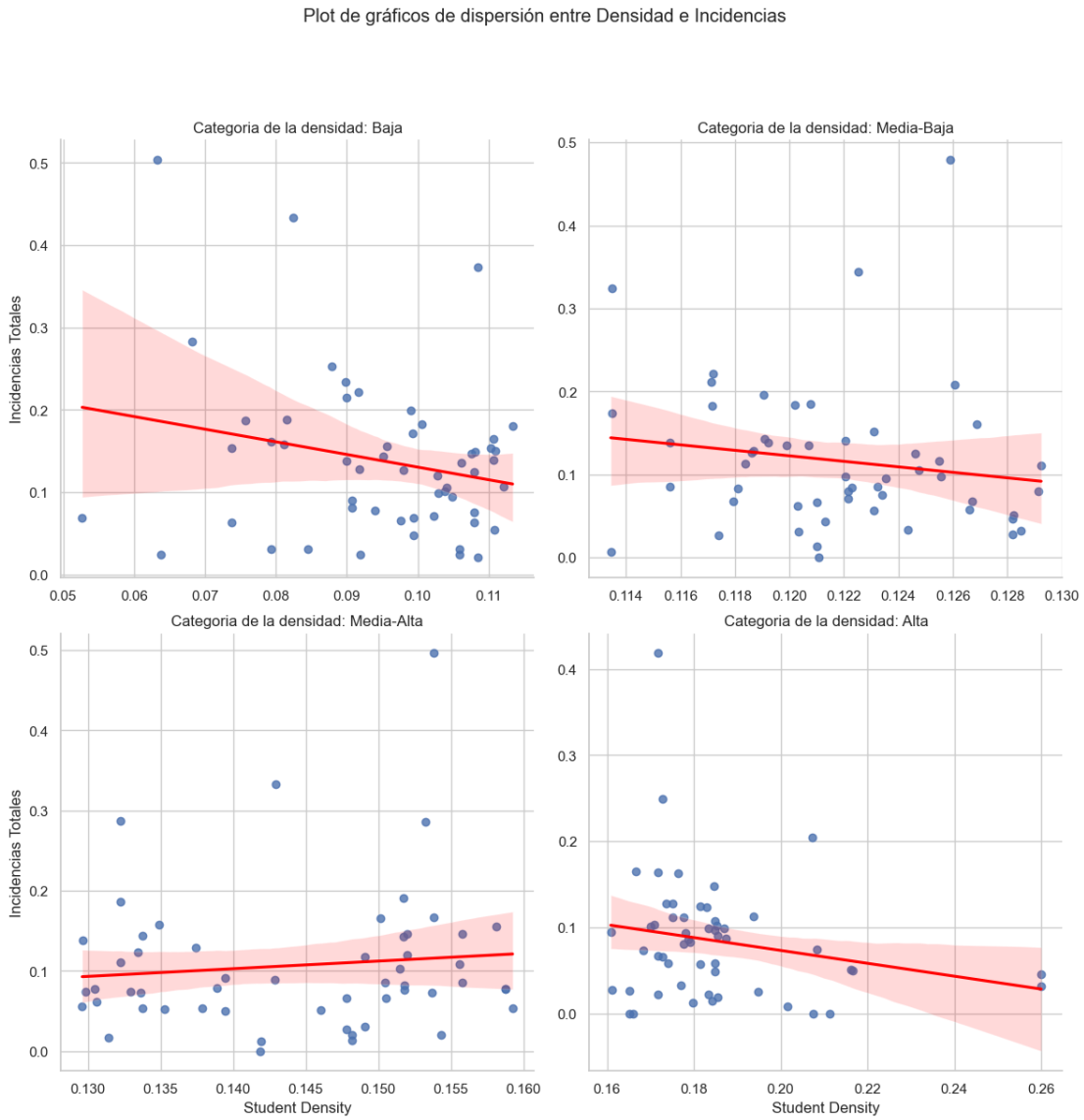
Finalmente, podemos discutir que las variables afectadas por la densidad de los alumnos en las diferentes aulas son el CO_2 y la temperatura, dando a entender que el resto de las variables no se verían apenas afectadas. Teniendo en cuenta los estudios realizados anteriormente, obtuvimos que esas dos variables tenían cierta repercusión con el incremento de incidencias y esto sugeriría, que las variables arquitectónicas tienen un impacto indirecto con las incidencias reportadas.

4.4 Correlaciones variables arquitectura - salud

Hemos realizado la misma estrategia que en el estudio anterior (**Correlaciones variables ambiente-arquitectura**) para obtener las correlaciones. Hemos agrupado el número de incidencias reportadas de manera anual y se ha obtenido la media para comparar ese valor con la densidad del aula por curso. La densidad del aula se puede clasificar en

diferentes categorías así que realizaremos el estudio sobre, como las cuatro categorías pueden afectar al número de reportes de incidencias.

Figura 8. Plot de gráficos de dispersión Densidad – Incidencias



Observamos en la (Figura 8) que la mayoría tienen una tendencia negativa a excepción de la categoría "Media-Alta" y esto sugiere que a medida que hay más alumnos, se toman más precauciones para evitar un aumento de incidencias. Aun así, calcularemos el p-valor para poder confirmar que las correlaciones tienen relación estadísticamente significativa.

Tabla 7. Correlaciones por categorías de la densidad – incidencias

CATEGORIA	CORR SPEARMAN	CORR PEARSON	P-VALOR
BAJA	-0,152	-0,228	0,094
MEDIA-BAJA	-0,245	-0,158	0,254

MEDIA-ALTA	0,154	0,105	0,451
ALTA	-0,192	-0,21	0,128

Tal y como podemos visualizar en la (Tabla 8), las dos correlaciones obtenidas coinciden con la tendencia y tienen un ligero cambio en el valor calculado, sin embargo, las cuatro categorías tienen un p-valor superior a 0.05 indicando que habría que aceptar la hipótesis nula y sugerir que las correlaciones obtenidas no tienen significancia estadística.

Estos datos sugieren que la densidad de alumnos en las aulas no tiene un impacto significativo en el número de incidencias reportadas. Sin embargo, dado que el p-valor es mayor que el umbral para rechazar la hipótesis nula, podemos inferir que estos datos podrían haberse generado aleatoriamente. No obstante, no es necesario descartar estas variables, ya que podrían ser útiles al combinarlas con otras variables o al crear un modelo de predicción, ya que podrían contribuir a mejorar la precisión.

4.5 Discusión

Tras analizar los resultados obtenidos en el estudio sobre las correlaciones de las variables principales del proyecto, hemos identificado múltiples correlaciones significativas en diversas variables ambientales, de salud y arquitectónicas, proporcionando información valiosa. Mediante diversas técnicas y la creación de nuevas variables para representar características que no teníamos, se han logrado identificar interacciones clave tanto en factores individuales como en su interacción conjunta.

Entre los hallazgos, las variables ambientales como el CO_2, la temperatura, la humedad, la presión y el ruido mostraron una correlación significativa con las incidencias totales. Estos resultados sugieren que un aumento en estas variables podría elevar el riesgo de transmisión de virus en las aulas, enfatizando la importancia de una gestión efectiva del ambiente interior para reducir la propagación de enfermedades. Además, la relación negativa entre el CO_2 y la temperatura o la humedad refuerza la idea de que una ventilación adecuada es crucial para mantener un ambiente saludable, lo que podría disminuir la propagación de virus y, a su vez, reducir el número de incidencias reportadas. Otro punto para destacar ha sido la correlación entre las variables arquitectónicas de las aulas con las variables ambientales y de salud. El estudio ha demostrado que la densidad de alumnos en las aulas tiene un impacto significativo en las variables ambientales como temperatura y CO_2, sugiriendo que las aulas con mayor densidad tienden a tener un aumento de estas dos variables. Sin embargo, se sugiere que no hay un impacto directo en los reportes. Aun así, estas variables se tendrán en cuenta en el futuro para la creación de interacciones entre otras variables e incluso podrían aportar en un modelo de predicción.

Finalmente, este estudio ha establecido una sólida base que nos permite profundizar y explorar el análisis multivariante, así como desarrollar estrategias efectivas de prevención en entornos educativos.

5 Predicción de incidencias

Tras obtener información sobre qué variables de nuestro *dataset* tienen un impacto significativo de manera independiente respecto al número de incidencias reportadas en una semana, hemos utilizado varias herramientas estadísticas que permiten examinar las relaciones entre las variables independientes y nuestra variable dependiente, que en este caso será la variable 'Incidencias Totales'. El objetivo del análisis multivariante es identificar cómo múltiples variables afectan al número de incidencias reportadas semanalmente y de qué manera, lo que incluso nos permitiría construir un modelo de predicción.

Por ello, hemos utilizado el análisis de regresión lineal, que es uno de los modelos estadísticos más utilizados para predecir una variable dependiente a partir de una o más variables independientes. Su fácil implementación, gracias a las librerías mencionadas anteriormente, me ha llevado a optar por esta vía e ir profundizando en este análisis para justificar todas las decisiones tomadas.

5.1 Modelo de Regresión

5.1.1 Regresión Lineal

Hemos utilizado un modelo de regresión lineal y realizado varias pruebas entre el conjunto de variables independientes. Inicialmente, hemos escogido el conjunto de variables que han tenido un impacto significativo en el análisis de correlaciones pero, aun así, hemos realizado las mismas pruebas para diferentes conjuntos de variables independientes. El conjunto significativo de variables independientes contiene: CO_2, HUM, TEMP, student_density, PRES y NOISE.

Tabla 8. Coeficientes del conjunto de variables significativo

VARIABLE	COEFICIENTES	P-VALOR
CO_2	-1,591E-05	0,0
HUM	-0,003	0,0
TEMP	-0,009	0,0
STD_DENSITY	-0,561	0,0
PRES	0,001	0,0
NOISE	0,005	0,0

Tal y como vemos en la (Tabla 8), se han obtenido unos coeficientes diferentes a las correlaciones obtenidas en las tablas (Tabla 4) y (Tabla 7), lo cual puede ser completamente normal, ya que en este tipo de técnicas multivariante se capturan las relaciones combinadas con el resto de las variables. Sin embargo, hemos obtenido coeficientes contradictorios y estadísticamente significativos, como en el caso de la variable CO_2, que ahora mostraría un impacto negativo, reduciendo así el número de incidencias generadas. Esto contradiría varios estudios previos analizados [13].

Revisaremos las métricas para evaluar que de bueno es nuestro modelo de regresión:

MSE: 0,024

RMSE: 0,157

MAE: 0,113

R^2: 0,063

Tras evaluar al modelo obtenemos un coeficiente de determinación (0,063) bajo lo cual está indicando que el modelo no captura correctamente la variabilidad de los datos y los errores son relativamente altos lo que sugiere que este modelo tiene limitaciones significativas en su capacidad predictiva. Se han realizado varias pruebas añadiendo más variables y que no estén altamente correlacionadas entre sí, como por ejemplo las diferentes partículas de aire o la variable de luminosidad con el índice ultravioleta.

Tabla 9. Coeficientes del conjunto más grande de variables

VARIABLE	COEFICIENTES	P-VALOR
CO_2	-1.591E-05	0,0
HUM	-0,004	0,0
TEMP	-0,009	0,0
STD_DENSITY	-0,572	0,0
PRES	0,001	0,004
NOISE	0,004	0,0
INDEX_UV	-0,004	0,484
PM25	-0,002	0,0
LOCATION	0,009	0,05
TIPO	-0,014	0,005

Tras analizar los resultados de la (Tabla 9), obtenemos que las variables ambientales, combinadas con las arquitectónicas, tienen un impacto mínimo sobre las incidencias generadas semanalmente, aunque son estadísticamente significativas excepto la variable que captura el índice ultravioleta. Podemos observar, que las variables localización y tipo tienen un impacto dependiendo de donde estén, y reduciendo el número de incidencias generadas en los pueblos interiores. Aun así, hemos revisado las siguientes métricas para evaluar el modelo:

MSE: 0,025

RMSE: 0,158

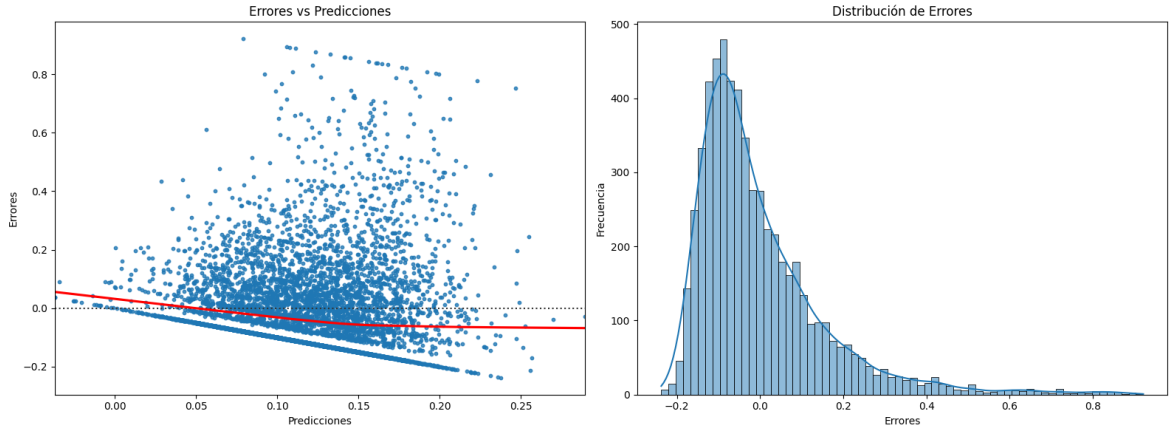
MAE: 0,113

R^2: 0,069

Obtenemos unos datos muy similares a los primeros y eso podría ser debido a que el modelo no es adecuado para nuestro conjunto de datos. Aun así, se ha profundizado más este análisis para poder descartar este modelo. Se han revisado los errores para valorar si tenemos una relación lineal entre el conjunto de variables independientes y el número de incidencias generadas. El objetivo de este pequeño estudio sobre los errores era obtener unos resultados

que nos facilite el tomar una decisión y poder descartar un modelo lineal y adentrarnos a un modelo que capture relaciones no lineales [14]. Tras varias pruebas, obtenemos las siguientes gráficas que muestran una distribución de los errores y un patrón claro de estos:

Figura 9. Plot sobre los errores del modelo de regresión lineal



En la (Figura 9) podemos observar dos gráficas que analizan los errores. La primera gráfica nos ayudará a obtener si los errores tienen algún tipo de patrón en función de los valores obtenidos por el modelo. Un patrón no aleatorio indicaría que la relación entre las variables no es capturada completamente por un modelo lineal, lo cual, es lo que nos está pasando. Obtenemos que hay un patrón que mientras aumentamos el valor de las predicciones aumenta la dispersión de los errores y esto estaría indicando heteroscedasticidad. Además, en la línea roja se nota cierta curvatura hacia abajo en los extremos de las predicciones, lo que podría sugerir una relación no lineal entre las variables independientes y la variable dependiente.

La segunda gráfica nos ha permitido evaluar la normalidad de los errores, lo cual es una de las condiciones clave en una regresión lineal. Si los errores no siguen una distribución normal, podría sugerir que el modelo no es adecuado para nuestros datos o que es necesario considerar una transformación de estos. En nuestro caso, observamos que no hay una distribución normal, ya que los errores no presentan simetría; la distribución está sesgada hacia la derecha, lo que indica que hay más errores positivos grandes que errores cercanos a cero.

Dado estos resultados iniciales en el análisis multivariante, podemos descartar el modelo de regresión lineal para nuestros datos y considerar modelos que capturen relaciones no lineales.

5.1.2 Modelos Machine Learning

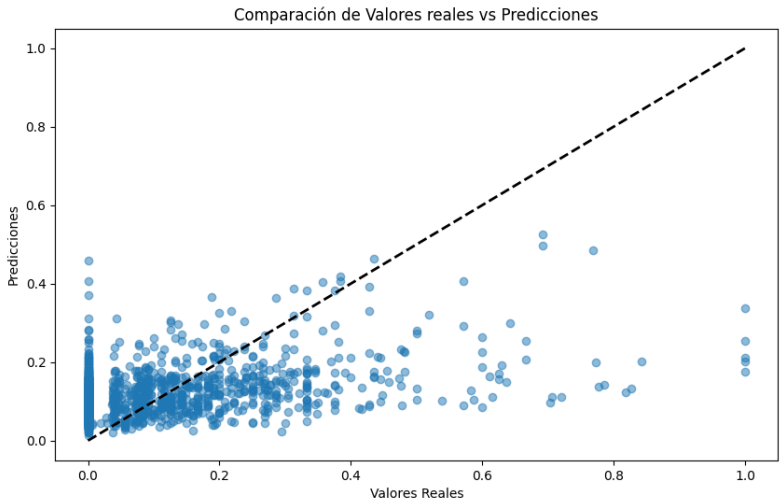
Se han seleccionado varios modelos no lineales para evaluarlos con las métricas utilizadas previamente. El objetivo era realizar una breve comparación entre estos modelos para poder elegir el que mejor se ajuste a nuestros datos. Se han utilizado las herramientas de búsqueda de parámetros para la optimización del rendimiento de los modelos y hemos obtenido estos resultados:

Tabla 10. Métricas de los múltiples modelos de regresión no lineales

MODELO	MSE	RMSE	R^2	MAE
REGRESIÓN POLINÓMICA	0,024	0,154	0,113	0,112
ÁRBOL DE DECISIÓN	0,036	0,189	-0,336	0,125
RANDOM FOREST	0,019	0,139	0,166	0,105
SVR	0,023	0,15	0,158	0,109
MLP	0,024	0,153	0,067	0,112

Las métricas de la (Tabla 10) sugieren que el modelo de *Random Forest* es el más robusto y eficaz para predecir el número de incidencias generadas a partir de las variables independientes, en comparación con el resto de los modelos. Ha obtenido el menor número de errores de predicción, explica hasta un 16% de la variabilidad en el modelo y, además, presenta el promedio de errores absolutos más bajo, lo que sugiere que el *Random Forest* es el modelo más adecuado para este conjunto de datos, logrando un equilibrio entre precisión y capacidad de generalización. Aunque el *SVR* y el modelo de *Regresión Polinómica* tienen un rendimiento similar y razonable, hemos optado por el modelo mencionado anteriormente. Teniendo el modelo seleccionado, se ha explorado posibles mejoras del modelo y visualizaciones de los resultados para poder obtener conclusiones. Inicialmente, hemos realizado una gráfica para ver una comparativa entre predicciones y valores reales. Esto nos ha dado una idea sobre los puntos débiles que tiene el modelo.

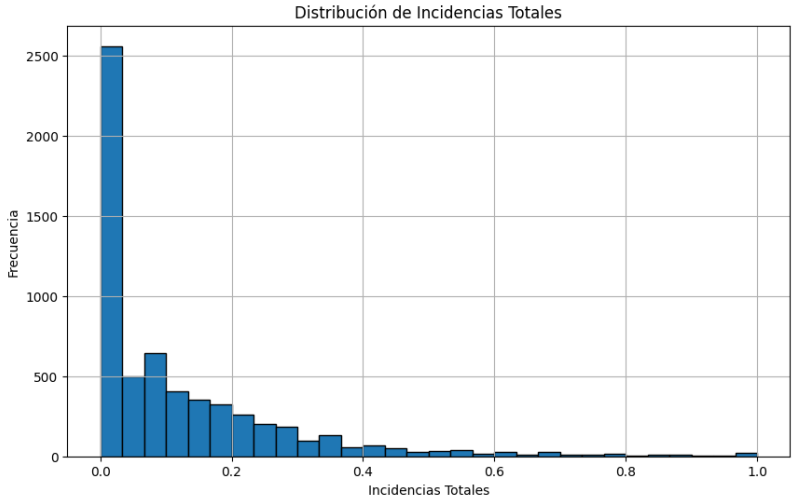
Figura 10. Gráfica de dispersión entre valores reales y predicciones



En un modelo perfecto, los puntos estarían situados a lo largo de la línea diagonal discontinua, lo cual no ocurre en este caso. Podemos observar lo alejadas que están las predicciones del valor ideal, lo que nos ha permitido evaluar y sacar conclusiones. Se aprecia que la mayoría de los puntos están concentrados en la esquina inferior izquierda, donde se encuentran los valores pequeños, lo que indica poca dispersión en esta región. Sin embargo, a medida que los valores aumentan, las predicciones se dispersan más.

Otro detalle interesante es que el modelo apenas puede capturar incidencias superiores al 50%, es decir, predice valores más bajos cuando en realidad no lo son. Esto podría deberse a que el modelo ha sido entrenado con una gran cantidad de datos con valores pequeños, lo cual es motivo de preocupación. Tras realizar un análisis simple de la distribución de incidencias totales, podemos observar cómo están distribuidas y obtuvimos la siguiente gráfica:

Figura 11. Gráfica de la distribución de las incidencias



La (Figura 11) nos sugiere qué el modelo tiene dificultades para predecir valores superiores al 50%, lo cual se debe a la escasez de muestras en ese rango. Podemos observar que apenas contamos con datos que representen grandes incidencias, lo que afecta considerablemente el rendimiento del modelo. Además, el modelo fue entrenado con la técnica del 80/20, que selecciona un 80% del conjunto de datos para entrenar y aprender sobre la variable dependiente, mientras que el 20% restante se utiliza para evaluar la capacidad del modelo para predecir en función del entrenamiento recibido. A partir de aquí, hemos valorado y seleccionado varias técnicas que podrían ser realmente útiles:

1. *Sobremuestreo / Submuestreo*: Podemos aumentar el número de muestras con valores altos mediante técnicas como SMOTE (Synthetic Minority Over-sampling Technique) que realizaría pequeñas modificaciones a los datos que tenemos para considerarlo como nuevos datos. La otra técnica se basa en reducir el número de muestras de la clase dominante, que en este caso corresponde a las incidencias bajas.

2. Pesos de clase: Modificar el modelo para que preste más atención a los ejemplos con valores altos.

3. *StratifiedKFold:* Es una variación del método K-Fold Cross Validation en el que nos asegura que cada subconjunto tenga aproximadamente la misma proporción de cada clase de la variable incidencias.

4. *Modelos ensamblados*: Un modelo ensamblado combina varias técnicas o modelos para manejar mejor la variabilidad y el desequilibrio en los datos.

Esto nos ha llevado a considerar varias decisiones. La primera opción, aumentar o disminuir la cantidad de muestras de incidencias, la descartamos debido a su impacto significativo negativo que tiene. Aumentar la cantidad de muestras sería similar a falsificar la información, lo cual no es deseable y además de que incrementaría los costos de computación. Por otro lado, reducir la clase dominante sobre la incidencia podría eliminar información realmente útil.

Se han realizado varias pruebas combinando la técnica *Stratified K-Fold* con la ponderación de pesos en valores altos. Las métricas apenas han mejorado, pero podrían servir. Se ha obtenido un ligero incremento en la cantidad de incidencias altas reportadas. Finalmente, nos quedaría valorar la opción de utilizar modelos ensamblados para ver si podemos optimizar nuestro modelo.

Random Forest es una implementación de un modelo ensamble de tipo *Bagging* ya que se basa en entrenar múltiples árboles de decisión en diferentes subconjuntos, así que nos hemos basado principalmente en los modelos *Stacking* y *Boosting*. Inicialmente, hemos creado un modelo *boosting* y comprobado el rendimiento evaluando las siguientes métricas:

Tabla 11. Métricas de los múltiples modelos de regresión no lineales mejorados

MODELO	MSE	RMSE	R^2	MAE
RANDOM FOREST	0,019	0,139	0,166	0,105
XGBREGRESSOR	0,021	0,144	0,175	0,102
LGBMREGRESSOR	0,021	0,146	0,167	0,103
R.F + XGBREGRESSOR	0,02	0,143	0,188	0,101
R.F + LGBMREGRESSOR	0,021	0,145	0,168	0,102

Tras los resultados obtenidos en la (Tabla 11), concluimos que aplicar las técnicas mencionadas anteriormente, tanto en el proceso de entrenamiento como en la combinación de modelos para crear un modelo final capaz de predecir, ha influido de manera positiva, aunque de forma ligera. En caso de decidir utilizar esta mejora, deberíamos considerar si el tiempo adicional de computación y entrenamiento justifica el ligero incremento en el rendimiento.

5.2 Modelo de Clasificación

El modelo de clasificación nos permite predecir una clase o una etiqueta a partir de unos datos de entradas. Para desarrollar este tipo de modelo, hemos creado una categoría para la variable **Incidencias totales** con el fin de clasificar las incidencias. Inicialmente, se ha creado un modelo de clasificación binario, ya que se pueden identificar dos clases: una que indique si hay alguna incidencia reportada y otra que indique si no la hay, es decir, si el número de incidencias calculado es mayor que 0. Posteriormente, se intentó mejorar ese modelo y desarrollar un modelo multiclase capaz de identificar diferentes tipos de incidencias. Hemos trabajado en este modelo teniendo en cuenta el desbalanceo de muestras en la variable dependiente que detectamos en el estudio (**Modelo de Regresión**).

5.2.1 Modelo de Clasificación Binario

Para crear este modelo, ha sido necesario definir una categoría que indique si hay incidencia o no. Una vez creada esta categoría, hemos revisado la distribución de incidencias, y en este caso, el conjunto de datos no ha presentado un desbalanceo tan marcado como en el anterior estudio, ya que el 59% de los reportes indican incidencia, mientras que el 41% no la indican. Después de realizar varias pruebas para testear y evaluar el rendimiento de los diferentes modelos, hemos obtenido las siguientes métricas:

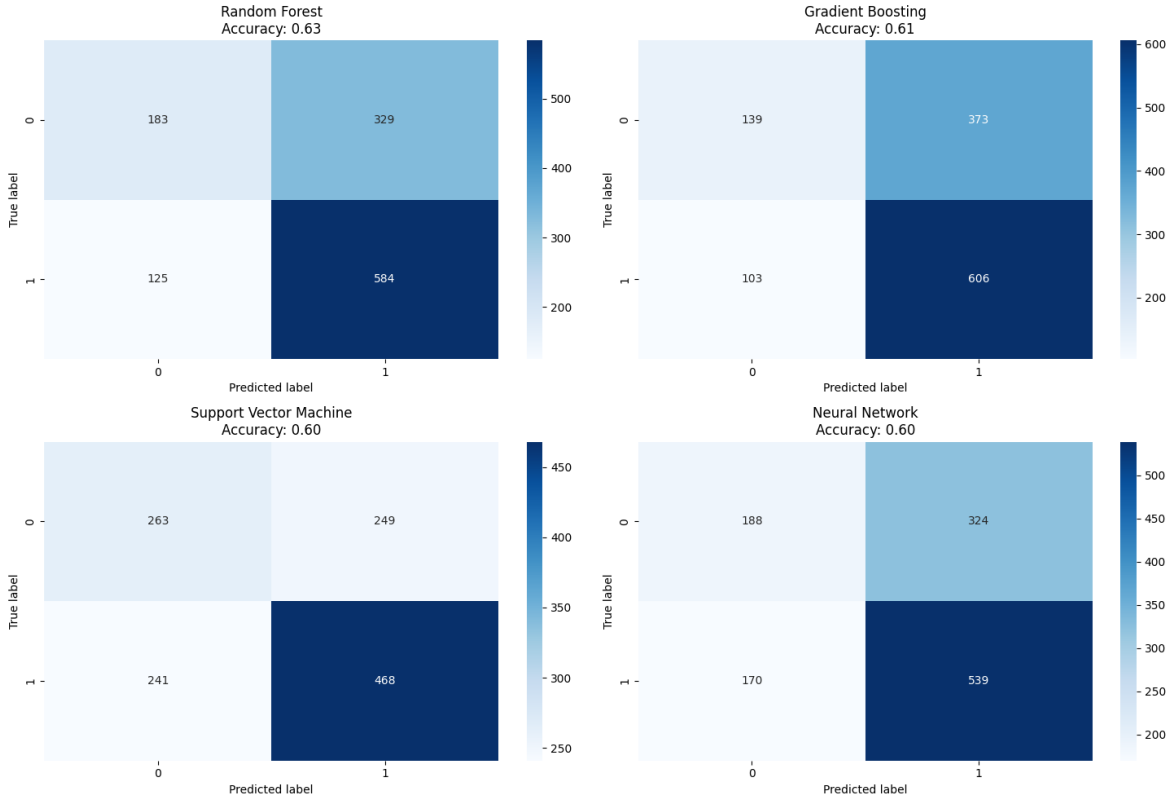
Tabla 12. Métricas de los modelos de clasificación binaria

MODELO	ACCURACY	PRECISION	RECALL	F1
DECISION TREE	0,541	0,607	0,597	0,602
RANDOM FOREST	0,628	0,64	0,824	0,721
GRADIENT BOOSTING	0,61	0,619	0,855	0,718
SVM	0,599	0,653	0,66	0,656
MLP	0,595	0,626	0,753	0,684

Tras un primer contacto con las métricas de evaluación de los modelos que hemos obtenido en la (Tabla 12), hemos observado que el modelo *Decision Tree* es el que peor resultado muestra, casi comparándolo con un modelo basando en la aleatoriedad (50/50). Por ello, se ha descartado directamente para la siguiente métrica. *Gradient Boosting* y *Random Forest* son unos modelos que han obtenido un rendimiento ligeramente superior a los modelos *SVM* y la red neuronal *MLP - Classifier*.

Aun así, se ha generado un plot de matrices de confusión para evaluar cuál modelo nos resulta más adecuado.

Figura 12. Plot de matrices de confusión de los diferentes modelos binarios

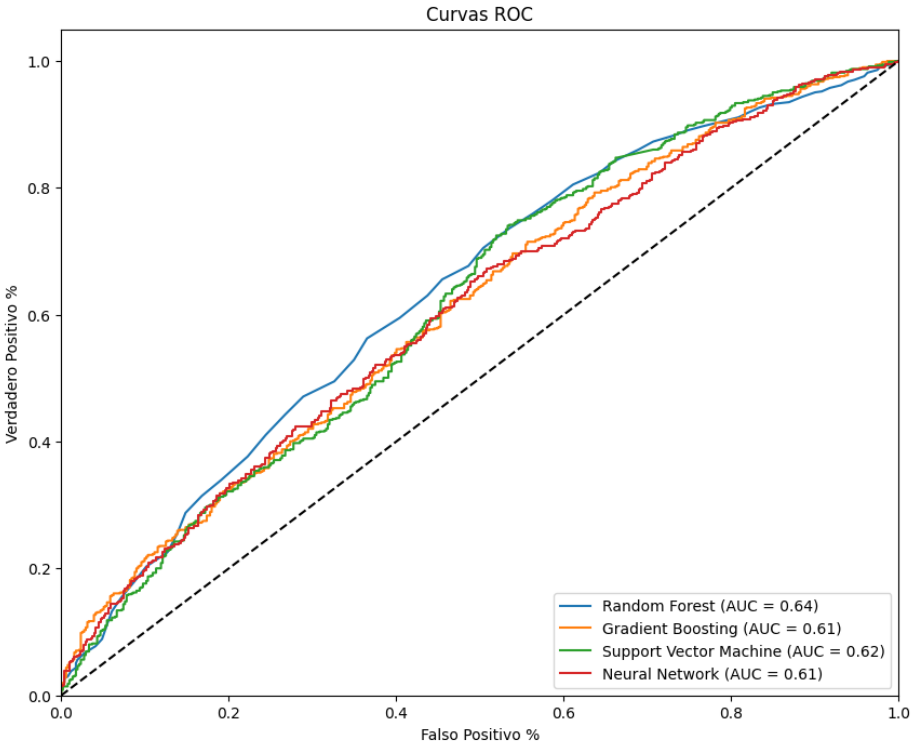


Tras interpretar el conjunto de matrices de confusión de la (Figura 12), se observa una gran cantidad de positivos, tanto verdaderos como falsos, y una cantidad reducida de negativos. Es importante tener en cuenta que cualquier falso resultado puede impactar negativamente en nuestra evaluación final. Si queremos crear un modelo que alerte sobre las incidencias generadas en la semana para implementar un sistema de prevención de la propagación de virus respiratorios, un falso negativo nos penalizaría más que un falso positivo. Esto se debe a que un falso negativo implicaría ignorar una incidencia cuando realmente existe, mientras que un falso positivo simplemente activaría el sistema de prevención sin generar problemas prácticos.

Observamos que el modelo que mejor se ajusta a estos criterios es el *Random Forest*, lo cual coincide con los resultados de las métricas realizadas. *Gradient Boosting* muestra resultados similares con el modelo ganador con proporciones similares de falsos negativos. La red neuronal y SVM son los que muestran el peor rendimiento, con una gran cantidad de falsos negativos, igualando casi la cantidad de verdaderos negativos.

Finalmente, se han utilizado las métricas ROC y AUC para comparar el rendimiento de los modelos con un modelo basado en la aleatoriedad, es decir, un modelo con un 50% de probabilidad.

Figura 13. Gráfica de curvas Roc de los diferentes modelos de clasificación



La (Figura 13) nos muestra las diferentes curvas ROC que han generado los diferentes modelos. La línea discontinua negra corresponde a un AUC de 0.5, lo que indica un sistema completamente aleatorio. Cuanto más alejada esté la línea que representa el AUC del modelo de esa línea discontinua, mejor será el modelo en términos de capacidad para distinguir entre clases. Inicialmente, observamos que todas las líneas de los modelos se alejan de la línea negra, lo cual es positivo, pero no lo suficiente como para considerar que se trata de un modelo de gran precisión en la distinción entre clases. El modelo *Random Forest* sigue siendo ligeramente superior al resto, por lo que podemos confirmar que será nuestro modelo principal.

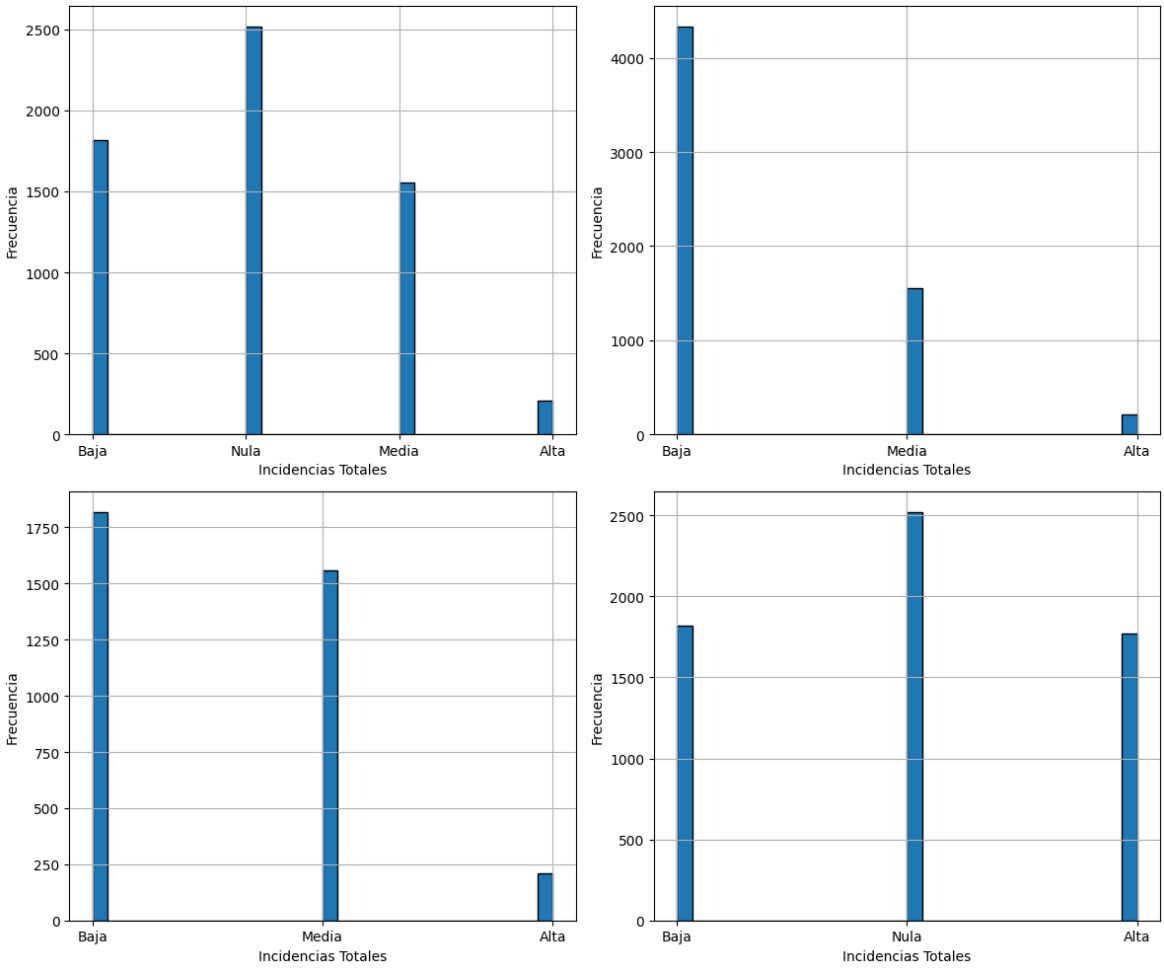
Realizamos algunos cambios en el proceso de aprendizaje y añadimos técnicas adicionales para mejorar el rendimiento general de los otros modelos, comparándolos como hicimos en el estudio del modelo de regresión. Sin embargo, no se han incorporado en este trabajo, ya que no se obtuvo una mejora considerable, excepto en el caso de la red neuronal, que aún no supera al modelo *Random Forest*. Por ello, decidimos no replicar exactamente el mismo procedimiento para un único cambio.

5.2.2 Modelo de Clasificación Multiclase

El objetivo de crear este modelo clasificador es poder predecir qué tipo de incidencias estamos teniendo, para así sugerir un aumento de las prevenciones en las aulas y reducir el riesgo de contagio. Teniendo en cuenta el problema que encontramos con el desbalanceo de incidencias que podemos observar en la (Figura 11. Gráfica de la distribución de las incidenciasFigura 11), con una gran cantidad de incidencias pequeñas. Es crucial seleccionar un umbral adecuado entre las diferentes clases para obtener una distribución lo más equilibrada posible, evitando que los modelos presenten problemas con la clase dominante

en el conjunto de entrenamiento. Se realizaron varias pruebas con diferentes umbrales definidos, y hemos obtenido el siguiente gráfico:

Figura 14. Distribución de las diferentes clases de incidencias con diferentes umbrales



Tal y como podemos observar en la (Figura 14), observamos que la clase 'Alta' presenta un problema en todas las distribuciones, excepto en la última, lo cual representa un desafío para el modelo. Por otro lado, la clase 'Nula' contiene una gran cantidad de datos, lo que la convierte en la clase dominante. En la segunda figura, decidimos unir las clases 'Nula' y 'Baja', pero esta opción fue descartada de inmediato debido a la gran diferencia en la cantidad de datos. En la tercera opción, se eliminó la clase 'Nula', pero esto no resolvió el problema de la clase 'Alta' y su escasa cantidad de información. Finalmente, la última gráfica muestra una modificación de los umbrales, donde se realizaron varios ajustes hasta obtener una distribución más equilibrada. Los umbrales que hemos utilizado finalmente para poder obtener una distribución más equilibrada son los siguientes:

- **Nula**: x = 0
- **Baja**: 0 < x <= 0.15 (incidencia menor o igual a un 15%)
- **Alta**: x < 0.15 (incidencia superior al 15%)

Con estos umbrales, clasificamos las clases de tal manera que los datos están distribuidos de la forma más equilibrada posible, manteniendo la coherencia en la definición

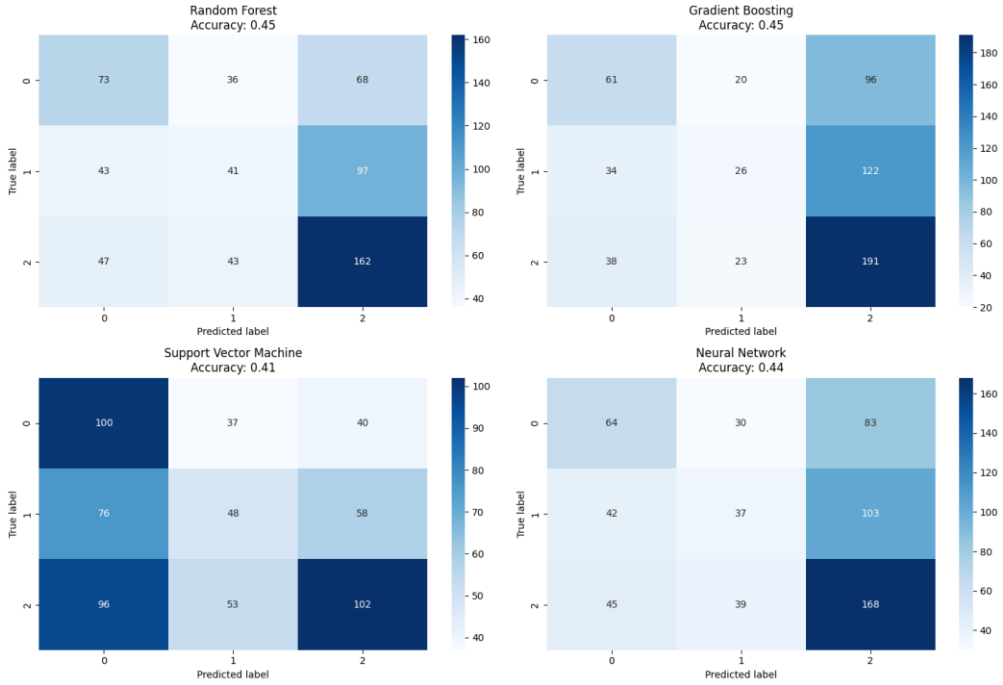
de las clases. A partir de aquí, evaluamos los modelos anteriores para determinar su rendimiento en un contexto multiclase. Se realizaron varias pruebas con diferentes tipos de entrenamiento, y finalmente, presentamos las métricas obtenidas utilizando el método de *Stratified K-Fold*, ya que ofreció los mejores resultados.

Tabla 13. Métricas de los modelos de clasificación multiclase

MODELO	ACCURACY	PRECISION	RECALL	F1
DECISION TREE	0,373	0,373	0,373	0,373
RANDOM FOREST	0,451	0,436	0,451	0,436
GRADIENT BOOSTING	0,452	0,434	0,452	0,41
SVM	0,411	0,423	0,411	0,406
MLP	0,442	0,426	0,442	0,422

Tras revisar los resultados de las diferentes métricas de la (Tabla 13), podemos sugerir que el modelo de *Decision Tree* sigue generando el rendimiento más bajo en comparación con el resto de los modelos, mientras que el *Random Forest* es el modelo con el mejor rendimiento. Sin embargo, los cuatro modelos presentan un rendimiento muy similar entre sí. Siguiendo la misma estructura del estudio anterior, hemos realizado una matriz de confusión para los cuatro modelos que han mostrado un rendimiento similar, con el fin de comparar las predicciones de las diferentes clases. Esto nos ha permitido identificar los problemas que enfrentan los modelos con las diferentes clases y determinar qué clase tiene más peso en cada modelo. En este caso, la matriz será de 3x3, generando así: un valor de acierto para la clase correspondiente, un valor erróneo para una clase, y otro valor erróneo para la segunda clase.

Figura 15. Plot de matrices de confusión de los diferentes modelos multiclase



Interpretando el conjunto de matrices de confusión de la (Figura 15), podemos observar que los modelos no están proporcionando una precisión alta, ya que las precisiones están en un rango de 0,41 a 0,45, superando solo ligeramente a un modelo basado en el azar (0,33). Las clases 'Nula' y 'Alta' tienen un gran número de aciertos, pero también una gran cantidad de errores en la mayoría de los modelos. La clase 'Baja' presenta problemas en todas las predicciones de todos los modelos, lo que sugiere que estos tienen dificultades para capturar el efecto de nuestro conjunto de variables independientes en esta clase.

Aunque estos modelos han superado ligeramente el rendimiento de un modelo basado en la aleatoriedad, no ofrecen una precisión alta en la predicción de las diferentes clases de incidencias. Se intentaron realizar mejoras para aumentar el rendimiento de los diferentes modelos, pero sin éxito o con un incremento marginal en el rendimiento que a su vez aumentó la complejidad del modelo y el tiempo computacional.

Finalmente, estas predicciones no son completamente fiables para poder realizar un sistema prevenciones basándonos en las predicciones, pero, sin embargo, los resultados obtenidos son y podrán ser útiles para futuras investigación o expansiones de este estudio.

5.3 Discusión

El análisis de los resultados que hemos obtenido en este estudio analizando el impacto de un conjunto de variables en la generación de incidencias revela varias observaciones importantes que destacan tanto las limitaciones de los modelos utilizados como posibles implicaciones para futuros estudios.

Inicialmente, utilizamos el modelo de regresión lineal, una herramienta bastante útil y común debido a su simplicidad y facilidad de implementación. Sin embargo, no obtuvimos buenos resultados. Los bajos valores del coeficiente de determinación y las altas métricas de error sugerían que el modelo no era capaz de capturar correctamente la relación entre las

variables independientes y las incidencias. Estos resultados, junto con un análisis de los errores que indicaba patrones y distribuciones asimétricas, sugerían la existencia de relaciones no lineales, lo cual sería la principal razón del bajo rendimiento de este modelo.

Incluyendo modelos no lineales al análisis se observaron un incremento al rendimiento, especialmente con el modelo *Random Forest*, que mostraba una mayor capacidad de predicción del número de incidencias generadas semanalmente. Sin embargo, el modelo seguía siendo limitado ya que el coeficiente de determinación apenas podía capturar un 16% de variabilidad. Este resultado subraya la complejidad de las relaciones entre las variables del dataset y sugiere que, aunque *Random Forest* es robusto, puede no ser suficiente para capturar toda la variabilidad inherente a los datos.

Se encontraron varios problemas durante el estudio, especialmente el desbalanceo en la distribución de incidencias. En un conjunto de datos reales, es muy difícil lograr una distribución equitativa. La escasez de incidencias altas generaba problemas en todos los modelos en esos rangos, lo que conducía a una subestimación de los valores más altos. Se exploraron posibles soluciones, como la manipulación de datos, ponderación de pesos o técnicas de validación cruzada para mejorar el rendimiento. Sin embargo, cualquier forma de manipulación de datos, ya fuera generando o reduciendo muestras, se descartó por completo, y el resto de las técnicas no lograron mejoras significativas, sacrificando tiempo de computación y aumentando la complejidad del modelo.

En cuanto a los resultados de los modelos de clasificación, el modelo Random Forest también destacaba como el más eficaz en comparación al resto, aun así, obteniendo rendimientos muy similares entre ellos. Sin embargo, los modelos de clasificación binaria y multiclase mejoraban ligeramente a un modelo basado en aleatoriedad, teniendo que la clasificación binaria era relativamente mejor que la clasificación multiclase. La proporción de falsos negativos es especialmente problemática, dado que, si quisiéramos crear un sistema de prevenciones o sugerencias para evitar la propagación del virus dentro de las aulas, un falso negativo podría tener consecuencias significativas al no detectar un problema cuando realmente existe.

Se han considerado y aplicado diferentes mejoras y técnicas durante los análisis para mejorar los modelos, pero los incrementos en el rendimiento fueron mínimos, a costa de aumentar la complejidad del modelo y el tiempo de computación. Esto plantea la cuestión de si estos esfuerzos adicionales en términos de tiempo y recursos computacionales justifican las mejoras obtenidas.

En conclusión, este estudio ha permitido explorar y justificar las capacidades y limitaciones de varios modelos predictivos aplicados a un conjunto de datos complejos y reales. Los resultados eran, en cierto modo, previsibles, dado que las incidencias pueden originarse fuera de las aulas y ser causadas por factores externos. Es muy difícil predecir este tipo de incidencias, ya que, aunque los niños pasen gran parte del día en las aulas, pueden participar en actividades extraescolares o incluso contagiarse en su propia casa.

6 Conclusiones

Este proyecto ha representado un gran desafío, tanto a nivel técnico como personal, ya que partí sin conocimientos previos sobre análisis de datos ni técnicas de *Machine Learning*. Mi principal motivación era aprender uno de los ámbitos más populares que hay actualmente en la informática ya que el plan de estudio que yo estaba cursando no tocaba estos temas. Por ello, he tenido que aprender de manera autodidacta, investigando diferentes artículos científicos para entrar en contexto con la problemática de los datos ambientales y experimentando las partes prácticas mediante un enfoque de prueba y error.

Una vez preparamos los datos para el análisis, realizamos un primer estudio sobre las correlaciones entre las variables principales. Se realizaron varias pruebas, y finalmente se decidió combinar las correlaciones Pearson y Spearman para obtener una visión más robusta y completa entre las variables, al no tener claro que tipo de relación tenían. Con diferentes herramientas de visualización pudimos visualizar e interpretar los resultados y finalmente, tomar estos resultados como base a un modelo multivariante.

El análisis realizado nos ha permitido identificar las relaciones y patrones entre las variables del conjunto con el número de incidencias semanales, lo que ha sido un reto para mí, debido a la inexperiencia y los problemas como la complejidad y el desbalanceo de datos. A pesar de las limitaciones observadas, como el bajo rendimiento del modelo de regresión lineal y las dificultades para predecir valores altos de incidencias, se lograron algunos avances importantes, especialmente con la implementación de modelos no lineales como *Random Forest*. Aunque se alcanzaron algunas mejoras a través del uso de técnicas como *Stratified K-Fold* y la ponderación de pesos, los incrementos en el rendimiento fueron mínimos, lo que indicaría que aún faltaría detalles relevantes o profundizar todavía más este campo para obtener unos resultados óptimos.

Los resultados obtenidos podrían considerarse lógicos. Las variables que disponemos para analizar pueden tener un peso significativo en la facilitación de la transmisión de virus respiratorios, pero no lo explican todo. Hay que tener en cuenta que, aunque los alumnos pasan gran parte del día en las aulas, también existe la posibilidad de que se contagien de otras maneras: en casa con sus familiares, en actividades extraescolares, o incluso durante el fin de semana. Por lo tanto, hemos podido validar y justificar que los datos que estamos recibiendo pueden tener un impacto moderado en la propagación de los virus respiratorios, facilitando así su transmisión en las aulas cerradas. El trabajo realizado y los resultados obtenidos podrán utilizarse como una base para futuras investigaciones sobre la propagación de los virus respiratorios.

Personalmente, he disfrutado mucho realizando este trabajo. Me ha permitido darme cuenta de que quiero especializarme en el ámbito del análisis de datos y seguir formándome en este campo. He estado aprendiendo sobre este campo a través de artículos científicos, clases grabadas y la propia documentación de las herramientas que he utilizado. Estoy muy satisfecho con el trabajo realizado, y no solo por los resultados, sino también por haber aprendido nuevas metodologías y herramientas que me serán útiles en futuros proyectos e investigaciones.

Por último, me gustaría agradecer la oportunidad que me ha brindado el proyecto ACTUA por acoger a un alumno sin experiencia ni conocimientos previos para realizar un proyecto de análisis de datos con datos reales. También quiero agradecer a Agustí Solanas, investigador del proyecto, por su tiempo y sus consejos para guiar mi trabajo, y a Edgar Batista, técnico del grupo, por supervisar mi trabajo de manera constante, sugerir ideas y guiarme de tal manera que no me he sentido perdido en ningún momento.

7 Referencias

[1] E. Batista, O. Villanova, J. Rosell-Llompart, F. Huera-Huarte, A. Martínez-Ballesté y A. Solanas, «On the Deployment of Low-Cost Sensors to Enable Context-Aware Smart Classrooms,» de *International Conference on Applications in Electronics Pervading Industry, Environment and Society*, Genova, Italy, 2022.

[2] E. Batista, F. Huerta, A. Martínez, J. Rosell y A. Solanas, El projecte ACTUA: Investigant la transmissibilitat dels virus respiratoris a les aules, 2023.

[3] J. A. Rodrigo, «https://cienciadedatos.net,» Julio 2017. [En línea]. Available: https://cienciadedatos.net/documentos/17_mann%E2%80%93whitney_u_test.

[4] Y.-y. Song y Y. Lu, «Decision tree methods: applications for classification and prediction,» *Shanghai Archives of Psychiatry*, vol. 27, nº 2, pp. 130-135, 25 Apr 2015.

[5] S. J. Rigatti, «Random Forest,» *Journal of Insurance Medicine*, vol. 47, nº 1, pp. 31-39, 1 Jan 2017.

[6] D. A. Pisner y D. M. Schnyer, «Support vector machine,» de *Machine Learning: Methods and Applications to Brain Disorders*, Academic Press, 2020, pp. 101-121.

[7] H. Taud y J. Mas, «Multilayer Perceptron (MLP),» de *Geomatic Approaches for Modeling Land Change Scenarios*, Springer, 2017, pp. 451-455.

[8] S.-L. Developers, «https://scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/modules/ensemble.html.

[9] «https://es.wikipedia.org/,» [En línea]. Available: https://es.wikipedia.org/wiki/Optimizaci%C3%B3n_de_hiperpar%C3%A1metros.

[10] R. Díaz, «https://www.themachinelearners.com,» [En línea]. Available: https://www.themachinelearners.com/metricas-de-clasificacion/.

[11] J. I. Barrios, «https://www.juanbarrios.com,» Mayo 2022. [En línea]. Available: https://www.juanbarrios.com/evaluando-los-modelos-de-regresion/.

[12] Scikit-Learn, «scikit-learn.org,» [En línea]. Available: https://scikit-learn.org/stable/index.html#.

[13] U. C. d. Cuenca, «Determinantes de salud ambiental,» 2023.

[14] J. R. Abuín, «Regresión Lineal Múltiple,» 2007.