IA Aplicada a Ingeniería con Python
2023

Trabajo Práctico II

Introducción a la Inteligencia Artificial

Fernando Ezequiel Pose

Índice

1. Objetivos

Afianzar los conceptos introductorios respecto los clasificadores y regresores supervisados, mediante su implementación práctica. Identificar el desempeño de los modelos de acuerdo al tipo de dataset, interpretar las métricas y resultados obtenidos. Aplicar los conceptos de descriptores y definirlos.

Promover el interés respecto a la utilidad de los modelos de clasificación y regresión, paramétricos y no paramétricos, mediante aplicaciones prácticas reales.

2. Recomendaciones para la resolución del trabajo

Evitar copiar y/o modificar soluciones de pares (compañeros, sitios de Internet, etc.), en lugar de ello, esforzarse por elaborar una producción original propia a partir del análisis y reflexión de cada una de las consignas, teniendo a mano la teoría provista en clases, la bibliografía ofrecida y todo otro material complementario que juzgue necesario para enriquecer su producción.

Reflexionar sobre los conceptos o justificaciones que se ofrecen como solución a la consigna presentada. Es decir, pueden intercambiarse opiniones, debates o puestas en común respecto a un determinado punto, pero la producción entregada debe basarse en su concepción personal del marco teórico, experiencias generales e interpretación de las consignas.

3. Producción esperada para acreditar la actividad

Presentar un informe de estilo monográfico con formato libre en la tarea designada en el aula virtual del curso, incluyendo el contenido solicitado en cada punto de la guía. Realizarlo en tiempo y forma, dentro del plazo máximo de una semana desde la disponibilidad del presente documento.

Además del informe, adjuntar los códigos utilizados para llevar a cabo las experiencias. Los resultados presentados, deben poder ser replicables.

Priorizar la calidad por sobre la cantidad, cuidando la prolijidad general en la confección, incluyendo una portada debidamente identificatoria del trabajo.

4. Ejercicio 1

4.1. Enunciado

Se tiene un dataset con información referente a la lectura de gases realizada por una matriz de 6 sensores de bajo costo en instantes sucesivos de tiempo. A estas lecturas, se asocia la información de la actividad realizada en el recinto sensado.

El conjunto de sensores utilizados se puede agrupar en dos categorías principales:

- Sensores MQ (MQ2, MQ9, MQ135, MQ137, MQ138) que tienen gran sensibilidad, baja latencia y bajo costo; cada sensor puede responder a diferentes gases;
- Sensor analógico de gas CO2 (MG-811) que tiene una excelente sensibilidad al dióxido de carbono y apenas se ve afectado por la temperatura y la humedad del aire.

El conjunto de datos contiene 1845 muestras recolectadas que describen 4 situaciones objetivo:

1. **Situación normal** - Actividad: aire limpio, una persona que duerme, estudia o descansa. Muestras disponibles: 595.
2. **Preparación de comidas** - Actividades: cocinar carne o pasta, verduras fritas. Una o dos personas en la habitación, circulación de aire forzado. Muestras disponibles: 515.
3. **Presencia de humo** - Actividad: quemar papel y madera por un corto período de tiempo en una habitación con ventanas y puertas cerradas – Muestras disponibles: 195.
4. **Limpieza** - Actividad: uso de detergentes en aerosol y líquidos con amoníaco y/o alcohol. La circulación de aire forzado se puede activar o desactivar - Muestras disponibles: 540.

Cada muestra está compuesta por 7 valores; los primeros seis valores son las salidas de los sensores, mientras que el último es el índice de la acción que generó los valores adquiridos por los mismos. Las cuatro situaciones diferentes están asociadas con una composición del aire distinta, teniendo en cuenta que cualquier actividad produce sustancias químicas (respiración humana, exhalaciones de procesos metabólicos, liberación de volátiles por combustión y/o oxidación, evaporación de detergentes domésticos, etc.).

Los datos se encuentran en el archivo adjunto denominado “dataset_ADL_clasificacion.csv” (sección 6). A continuación, se presenta una porción de este dataset:

MQ2	MQ9	MQ135	MQ137	MQ138	MG-811	Situación
670	696	1252	1720	1321	2431	4
641	674	1156	1652	1410	2433	1
642	646	1159	1643	1455	2361	3
640	590	1105	1608	1459	2427	4
616	627	1192	1637	1466	2447	2

Cuadro 1: Fragmento del dataset, 1, calidad del aire.

En la columna *”Situación”* se encuentra codificada la clase que indica si se trata de una **actividad normal (1)**, **preparación de comidas (2)**, **presencia de humo (3)** o de **limpieza (4)**.

El objetivo es determinar si con estas observaciones es posible obtener un modelo que permita clasificar cada una de las 4 situaciones objetivo, a partir de los valores arrojados por los sensores.

Detallar y fundamentar cada aspecto de la solución propuesta, incluyendo los aspectos que crea conveniente y respondiendo como mínimo las siguientes premisas:

1. ¿Qué clasificador o meta clasificador se ajusta mejor a la solución buscada? ¿Por qué?
2. En base a los resultados obtenidos, ¿Sería posible utilizar el modelo para predecir el comportamiento de nuevas mediciones?

3. Evaluar aspectos relacionados al rendimiento en el proceso de entrenamiento y vincularlos con la relación costo/beneficio.
4. Ídem para los tiempos de inferencia o consulta, ¿Cuál es fundamental si pretendo utilizar un sistema en tiempo real con el modelo desarrollado?
5. De los resultados visualizados en la matriz de confusión, ¿Para todas las clases el modelo seleccionado presenta el mismo comportamiento?
6. ¿Existen problemas evidentes en este dataset? Si es así, ¿Cuáles son y cómo los solucionaría?
7. ¿Son necesarios todos los sensores para que el modelo pueda identificar de manera eficiente las clases de interés?

4.2. Resolución

4.2.1. Análisis exploratorio de los datos

En esta sección, se lleva a cabo el análisis exploratorio de los datos (EDA, *Exploratory Data Analysis*). La Tabla 2 presenta una concisa descripción del conjunto de datos, el cual consta de 1831 muestras (filas) y 7 variables (columnas) donde seis variables representan los diferentes tipos de sensor y uno la situación sobre la cual toman la medición. Si bien esta tabla no da información acerca de la presencia de muestras duplicadas, se destaca la ausencia de valores nulos, confirmando la integridad de los datos. Finalmente, se puede observar que todos los datos son del tipo entero.

Cuadro 2: Descripción de los datos.

Index: 1831 entries, 0 to 1844			
Data columns (total 7 columns):			
	Column	Non-Null Count	Dtype
0	MQ2	1831 non-null	int64
1	MQ9	1831 non-null	int64
2	MQ135	1831 non-null	int64
3	MQ137	1831 non-null	int64
4	MQ138	1831 non-null	int64
5	MG-811	1831 non-null	int64
6	Situacion	1831 non-null	int64

La Tabla 3 proporciona estadísticas adicionales (media, desviación estándar, valores mínimo y máximo, y los cuartiles 25, 50 y 75) de los datos. Si bien el sensor analógico (MG-811) exhibe valores elevados en comparación con los sensores digitales (sensores del tipo MQ), la magnitud de estos valores son similares, sugiriendo que inicialmente no sería necesario escalar o normalizar los datos.

Cuadro 3: Descripción estadística de los datos.

	MQ2	**MQ9**	**MQ135**	**MQ137**	**MQ138**	**MG-811**	**Situacion**
count	1831.00	1831.00	1831.00	1831.00	1831.00	1831.00	1831.00
mean	587.46	653.70	1166.34	1609.63	1302.43	2246.76	2.37
std	189.87	172.69	207.82	118.39	278.67	181.03	1.21
min	263.00	346.00	753.00	1323.00	773.00	1797.00	1.00
25 %	431.50	518.00	995.00	1509.00	1087.00	2137.00	1.00
50 %	551.00	623.00	1162.00	1611.00	1264.00	2265.00	2.00
75 %	712.50	745.50	1308.00	1692.00	1552.50	2372.00	4.00
max	1266.00	1388.00	1738.00	1926.00	1948.00	2703.00	4.00

De la Tabla 3 se puede observar un desvío estándar significativo en el sensor tipo MQ2 en comparación con su respectivo valor medio, lo que podría sugerir su exclusión a fin de mejorar el desempeño del modelo de clasificación.

La Figura 1 ilustra la matriz de correlación tanto de las variables de entrada como la salida. Si bien existe un valor alto de correlación entre algunos sensores (por ejemplo, MQ9 o MQ135 con MQ2), esto podría deberse al hecho de que todos los sensores comparten el mismo fabricante y miden variables similares en un espacio lo suficientemente reducido como para que los valores se aproximen entre sensores.

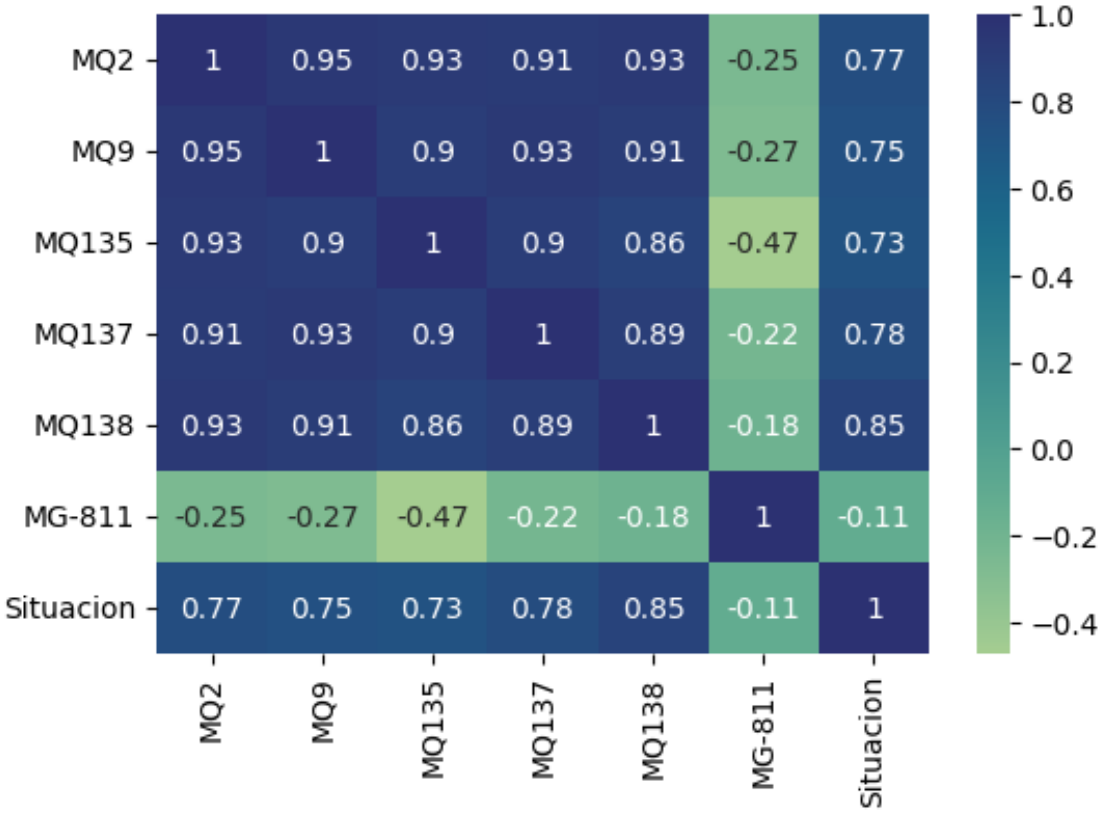


Figura 1: Matriz de correlación entre variables.

La Figura 2, muestra la relación entre las diferentes variables. Nuevamente, al igual que en 1, puede notarse una relación cuasi lineal entre los sensores tipo MQ135 y MQ2 sugiriendo que el sensor MQ2 podría no ser tenido en cuenta en el modelo de clasificación aunque, en una primera instancia, se evaluará un modelo de clasificación utilizando todos los sensores (analógicos y digitales) a fin de determinar si la inclusión o exclusión de algunos sensores mejora o empeora el rendimiento del modelo final.

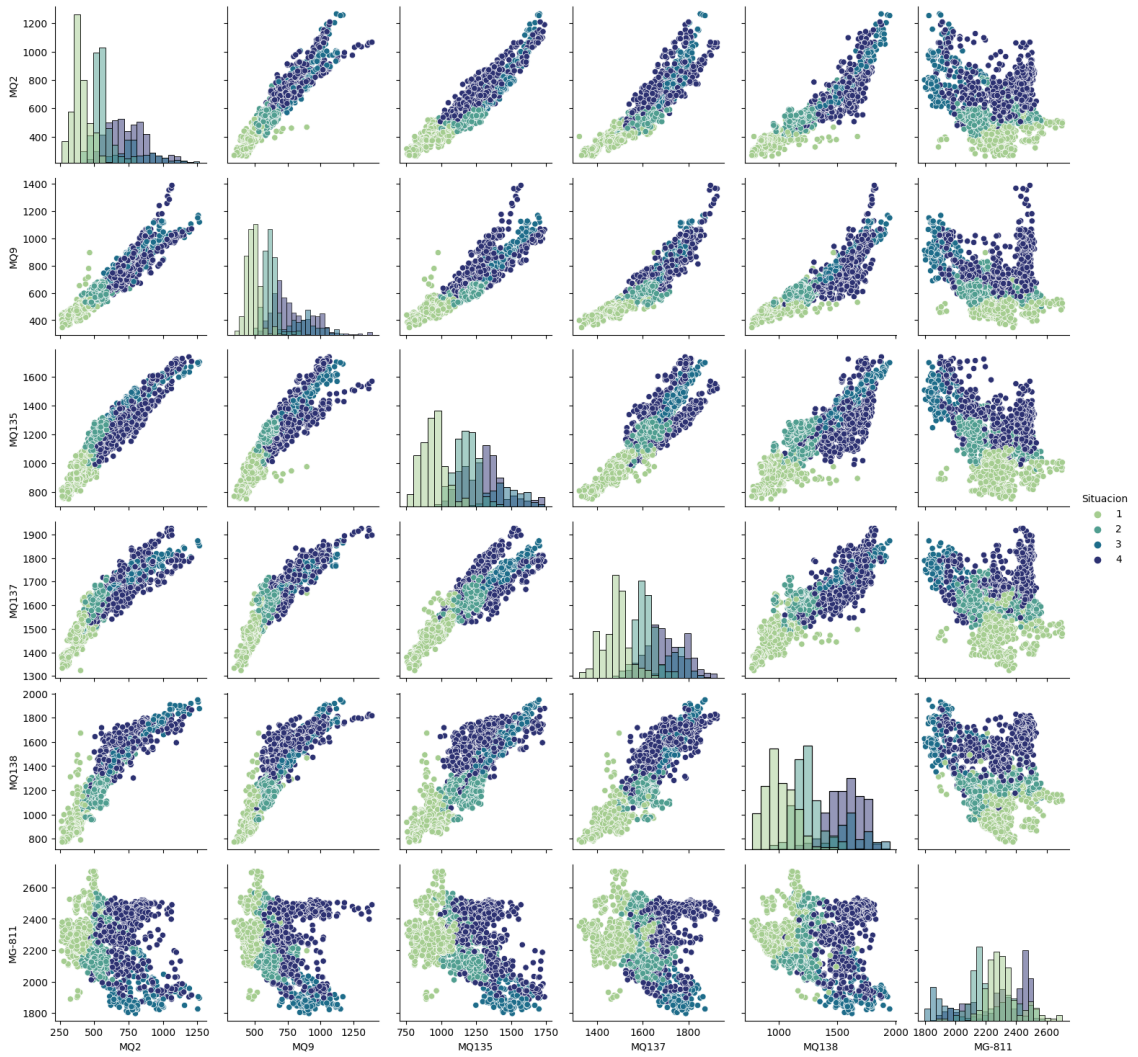


Figura 2: Relación entre las variables analizadas.

4.2.2. Preprocesamiento de los datos

En la fase inicial o etapa de preprocesamiento de datos, fueron identificadas y eliminadas 14 filas duplicadas. Como fue mencionado anteriormente en la sección 4.2.1, en este caso, no se llevó a cabo el escalado ni la normalización de los datos. Posteriormente, el conjunto de datos fue dividido en dos partes: el conjunto de entrenamiento, que comprendió el 70 % del total de datos, y el conjunto de prueba, que representó el 30 %. Es importante destacar que, antes de la división de datos, fue realizada una mezcla de las muestras en el dataset con el objetivo de romper cualquier estructura inherente producida por la carga de los mismos en la tabla, asegurando así que las muestras fueran lo más aleatorias posible. Este enfoque contribuye a garantizar la representatividad y generalización del modelo ante diferentes escenarios. Además, para todos los casos, siempre fue utilizado el valor 8 como semilla.

4.2.3. Primer análisis - LazyPredict

En búsqueda del mejor modelo, en primera instancia, se empleó la librería de código *LazyPredict* a fin de identificar los modelos candidatos más prometedores. Los resultados obtenidos se detallan en la Tabla 4.

Cuadro 4: Resultados de LazyPredict para los modelos candidatos.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LabelPropagation	0.97	0.98	None	0.97	0.11
LabelSpreading	0.97	0.98	None	0.97	0.15
KNeighborsClassifier	0.97	0.97	None	0.97	0.04
ExtraTreesClassifier	0.97	0.97	None	0.97	0.21
LGBMClassifier	0.95	0.95	None	0.95	0.38
BaggingClassifier	0.95	0.94	None	0.95	0.09
RandomForestClassifier	0.95	0.94	None	0.95	0.31
QuadraticDiscriminantAnalysis	0.92	0.93	None	0.92	0.01
SVC	0.93	0.92	None	0.93	0.06
DecisionTreeClassifier	0.92	0.92	None	0.92	0.02
ExtraTreeClassifier	0.91	0.9	None	0.91	0.01
LogisticRegression	0.89	0.88	None	0.89	0.04
LinearDiscriminantAnalysis	0.87	0.87	None	0.88	0.02
GaussianNB	0.86	0.85	None	0.86	0.02
NearestCentroid	0.84	0.84	None	0.84	0.02
CalibratedClassifierCV	0.84	0.84	None	0.84	0.35
Perceptron	0.83	0.84	None	0.83	0.02
LinearSVC	0.83	0.84	None	0.83	0.1
SGDClassifier	0.82	0.81	None	0.81	0.02
PassiveAggressiveClassifier	0.81	0.8	None	0.82	0.02
BernoulliNB	0.73	0.76	None	0.72	0.02
RidgeClassifier	0.78	0.75	None	0.77	0.02
RidgeClassifierCV	0.78	0.75	None	0.77	0.02
AdaBoostClassifier	0.65	0.68	None	0.6	0.26
DummyClassifier	0.26	0.25	None	0.1	0.01

Tras analizar los resultados, se decidió profundizar (en las siguientes secciones) en los siguientes modelo: **KNeighbors** (KNN), por liderar en *accuracy* junto con **LabelSpreading**, **LabelPropagation**, y **ExtraTrees**, el modelo **RandomForest** (RF) dado que se trata del segundo modelo con mayor *accuracy* y el modelo **Máquinas de vectores de soporte** (SVM) por ser el tercer modelo con mejor *accuracy*. Por último, si bien el modelo **eXtreme Gradient Boosting** (XGB) no se encuentra incluido en la biblioteca sklearn por lo que no fue previamente analizado de forma automática, el mismo trata de un modelo de ensamble basado en árboles de decisión el cual puede generar resultados robustos por lo que también será evaluado.

4.2.4. Desarrollo del modelo de predicción

En la primera etapa cada modelo de clasificación (ver modelos elegidos en 4.2.3), fue entrenado con los parámetros por defecto utilizando todos los sensores como variables de entrada. Para cada uno, fue generada una matriz de confusión (ver Figura 3) y se calculó las siguientes métricas: *accuracy*, *precision*, *recall* y *F1-Score* para las cuatro clases posibles.

Los resultados obtenidos para el modelo RF se presentan en la Tabla 5.

Cuadro 5: Reporte con las principales métricas de clasificación para el modelo RF

	precision	recall	f1-score	support
1	0.99	0.96	0.98	196
2	0.94	0.94	0.94	155
3	0.93	0.91	0.92	58
4		0.95	0.93	141
accuracy			0.95	550
macro avg	0.94	0.94	0.94	550
weighted avg	0.95	0.95	0.95	550

Puede notarse, para las cuatro clases, que todos los índices se sitúan por encima del 90 %. Además, mientras que la Clase 1 (situación normal) obtuvo los mejores resultados, la Clase 3 (presencia de humo) presentó los peores resultados lo cual era de esperable de suceder ya que el dataset utilizado se encuentra desbalanceado siendo la clase 1 la de mayor cantidad de muestras y por lo tanto mayor cantidad de muestras para entrenar y la 3 la de menor cantidad de muestras.

En la Tabla 6 se muestran los resultados obtenidos con el modelo KNN.

Cuadro 6: Reporte con las principales métricas de clasificación para el modelo KNN

	precision	recall	f1-score	support
1	0.98	0.98	0.98	196
2	0.96	0.97	0.96	155
3	0.93	0.98	0.96	58
4	0.98	0.94	0.96	141
accuracy			0.97	550
macro avg	0.96	0.97	0.97	550
weighted avg	0.97	0.97	0.97	550

Al igual que para el caso del modelo RF, la Clase 1 obtuvo los mejores resultados en términos de índices de desempeño. Además, en comparación con el modelo RF, la Clase 3 presentó mejoras en la sensibilidad del modelo. El accuracy (contemplando todas las cuatro clases) también aumentó en este modelo.

En la Tabla 7 se presentan Los resultados obtenidos con el modelo XGB.

Cuadro 7: Reporte con las principales métricas de clasificación para el modelo XGB

	precision	recall	f1-score	support
1	0.98	0.97	0.98	196
2	0.93	0.96	0.94	155
3	0.92	0.97	0.94	58
4	0.97	0.93	0.95	141
accuracy			0.96	550
macro avg	0.95	0.96	0.95	550
weighted avg	0.96	0.96	0.96	550

Este modelo presentó resultados similares al modelo RF y el modelo KNN. Respecto a la Clase 3, obtuvo un mejor rendimiento en términos de precisión y sensibilidad respecto al modelo RF, pero ligeramente menor al modelo KNN.

En la Tabla 8 se presentan los resultados obtenidos con el modelo SVM.

Cuadro 8: Reporte con las principales métricas de clasificación para el modelo SVM

	precision	recall	f1-score	support
1	0.95	0.94	0.95	196
2	0.86	0.85	0.85	155
3	0.83	0.86	0.85	58
4	0.83	0.86	0.85	141
accuracy			0.88	550
macro avg	0.87	0.88	0.87	550
weighted avg	0.88	0.88	0.88	550

La Figura 3 ilustra las matrices de confusión para los modelos RF (a), KNN (b), XGB (c), y SVM (d). Estas matrices proporcionan una representación visual de la clasificación de las muestras en las diferentes clases, permitiendo una evaluación detallada del desempeño de cada modelo en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada clase. Las matrices respaldan y complementan los resultados numéricos previamente discutidos en los reportes de clasificación de cada modelo.

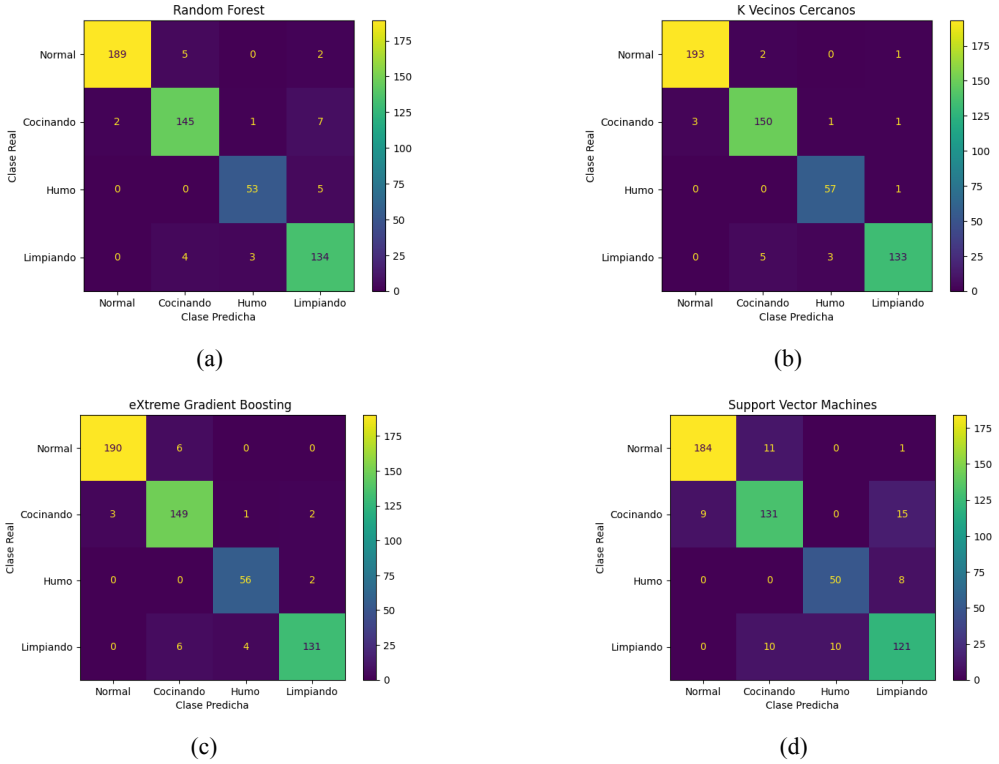


Figura 3: Maatriz confusión para los 4 modelos analizados.

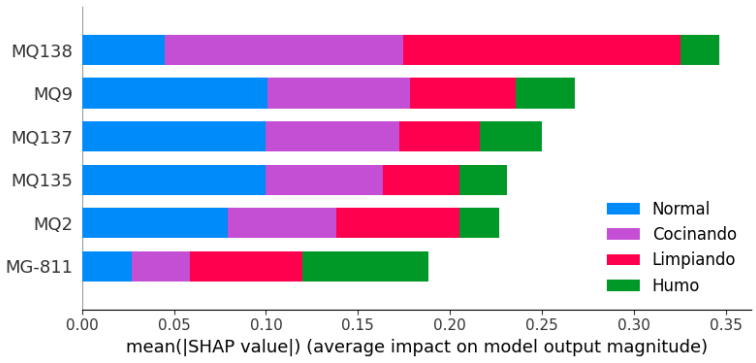
4.2.5. Desarrollo del modelo de predicción óptimo

En búsqueda de un modelo óptimo, en términos de menor cantidad de variables de entrada ó mejores rendimientos en la predicción, fue empleada la librería *SHAP* para evaluar las variables más relevantes durante el proceso de entrenamiento en modelos basados en árboles de decisión (modelos RF y XGB).

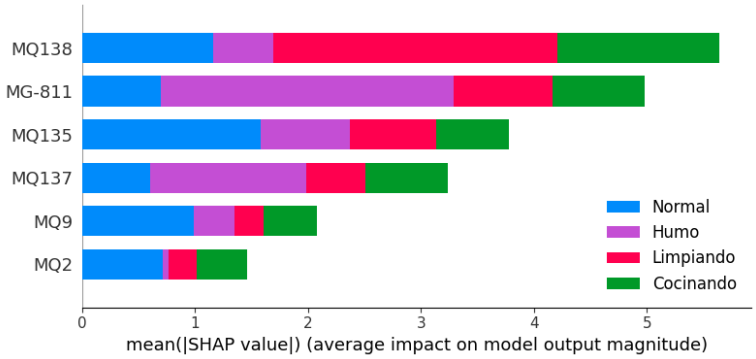
La Figura 4 muestra la importancia de las variables de entrada para los modelos RF y XGB. En el caso del modelo RF, se observó que los sensores MQ138 y MQ9 fueron los más relevantes en el proceso de entrenamiento, mientras que el sensor analógico MG-811 fue el de menor importancia. En el caso del modelo XGB, el sensor más significativo fue el MQ138, y el sensor MQ2 el menos importante. Cabe destacar que, a diferencia del modelo RF, el sensor analógico tuvo mayor relevancia en el modelo XGB.

Como parte de la optimización, a fin de obtener un modelo más eficiente y económico (uso de menos sensores), se exluyó el sensor MQ2. La exlusión de este sensor se refuerza a partir de las conclusiones obtenidas inicialmente en el análisis exploratorio de datos donde la correlación de este con otros sensores era significativa. En caunto al sensor MQ137, el mismo fue excluido debido a su baja importancia al momento de entrenar ambos modelos.

A fin de realizar una comparación consistente entre los modelos, se utilizaron los mismos conjuntos de datos de entrenamiento y prueba utilizados en la primera parte del ejercicio.



(a)



(b)

Figura 4: Análisis de importancia de las variables para los modelos RF y XGB.

La Tabla 9 presentan los resultados obtenidos utilizando un modelo de RF al excluir los sensores MQ2 y MQ137.

Cuadro 9: Reporte con las principales métricas de clasificación para el modelo RF

	precision	recall	f1-score	support
1	0.99	0.97	0.98	196
2	0.94	0.94	0.94	155
3	0.95	0.91	0.93	58
4	0.90	0.94	0.92	141
accuracy			0.95	550
macro avg	0.95	0.94	0.94	550
weighted avg	0.95	0.95	0.95	550

Se observa una mejora en la precisión del modelo para la Clase 3 en comparación con los resultados obtenidos al utilizar todos los sensores en el mismo modelo lo que sugiere que, en caso de finalmente utilizar un modelo RF, podrían ser utilizados una menor cantidad de sensores obteniendo un mejor rendimiento en la clasificación.

En la Tabla 10 se presentan los resultados obtenidos utilizando un modelo KNN al excluir los sensores MQ2 y MQ137.

Cuadro 10: Reporte con las principales métricas de clasificación para el modelo KNN

	precision	recall	f1-score	support
1	0.97	0.98	0.98	196
2	0.94	0.95	0.94	155
3	0.93	0.93	0.93	58
4	0.96	0.92	0.94	141
accuracy			0.95	550
macro avg	0.95	0.95	0.95	550
weighted avg	0.95	0.95	0.95	550

En este caso, el modelo KNN en la Clase 3, experimentó una ligera disminución en la exactitud al reducir su sensibilidad 0.05 % manteniendo el valor de precisión constante.

La Tabla 11 presenta los resultados obtenidos en el modelo XGB al excluir los sensores MQ2 y MQ137.

Cuadro 11: Reporte con las principales métricas de clasificación para el modelo XGB

	precision	recall	f1-score	support
1	0.98	0.97	0.98	196
2	0.92	0.92	0.92	155
3	0.96	0.90	0.93	58
4	0.90	0.93	0.92	141
accuracy			0.94	550
macro avg	0.94	0.93	0.94	550
weighted avg	0.94	0.94	0.94	550

En este caso, se obtuvo un mayor valor de precisión, perdiendo exactitud en la clasificación de muestras pertenecientes a la Clase 3, lo que sugiere, en casos de presencia de humo, la posibilidad de no detectar todos los casos, aunque la mayoría de los casos reales de humo serían clasificados correctamente.

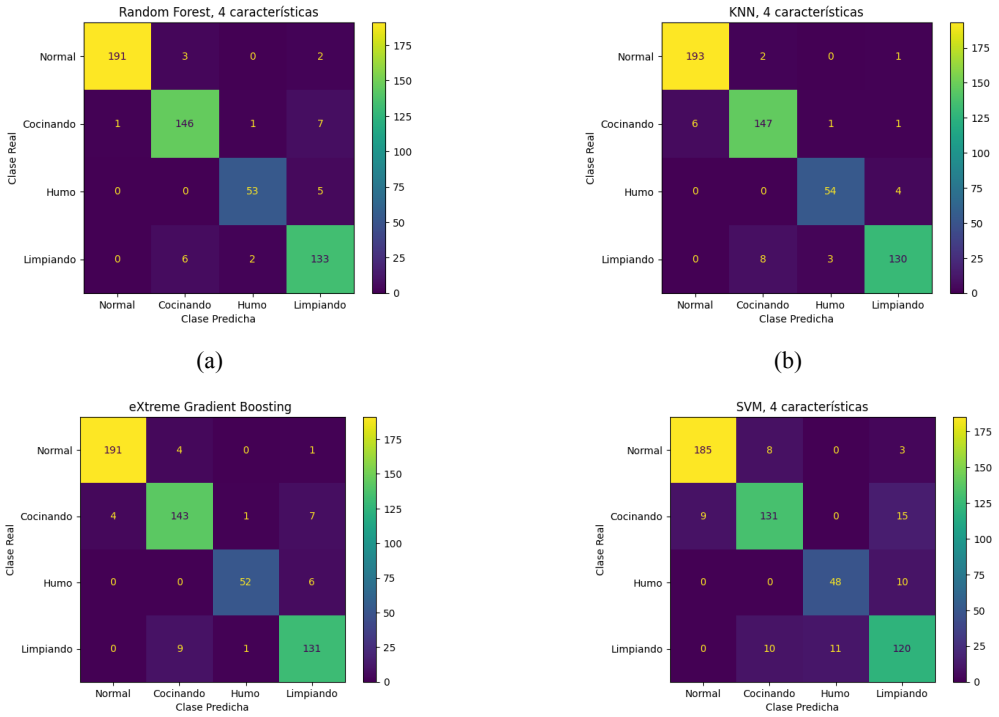
Finalmente, la Tabla 12, presenta los resultados obtenidos con el modelo SVM al excluir los sensores MQ2 y MQ137.

Cuadro 12: Reporte con las principales métricas de clasificación para el modelo SVM

	precision	recall	f1-score	support
1	0.95	0.94	0.95	196
2	0.88	0.85	0.86	155
3	0.81	0.83	0.82	58
4	0.81	0.85	0.83	141
accuracy			0.88	550
macro avg	0.86	0.87	0.87	550
weighted avg	0.88	0.88	0.88	550

Se puede notar que los resultados en la clasificación de la clase 3 empeoraron en comparación tanto con el mismo modelo utilizando todos los sensores como en los modelos RF, XGB y KNN excluyendo 2 sensores. Lo dicho, sugiere que este modelo en principio podría no ser el adecuado.

La Figura 5 ilustra las matrices de confusión obtenidas para los cuatro nuevos modelos desarrollados excluyendo los sensores MQ2 y MQ137.



(a) (b)

(c) (d)

Figura 5: Matriz confusión para los 4 modelos óptimos analizados.

4.2.6. Comparación de modelos y conclusiones

La Tabla 6 resume los resultados de todos los modelos desarrollaods proporcionando una visión general del rendimiento obtenido por cada uno en términos de accuracy, recall, precision, f1-score, MCC-score, tiempo de entrenamiento, tiempo de inferencia y tiempo total (inferencia y entrenamiento). Es importante tener en cuenta que todas las métricas se derivan de los resultados globales del modelo y no de cada clase en particular lo que implica que, el mismo podría ser más o menos efectivo detectando una clase en particular habiendo sido esto ya analizado en secciones previas.

	Accuracy	Recall	Precision	F1-Score	MCC score	Time to Train	Time to Predict	Total Time
KNeighborsClassifier	96.91%	96.91%	96.94%	96.91%	95.70%	0.008	0.034	0.042
XGBClassifier	95.64%	95.64%	95.72%	95.65%	93.94%	3.313	0.007	3.319
KNeighborsClassifier (4 características)	95.27%	95.27%	95.27%	95.26%	93.41%	0.005	0.034	0.038
RandomForestClassifier (4 características)	95.09%	95.09%	95.17%	95.11%	93.16%	0.335	0.012	0.348
RandomForestClassifier	94.73%	94.73%	94.81%	94.75%	92.67%	0.467	0.012	0.479
XGBClassifier (4 características)	94.00%	94.00%	94.05%	94.01%	91.62%	0.206	0.007	0.213
SVM	88.36%	88.36%	88.44%	88.39%	83.80%	0.033	0.026	0.059
SVM (4 características)	88.00%	88.00%	88.13%	88.04%	83.30%	0.032	0.019	0.051

Figura 6: Comparación de modelos basada en métricas globales.

Tras el análisis de la Tabla 6, el modelo KNN mostró el mejor rendimiento de clasificación con un bajo tiempo de entrenamiento. Sin embargo, este modelo requirió uno de los tiempos más largos al momento de predecir nuevos datos, tanto en su versión con 6 sensores como en la de 4.

Si el modelo fuese desarrollado para implementar en una habitación donde la monitorización del aire y la clasificación requiriera de un rápido tiempo de respuesta, podría considerarse el modelo XGB debido a su buen desempeño complementado con un bajo tiempo de predicción.

En términos de costos, el modelo RF sería la opción más adecuada, ya que obtuvo un mejor rendimiento que el modelo XGB en las mismas condiciones.

En resumen, la elección del modelo a utilizar dependerá de los objetivos específicos del sistema, los requisitos de tiempo de respuesta y el presupuesto disponible.

5. Ejercicio 2

5.1. Enunciado

La resistencia a la compresión simple es la característica mecánica principal del concreto. Se define como la capacidad para soportar una carga por unidad de área, y se expresa en términos de esfuerzo [MPa]. El cemento es el material más activo de la mezcla de concreto, por tanto, sus características y sobre todo su contenido (proporción) dentro de la mezcla tienen una gran influencia en la resistencia del concreto, a cualquier edad. A mayor contenido de cemento se puede obtener una mayor resistencia y a menor contenido, la resistencia del concreto va a ser menor. Además, existen varios "agregados" que afectan de manera no lineal las proporciones de los materiales para una resistencia determinada.

Se tiene un *dataset* con información referente a factores que, según los especialistas, se vinculan con la resistencia a la compresión del hormigón. La información de estas variables, su tipo y unidades, se especifican de la siguiente manera:

- ***Cement*** *(component 1) – quantitative – kg in a m3 mixture – Input Variable*
- ***Blast Furnace Slag*** *(component 2) – quantitative – kg in a m3 mixture – Input Variable*
- ***Fly Ash*** *(component 3) – quantitative – kg in a m3 mixture – Input Variable*
- ***Water*** *(component 4) – quantitative – kg in a m3 mixture – Input Variable*
- ***Superplasticizer*** *(component 5) – quantitative – kg in a m3 mixture – Input Variable*
- ***Coarse Aggregate*** *(component 6) – quantitative – kg in a m3 mixture – Input Variable*
- ***Fine Aggregate*** *(component 7) – quantitative – kg in a m3 mixture – Input Variable*
- ***Age*** *– quantitative – Day (1 365) – Input Variable*
- ***Concrete compressive strength*** *– quantitative – MPa – Output Variable*

Los datos se encuentran en el archivo adjunto denominado "dataset_hormigon_regresion.csv"(sección 6). A continuación, se presenta una porción de este dataset:

cement	slag	flyash	water	superplasticizer	coarseaggregate	fineaggregate	age	csMPa
540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.99
540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.89
332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.27
332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.05
198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.30

Cuadro 13: Fragmento del dataset, 2, resistencia del concreto.

En la columna "csMPA" se encuentran los valores numéricos correspondientes a la resistencia a la compresión del hormigón, obtenida para la combinación de los demás materiales en las cantidades indicadas en la fila correspondiente.

El objetivo es determinar si con estas observaciones es posible obtener un modelo que permita **predecir** el valor de resistencia a partir de los componentes y proporciones utilizados en la elaboración del hormigón.

Detallar y fundamentar cada aspecto de la solución propuesta, incluyendo los aspectos que crea conveniente y respondiendo como mínimo las siguientes premisas:

1. ¿Qué regresor se ajusta mejor a la solución buscada? ¿Por qué?
2. En base a los resultados obtenidos, ¿Sería posible utilizar su modelo para predecir la resistencia que tendrá el concreto con una "receta" diferente a las utilizadas?
3. ¿Qué métricas utiliza para evaluar el desempeño de un modelo frente a un problema de este tipo? ¿Podría utilizar las mismas métricas empleadas para el modelo del ejercicio 1?

4. ¿Existen problemas evidentes en este dataset? Si es así, ¿Cuáles son y cómo los solucionaría?
5. El preprocesamiento de los datos, ¿Afecta de alguna manera el desempeño del modelo?
6. Un preprocesamiento incorrecto de los datos, ¿Qué inconvenientes cree que podría generar?

5.2. Resolución

5.2.1. Análisis exploratorio de los datos

En esta sección, se lleva a cabo el análisis exploratorio de los datos del ejercicio 2. La Tabla 14 presenta una concisa descripción del conjunto de datos, el cual consta de 1005 muestras (filas) y 9 variables (columnas) donde una de las variables representa la resistencia a la compresión del hormigón. Si bien esta tabla no da información acerca de la presencia de muestras duplicadas, se destaca la ausencia de valores nulos, confirmando la integridad de los datos. Finalmente, se puede observar que todos los datos son del tipo decimal (*float*) menos la variable "edad" la cual es del tipo entero.

Cuadro 14: Descripción de los datos.

Index: 1005 entries, 0 to 1029			
Data columns (total 9 columns):			
	Column	Non-Null Count	Dtype
0	cement	1005 non-null	float64
1	slag	1005 non-null	float64
2	flyash	1005 non-null	float64
3	water	1005 non-null	float64
4	superplasticizer	1005 non-null	float64
5	coarseaggregate	1005 non-null	float64
6	fineaggregate	1005 non-null	float64
7	age	1005 non-null	int64
8	csMPa	1005 non-null	float64

La Tabla 15 proporciona estadísticas adicionales (media, desviación estándar, valores mínimo y máximo, y los cuartiles 25, 50 y 75) de los datos. Se observa una discrepancia significativa entre el desvío estándar y la media para todas las variables incluso, en algunos casos, con el desvío estándar superando el valor de media destacando la presencia de una variabilidad significativa en los datos, lo que debe ser cuidadosamente considerado durante el análisis y desarrollo del modelado ya que una variabilidad alta en algunas variables podría influir en la capacidad del modelo para generalizar eficazmente. Finalmente, dado que la magnitud de las variables son similares, sugiriendo que inicialmente no sería necesario escalar o normalizar los datos.

Cuadro 15: Descripción estadística de los datos.

	cement	**slag**	**flyash**	**water**	**superplasticizer**	**coarseaggregate**	**fineaggregate**	**age**	**csMPa**
count	1005.00	1005.00	1005.00	1005.00	1005.00	1005.00	1005.00	1005.00	1005.00
mean	278.63	72.04	55.54	182.08	6.03	974.38	772.69	45.86	35.25
std	104.34	86.17	64.21	21.34	5.92	77.58	80.34	63.73	16.28
min	102.00	0.00	0.00	121.80	0.00	801.00	594.00	1.00	2.33
25 %	190.70	0.00	0.00	166.60	0.00	932.00	724.30	7.00	23.52
50 %	265.00	20.00	0.00	185.70	6.10	968.00	780.00	28.00	33.80
75 %	349.00	142.50	118.30	192.90	10.00	1031.00	822.20	56.00	44.87
max	540.00	359.40	200.10	247.00	32.20	1145.00	992.60	365.00	82.60

La Figura 7 ilustra la matriz de correlación entre las variables de entrada y la salida. Si bien no se observa valores altos de correlación entre las variables, se podría destacar la correlación presente entre la resistencia a la compresión del hormigón y el cemento, el agua y la edad. Sin embargo, el valor de estas correlaciones no alcanza niveles sustanciales. Este análisis sugiere que la relación lineal entre las tres variables y la resistencia del concreto puede no ser dominante y que la contribución de cada componente puede ser independiente o no lineal, aspectos que deben considerarse al desarrollar el modelo de regresión.

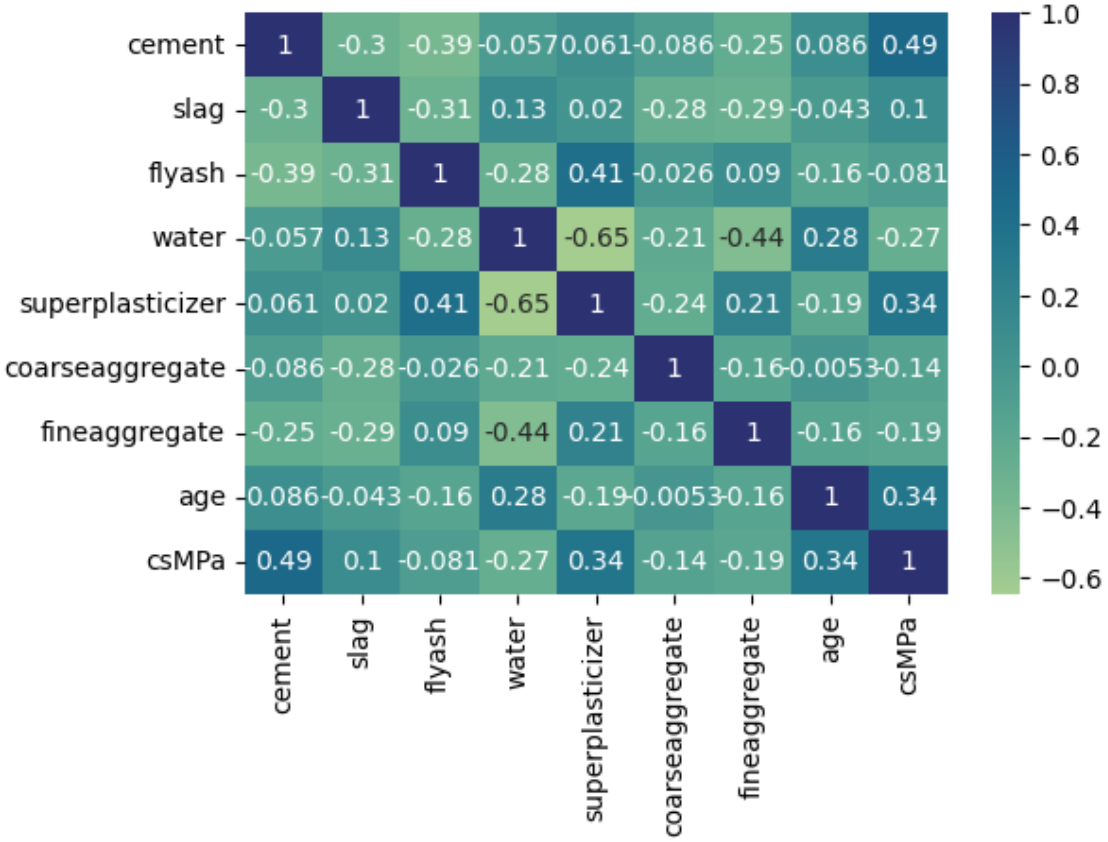


Figura 7: Matriz de correlación entre variables.

La Figura 8, muestra la relación entre las diferentes variables. Nuevamente, al igual que en 7, puede notarse que no existen valores altos de correlación entre las variables por lo que, en una primera instancia, se evaluará un modelo de regresión utilizando todas las variables disponibles a fin de determinar si la inclusión o exclusión de alguna variable mejora o empeora el rendimiento del modelo final.

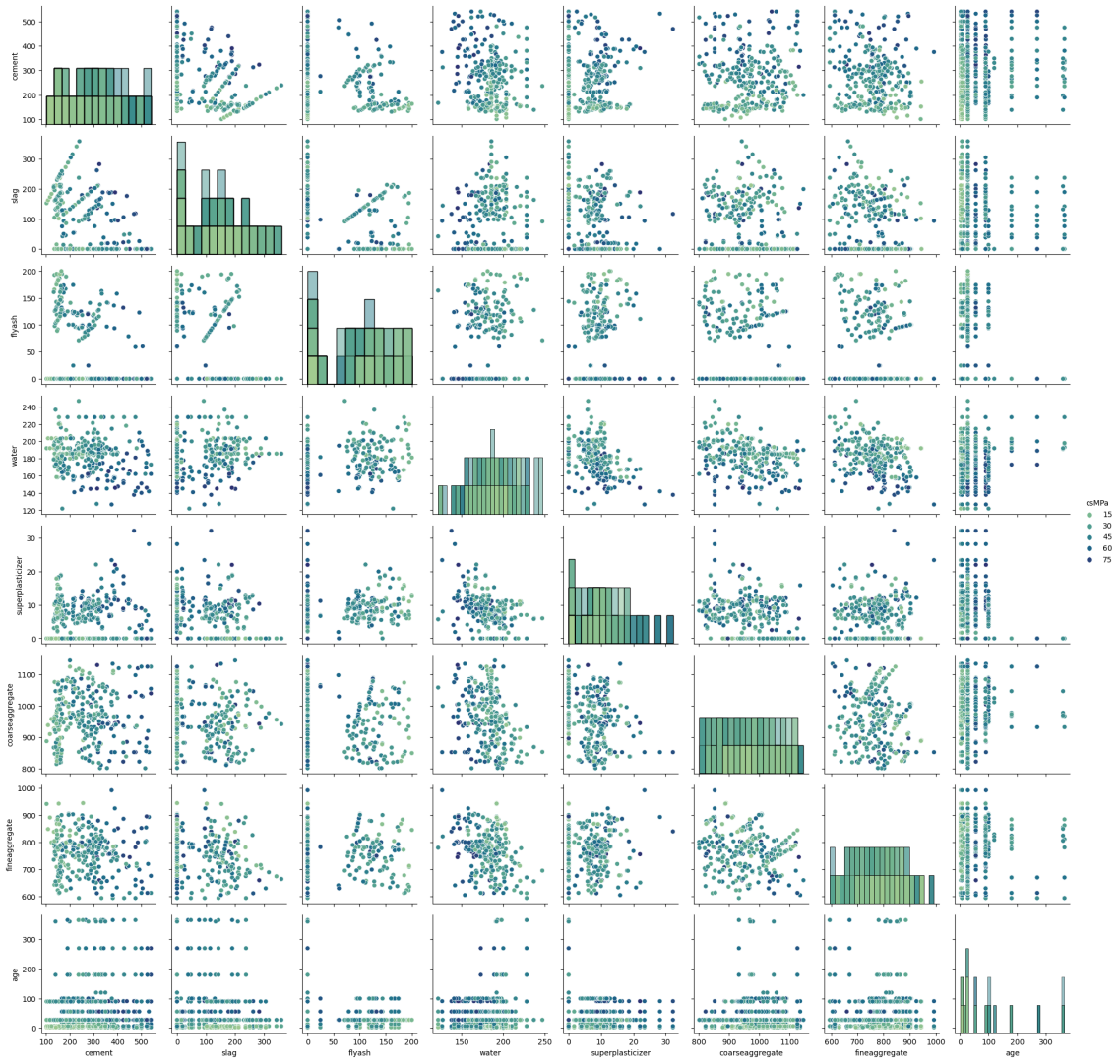


Figura 8: Relación entre las variables analizadas.

5.2.2. Preprocesamiento de los datos

En la fase inicial o etapa de preprocesamiento de datos, fueron identificadas y eliminadas 25 filas duplicadas. Como fue mencionado anteriormente en la sección 5.2.1, en este caso, no se llevó a cabo el escalado ni la normalización de los datos. Posteriormente, el conjunto de datos fue dividido en dos partes: el conjunto de entrenamiento, que comprendió el 70 % del total de datos, y el conjunto de prueba, que representó el 30 %. Es importante destacar que, antes de la división de datos, fue realizada una mezcla de las muestras en el dataset con el objetivo de romper cualquier estructura inherente producida por la carga de los mismos en la tabla, asegurando así que las muestras fueran lo más aleatorias posible. Este enfoque contribuye a garantizar la representatividad y generalización del modelo ante diferentes escenarios. Además, para todos los casos, siempre fue utilizado el valor 8 como semilla.

5.2.3. Primer análisis - LazyPredict

En búsqueda del mejor modelo de regresión, en primera instancia, se empleó la librería de código *LazyPredict* a fin de identificar los modelos candidatos más prometedores. Los resultados obtenidos se detallan en la Tabla 16. Este enfoque inicial proporciona una visión general de los modelos que podrían ser considerados como punto de partida en el desarrollo del modelo de regresión para predecir la resistencia a la compresión del hormigón.

Cuadro 16: Resultados de LazyPredict para los modelos candidatos.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ExtraTreesRegressor	0.92	0.92	4.54	0.3
LGBMRegressor	0.92	0.92	4.56	0.13
HistGradientBoostingRegressor	0.91	0.91	4.65	0.29
XGBRegressor	0.9	0.91	4.89	0.17
GradientBoostingRegressor	0.9	0.9	5.02	0.23
RandomForestRegressor	0.88	0.88	5.43	0.41
BaggingRegressor	0.86	0.87	5.78	0.08
ExtraTreeRegressor	0.83	0.84	6.39	0.01
DecisionTreeRegressor	0.81	0.82	6.83	0.02
AdaBoostRegressor	0.75	0.76	7.87	0.2
KNeighborsRegressor	0.66	0.66	9.21	0.01
SVR	0.59	0.6	10.01	0.05
NuSVR	0.56	0.58	10.35	0.06
TransformedTargetRegressor	0.54	0.55	10.62	0.01
LinearRegression	0.54	0.55	10.62	0.01
RidgeCV	0.54	0.55	10.62	0.01
Ridge	0.54	0.55	10.62	0.01
BayesianRidge	0.54	0.55	10.62	0.02
ElasticNetCV	0.54	0.55	10.63	0.09
LarsCV	0.54	0.55	10.63	0.06
LassoLarsCV	0.54	0.55	10.63	0.02
LassoCV	0.54	0.55	10.63	0.06
LassoLarsIC	0.54	0.55	10.63	0.02
SGDRegressor	0.54	0.55	10.69	0.02
Lasso	0.51	0.52	11.04	0.01
LassoLars	0.51	0.52	11.04	0.01
LinearSVR	0.5	0.52	11.07	0.01
HuberRegressor	0.49	0.51	11.17	0.02
Lars	0.49	0.51	11.19	0.04
OrthogonalMatchingPursuitCV	0.49	0.5	11.19	0.02
ElasticNet	0.46	0.48	11.48	0.02
PoissonRegressor	0.43	0.45	11.81	0.01
TweedieRegressor	0.41	0.43	12.02	0.01
GammaRegressor	0.4	0.41	12.19	0.01
MLPRegressor	0.37	0.39	12.42	0.82
RANSACRegressor	0.25	0.27	13.61	0.14
OrthogonalMatchingPursuit	0.17	0.19	14.31	0.01
DummyRegressor	-0.04	-0.01	15.99	0.01
PassiveAggressiveRegressor	-0.05	-0.02	16.1	0.01
GaussianProcessRegressor	-1.84	-1.76	26.42	0.09
KernelRidge	-4.18	-4.04	35.71	0.05

Tras analizar los resultados, se decidió profundizar (en las siguientes secciones) en los siguientes modelo: *LightGBM* (LGBM), por liderar en r^2 junto con *Extra Trees*, el modelo *XGBoost Extreme Gradient Boosting* (XGB) dado que se trata del segundo modelo con mayor r^2 y el modelo *Random Forest* (RF) por ser el tercer modelo con mejor r^2.

5.2.4. Desarrollo del modelo de predicción

En la primera etapa cada modelo de regresión (ver modelos elegidos en 5.2.3), fue entrenado con los parámetros por defecto utilizando todas las variables como entradas. Para cada modelo, se calcularon las siguientes métricas de evaluación: Error Cuadrático Medio (MSE, *Mean Squared Error*), la Raíz del Error Cuadrático Medio (RMSE, *Root Mean Squared Error*), el Error Absoluto Medio (MAE, *Mean Absolut Error*), el Coeficiente de Determinación (r^2, *R-Squared*), y se registraron los tiempos de entrenamiento, predicción y el tiempo total de entrenamiento más predicción.

La tabla 20 ilustra los resultados obtenidos para cada modelo utilizando los hiperparámetros por defecto. Se puede ver que, tal como previamente se había observado utilizando la librería *LazyPredict* en 5.2.3, el mejor modelo resulto ser el modelo LGBM quien además fue el que menos tiempo requirió tanto para entrenarse como para realizar predicciones sobre nuevas muestras, en este caso igualando al modelo XGB. Por lo tanto, para las condiciones de análisis propuestas, el mismo conjunto de datos para entrenamiento y evaluación y haciendo uso de los hiperparámetros por defecto en cada modelo, LGBM resultó ser el modelo con mejor desempeño en la inferencia de nuevos valores. En la siguiente sección, se realizarán ajustes en los modelos analizados con el objetivo de una mejora en sus desempeños.

Cuadro 17: Comparación de modelos analizados en el estudio de la resistencia a la compresión del hormigón.

	MSE	RMSE	MAE	R-Squared	Time to Train	Time to Predict	Total Time
LGBM	2.1e+01	4.6	3.3	91.60 %	0.1	0.005	0.105
XGB	2.4e+01	4.9	3.4	90.55 %	1.8	0.005	1.774
Random Forest	2.9e+01	5.4	3.9	88.37 %	0.56	0.010	0.569

5.2.5. Desarrollo del modelo de predicción óptimo

En búsqueda del modelo óptimo, en términos de menor cantidad de variables de entrada ó mejor rendimiento en la predicción, en esta sección, se analizarán el modelo RF utilizando solo las variables más importantes al momento de entrenar el modelo. Luego, se utilizará *GridSearchCV* con el fin de buscar nuevos valores de los hiperparámetros los cuales sean capaces de mejorar el desempeño en la predicción de los modelos RF y LGBM.

A continuación, se analiza el modelo RF utilizando únicamente las 3 variables más importantes con los hiperparámetros por defecto. Para esto, la Tabla 18 detalla la importancia asignada a cada variable durante la etapa de entrenamiento del modelo. Puede observase que la variable edad, el cemento y el agua son las tres más relevantes.

Cuadro 18: Importancia de las Features en el Modelo RF.

	Valores
age	0.36
cement	0.31
water	0.10
slag	0.08
superplasticizer	0.06
fineaggregate	0.04
coarseaggregate	0.03
flyash	0.02

Los resultados obtenidos al entrenar el modelo RF con estas tres variables se presentan en la Tabla 19. Luego, en la sección 5.2.6, se discuten los resultados obtenidos mediante un análisis comparativo con modelos restantes obtenidos.

Cuadro 19: Resultados del modelo RF con las 3 features más importantes.

	MSE	RMSE	MAE	R-Squared	Time to Train	Time to Predict	Total Time
Random Forest	5.1e+01	7.1	5.4	79.79 %	0.2449	0.0098	0.2547

El siguiente modelo analizado, es el modelo RF utilizando la técnica de búsqueda de hiperparámetros mediante búsqueda de cuadrícula. En este proceso, se fueron exploradas diversas combinaciones de valores de

hiperparámetros (ver Tabla 20) a través de *cross-validation* con 3 *folds*, resultando en la evaluación de un total de 288 combinaciones, realizando 864 ajustes.

Cuadro 20: Hiperparámetros evaluados en el Grid Search para el modelo RF.

n_estimators	(50, 75, 100, ..., 500)
max_features	(sqrt, log2)
min_samples_split	(2, 4)
min_samples_leaf	(1, 2)
bootstrap	(True, False)

Los resultados obtenidos al entrenar el modelo RF con el valor de los mejores hiperparámetros encontrados se presentan en la Tabla 21. Estos resultados serán analizados en la sección 5.2.6, donde se realizará una comparación detallada con los demás modelos evaluados.

Cuadro 21: Resultados del modelo RF con Grid Search.

	MSE	RMSE	MAE	R-Squared	Time to Train	Time to Predict	Total Time
Random Forest	2.1e+01	4.6	3.3	91.50 %	407.4155	0.0111	407.4267

Más información acerca de los valores de hiperparámetros conseguidos pueden encontrarse en el link al notebook utilizado en la resolución de este ejercicio el cual se encuentra en la sección 6.

Si bien en este punto del análisis el mejor modelo es el modelo LGBM, se utilizará nuevamente la técnica de la búsqueda de hiperparámetros con *GridSearchCV* (ver Tabla 22) de modo tal de incrementar el desempeño en la predicción del mismo. La búsqueda, al igual que la realizada para el modelo RF, se realizó mediante *cross-validation* con 3 *folds*, resultando en la evaluación de un total de 729 combinaciones, realizando 2187 ajustes.

Cuadro 22: Hiperparámetros evaluados en el Grid Search para el modelo LGBM.

num_leaves	(20, 30, 40)
learning_rate	(0.01, 0.05, 0.1)
max_depth	(5, 10, 15)
min_child_samples	(10, 20, 30)
subsample	(0.8, 0.9, 1.0)
colsample_bytree	(0.8, 0.9, 1.0)

Los resultados obtenidos al entrenar el modelo LGBM con el valor de los mejores hiperparámetros encontrados se presentan en la Tabla 23. Estos resultados serán analizados en la sección 5.2.6, donde se realizará una comparación detallada con los demás modelos evaluados.

Cuadro 23: Resultados del Modelo LGBM con Grid Search.

	MSE	RMSE	MAE	R-Squared	Time to Train	Time to Predict	Total Time
LGBM	1.8e+01	4.3	3.2	92.73 %	96.8278	0.0050	96.8328

5.2.6. Comparación de modelos y conclusiones

En la Tabla 9 se muestra una comparación del desempeño de los seis modelos entrenados basada en métricas globales.

De los resultados, se observa que el modelo LGBM con optimización mediante *GridSearch* tuvo el mejor rendimiento en términos de MSE, RMSE y MAE, alcanzando un R^2 del 92.73 %. Si bien el tiempo requerido para su entrenamiento fue significativo en comparación con los modelos restantes, requirió de un tiempo en la predicción que el modelo RF utilizando todas las variables para su entrenamiento y comparable al tiempo requerido por el modelo XGB.

El modelo RF entrenado con solo las tres variables más importantes (cemento, agua y edad), experimentó una disminución en su rendimiento mostrando el peor desempeño en términos de MSE, RMSE y MAE. Si bien respecto al modelo RF con 6 variables requirió de un tiempo menor de entrenamiento, reducción de

aproximadamente el 50 %, el tiempo de predicción no se redujo.

	MSE	RMSE	MAE	R-Squared	Time to Train	Time to Predict	Total Time
LGBMRegressor with Grid	1.8e+01	4.3	3.2	92.73%	9.7e+01	0.005	96.833
LGBMRegressor	2.1e+01	4.6	3.3	91.60%	0.1	0.005	0.105
RandomForestRegressor with Grid	2.1e+01	4.6	3.3	91.50%	4.1e+02	0.011	407.427
XGBRegressor	2.4e+01	4.9	3.4	90.55%	1.8	0.005	1.774
RandomForestRegressor	2.9e+01	5.4	3.9	88.37%	0.56	0.010	0.569
RandomForestRegressor (3 características)	5.1e+01	7.1	5.4	79.79%	0.24	0.010	0.255

Figura 9: Comparación de modelos basada en métricas globales.

En resumen, el modelo LGBM con optimización mediante *GridSearch* se destaca como la mejor opción a utilizar en la predicción de la resistencia a la compresión del hormigón bajo las condiciones de análisis tenidas en cuenta para el conjunto de datos analizado.

6. Material complementario

6.1. Notebooks

- Análisis de la calidad del aire (ejercicio 1)
- Análisis de la resistencia del concreto (ejercicio 2)

6.2. Base de datos

Gambi, Ennio (2020), “Air Quality dataset for ADL classification”, Mendeley Data, V1, doi: 10.17632/kn3x9rz3kd.1. Base de datos (ejercicio 1)

Yeh, I-C. ”Modeling of strength of high-performance concrete using artificial neural networks.Çement and Concrete research 28.12 (1998): 1797-1808. Base de datos (ejercicio 2)