Título: Trabajo Final de análisis de datos aplicando herramientas de ML.

Fabricio, Machado; Adrián, Fernandez; Hernán, Bernal**

UNRaf, Universidad Nacional de Rafaela, Santa Fe, Argentina.
Inteligencia Artificial y Aprendizaje Automático
(**) Autor Corresp.: bernalhd18@gmail.com

Resumen:

Este estudio, realizado en el marco de la materia "Inteligencia Artificial y Aprendizaje Automático," analiza un conjunto de datos sobre salud y prescripciones de pacientes con demencia. El objetivo fue explorar el conjunto de datos mediante análisis estadístico y descriptivo y desarrollar modelos óptimos de predicción y clasificación. Se probaron diversos algoritmos y, tras evaluar su desempeño con diferentes métricas, se seleccionaron los más adecuados para cada tarea.

Los resultados mostraron que el algoritmo de Regresión por Máquinas de Soporte Vectorial (SVR), en su versión simple y combinado con Adaboost, fue el más eficiente para regresión. En clasificación, los modelos de Bosques Aleatorios y Máquinas de Soporte Vectorial (SVM) destacaron, especialmente con selección de características. Los hallazgos subrayan la importancia de un análisis exploratorio detallado para adaptar los modelos al problema, ya que la alta dimensionalidad afectó negativamente el desempeño en regresión, mientras que una reducción excesiva de características dificultó la clasificación. Este trabajo demuestra la necesidad de equilibrar adecuadamente la reducción de dimensionalidad y la selección de características para optimizar el rendimiento de los modelos en conjuntos de datos médicos.

Palabras claves: Aprendizaje Automático, Demencia, Análisis de Datos, Algoritmos de Clasificación, Algoritmos de Predicción.

1. Introducción

La demencia es un síndrome caracterizado por el deterioro del juicio y las capacidades cognitivas, incluyendo la memoria, el pensamiento y el razonamiento, así como por cambios en el estado de ánimo y dificultades en el lenguaje y la comunicación, lo que afecta significativamente la vida diaria de quienes la padecen. Cabe destacar que no se trata de una enfermedad específica, sino de un conjunto de síntomas asociados a enfermedades subyacentes como la demencia vascular, el Alzheimer, la demencia frontotemporal o la demencia por cuerpos de Lewy.

Según un informe de la Comisión Lancet sobre prevención, intervención y atención de la demencia (2020), los factores de riesgo asociados con el desarrollo de este síndrome incluyen: bajo nivel educativo, pérdida auditiva, hipertensión arterial, obesidad, depresión, diabetes, inactividad física, aislamiento social, tabaquismo, consumo de alcohol, lesiones cerebrales traumáticas y exposición a la contaminación del aire.

Dado que actualmente no existe una cura para la demencia, resulta crucial identificar los factores que podrían desencadenar los síntomas, con el objetivo de implementar estrategias de prevención y tratamiento temprano. En este contexto, y aprovechando la creciente disponibilidad de datos clínicos almacenados electrónicamente, el presente estudio utiliza herramientas de aprendizaje automático para predecir y clasificar casos de demencia, abordando tanto la detección temprana como la evolución de la enfermedad.

Estudios previos [1] han explorado la relación entre factores demográficos, cognitivos, de salud, de estilo de vida y genéticos, y la prevalencia de demencia, empleando principalmente modelos de regresión logística. Este enfoque ha demostrado ser útil para analizar datos médicos complejos e identificar factores de riesgo clave en salud mental. Por otro lado, investigaciones específicas [2] han desarrollado modelos de clasificación diseñados para grupos con enfermedades preexistentes como diabetes, accidentes cerebrovasculares o fibrilación auricular. Estos modelos integraron técnicas de aprendizaje automático, como bosques aleatorios, máquinas de soporte vectorial y redes neuronales, en combinación con métodos estadísticos tradicionales, como la regresión logística y los modelos de Cox. Estas aproximaciones destacan la importancia de adaptar los modelos predictivos a las características particulares de cada grupo, mejorando su efectividad clínica.

A diferencia de los estudios previamente mencionados, este trabajo propone un enfoque integral que combina clasificación y predicción. Por un lado, se busca clasificar a los pacientes como diagnosticados o no diagnosticados con

demencia; por otro, se pretende predecir la edad probable en la que podrían desarrollar la enfermedad, considerando tanto factores relacionados con el estilo de vida como con los resultados médicos. Adicionalmente, se realizaron análisis estadísticos y descriptivos para explorar las relaciones entre las variables involucradas, lo que permite una comprensión más profunda de los factores que contribuyen al desarrollo de la demencia.

El objetivo principal de este trabajo es desarrollar modelos de predicción y clasificación que identifiquen, de manera precisa y confiable, tanto la presencia de demencia en pacientes como la edad probable en la que podría manifestarse. Para ello, se emplearán diferentes algoritmos de aprendizaje automático que integran características relacionadas con el estilo de vida y factores médicos. Posteriormente, se compararán los modelos generados utilizando diversas métricas de desempeño, con el fin de identificar el algoritmo más adecuado para cada tarea.

El artículo se organiza de la siguiente manera: en la sección 2 (Metodología), se describe el conjunto de datos utilizado, los procedimientos de limpieza y preprocesamiento aplicados, así como los algoritmos implementados, diferenciando entre los procesos de regresión y clasificación. En la sección 3, se presentan los resultados obtenidos y su discusión. Finalmente, la sección 4 resume los principales hallazgos y conclusiones del estudio.

2. Metodología

2.1 Conjunto de datos utilizado

El conjunto de datos utilizado constituye una fuente rica y valiosa para analizar los factores asociados con el inicio y la progresión de la demencia. Este recurso integra 1000 registros de pacientes y 24 variables que abarcan información clave sobre estilo de vida, características genéticas y resultados de salud, permitiendo a los investigadores explorar las complejas interacciones entre estas dimensiones.

Los datos fueron recopilados de diversas fuentes confiables, incluyendo bases de datos científicas reconocidas como PubMed y Google Scholar, el Sistema Nacional de Salud (NHS) del Reino Unido, y consultas directas con profesionales de la salud, garantizando la calidad y relevancia del conjunto de datos.

2.2 Proceso de preprocesamiento aplicado

La limpieza de datos se realizó en tres niveles siguiendo las recomendaciones de Hands-On Data Preprocessing in Python de Roy Jafari (2022):

Nivel 1: *Estandarización de la estructura de datos.*

Los atributos del dataset se renombraron utilizando el formato *snake case* para facilitar su codificación y análisis.

Se verificó la unicidad de las filas asegurando que cada entrada tuviera un identificador único y no existieran duplicados.

Nivel 2: *Reestructuración del conjunto de datos.*

Se evaluó la estructura del dataset inicial y se determinó que su disposición era adecuada para los análisis previstos. No fue necesario reestructurar el conjunto de datos.

Nivel 3: *Evaluación y corrección de valores.*

Errores en los datos:

Se confirmó que todos los valores presentaban tipos de datos adecuados.

Valores faltantes:

Los atributos *prescription* y *dosage_in_mg* presentaban 515 valores faltantes. Estos registros correspondían a pacientes sin diagnóstico de demencia, lo cual es consistente con la ausencia de prescripciones.

El atributo *chronic_health_conditions* presentaba 179 valores faltantes. Estos se atribuyeron a pacientes sin condiciones crónicas, lo que también tiene sentido clínico.

En ambos casos, los valores nulos se reemplazaron con el valor 'None' para preservar la coherencia semántica y evitar problemas durante el análisis.

Detección de valores atípicos:

Solo se registran outliers para el atributo *dosage_in_mg*.

Como se puede visualizar en la imagen a continuación, los outliers no parecen ser valores erróneos, dado que en relación con el atributo *prescription*, parece estar dentro de los valores esperados para la prescripción de "Donepezil".

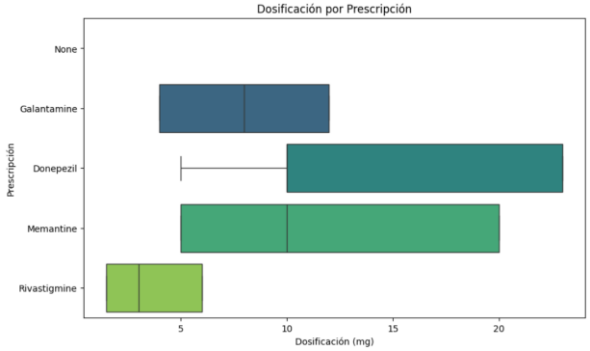

Fig. 1: Distribución de las dosis de medicamentos prescritos (en mg) según el tipo de prescripción.

2.3 Transformación de datos

Codificación de variables categóricas:

Para garantizar la correcta aplicación de los algoritmos de Machine Learning, se procedió a codificar las variables categóricas relevantes en el análisis. A cada categoría se le asignó un valor numérico entero, de manera que el modelo pueda interpretar adecuadamente estas variables como parte del vector de características sin generar inconsistencias.

Escalado de datos:
Con el objetivo de evitar sesgos o resultados incorrectos en el análisis, y mejorar la efectividad de los algoritmos, se aplicó un escalado a los datos mediante la clase *StandardScaler*. Este proceso asegura que todas las variables tengan un rango de valores similar, lo que facilita la convergencia de los modelos y mejora la precisión en las predicciones.

2.4 Herramientas analíticas aplicadas

Regresión:

Variables y justificación del modelo:

Para la aplicación de modelos de regresión, dado que ninguno de los atributos numéricos presentes en el dataset tiene una relevancia clara y directa que justifique una regresión sobre ellos en el contexto de los datos, hemos decidido suponer que el atributo “age” representa la edad en la cual el paciente fue diagnosticado con demencia.

Esta suposición le otorga un contexto clínico relevante, haciendo que la predicción de esta variable tenga un sentido más lógico dentro del análisis de los datos.

Modelos seleccionados:

Se implementaron tres modelos individuales y tres modelos de ensamble para abordar el problema de regresión. La selección responde a la necesidad de explorar distintos enfoques y obtener un modelo robusto:

Tipo de modelo	Modelo
Individual	Support Vector Regressor (SVR)
	Decision Tree Regressor
	Logistic Regression Multinomial
Ensamble	Random Forest Regressor
	SVR con Adaboost
	XGBoost Regressor

Datos de entrenamiento y validación:

Los modelos fueron entrenados utilizando subconjuntos generados mediante la función *train_test_split*, con un valor fijo de *random_state* para garantizar reproducibilidad y asegurar condiciones consistentes para una comparación justa entre los modelos.

La partición se realizó en una proporción de 80% para entrenamiento y 20% para validación.

A partir de estos datos de entrenamiento, se han generado dos subconjuntos de datos de entrenamiento para posterior comparación:

- Con la totalidad de los atributos.
- Con una selección de los primeros seis atributos más importantes en base al modelo *RandomForest* (feature importance).

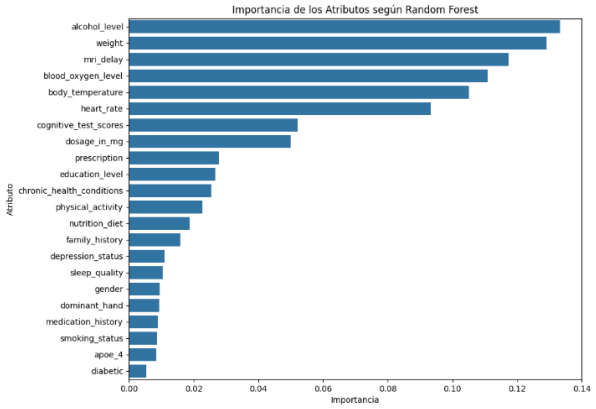

Fig. 2: Importancia de los atributos en la predicción según el modelo Random Forest. Los atributos están ordenados de mayor a menor importancia.

Selección de hiperparámetros:

Se utilizó la técnica de búsqueda en cuadrícula (*GridSearchCV*) para optimizar los parámetros de cada modelo.

La métrica de evaluación seleccionada fue el *neg_mean_squared_error*, dada su capacidad para penalizar los errores más significativos, lo que contribuye a obtener modelos más precisos.

Selección de métricas:

Para evaluar el desempeño de los modelos, se utilizaron las siguientes métricas:

- M: Error absoluto máximo.
- MAE (*Mean Absolute Error*): Error absoluto promedio.

- MSE (*Mean Squared Error*): Error cuadrático medio, que penaliza errores grandes de predicción.
- RMSE (*Root Mean Squared Error*): Raíz cuadrada del error cuadrático medio, utilizada para interpretar los errores en la misma escala que las variables originales.

Para cada modelo evaluado, se realizaron dos experimentos de entrenamiento:

Con la totalidad de los atributos disponibles: donde todos los atributos del dataset se usaron como entradas del modelo.

Con una selección de los atributos más importantes: definida en base a la métrica de importancia de variables generada por el modelo *Random Forest Regressor*.

El análisis se enfoca en dos aspectos principales:

Comparar el impacto de usar la totalidad de los atributos frente a utilizar únicamente los más relevantes.

Contrastar el desempeño entre los distintos modelos de aprendizaje automático.

Clasificación:

Variables y justificación del modelo:

Se seleccionaron 21 variables del conjunto de datos original, relacionadas con el estilo de vida y los resultados médicos de pacientes tanto con como sin diagnóstico de demencia. La selección de esta cantidad de variables responde al objetivo de construir un modelo capaz de generalizar frente a múltiples factores provenientes de diferentes ámbitos.

Se excluyeron del vector de características las variables *dementia*, *dosage_in_mg* y *prescription*, ya que éstas sólo estaban presentes en pacientes diagnosticados con demencia. Su inclusión habría introducido sesgos significativos en el análisis, afectando la precisión de la clasificación, dado que dichas variables actúan como marcadores explícitos del diagnóstico y no como predictores independientes.

Modelos seleccionados:

Se eligieron tres modelos de clasificación: dos algoritmos individuales (Máquinas de Soporte Vectorial y Regresión Logística) y un método de ensamble (Bosques Aleatorios). La selección de estos modelos se fundamentó en sus características particulares y en su capacidad para abordar problemas específicos relacionados con el conjunto de datos analizado:

- Bosques Aleatorios: Seleccionado por su capacidad para capturar patrones no lineales en los datos, debido a la ausencia de correlaciones significativas entre las variables de estudio.
- Máquinas de Soporte Vectorial (SVM): Adecuado para espacios de alta dimensionalidad y conjuntos de datos pequeños o medianos, como el utilizado en este estudio. La posibilidad de utilizar funciones kernel permite manejar datos no linealmente separables.
- Regresión Logística: Ideal para clasificar datos binarios ("demencia" o "no demencia").

Por su parte, en el caso de Regresión Logística, se llevaron a cabo tres enfoques principales para construir el modelo y evaluar su desempeño, con el objetivo de determinar el método más óptimo y efectivo. Estos enfoques incluyeron:

1. Clase *LogisticRegression* con validación cruzada: Evaluación del desempeño mediante validación cruzada para garantizar la capacidad del modelo de generalizar a nuevos datos.
2. Clase *LogisticRegressionCV*: Automatización de la búsqueda del mejor parámetro regularizador (*C*), optimizando el rendimiento y reduciendo el riesgo de sobreajuste.
3. Clase *LogisticRegression* sin validación cruzada: Evaluación del comportamiento del modelo sin ajustes adicionales, proporcionando una línea base para la comparación de resultados.

Datos de entrenamiento y validación:

Los datos se dividieron en un 80% para entrenamiento y un 20% para validación, utilizando la función train_test_split con un valor fijo de random_state para garantizar reproducibilidad. A partir de los datos de entrenamiento, se generaron tres configuraciones de modelos para cada algoritmo:

- Uso de la totalidad de los atributos disponibles.
- Reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA).
- Selección de los 12 atributos más importantes según el análisis de importancia de características (*feature importance*) del modelo Bosques Aleatorios.

Este diseño experimental permitió evaluar el desempeño de los algoritmos bajo diferentes configuraciones, considerando su capacidad para manejar datos de alta dimensionalidad y detectar patrones en vectores reducidos.

Selección de hiperparámetros:

Se empleó la técnica de búsqueda en cuadrícula (*GridSearchCV*) para optimizar los hiperparámetros de cada modelo:

- Máquinas de Soporte Vectorial (SVM): Se utilizó un kernel radial base (RBF), ideal para capturar relaciones no lineales en el espacio de características.
- Regresión Logística: Se aplicó una penalización con regularización Lasso (norma L1), que reduce a cero los coeficientes de características irrelevantes, mejorando la interpretabilidad y generalización.

Selección de métricas:

Para evaluar el desempeño de los modelos, se utilizaron las siguientes métricas:

- *Accuracy*: Proporción de predicciones correctas sobre el total de instancias evaluadas.
- *Precision*: Proporción de instancias clasificadas como positivas que realmente lo son.
- *Recall*: Proporción de verdaderos positivos identificados respecto al total de positivos reales.
- F1 *Score*: Media armónica entre *precision* y *recall*.
- Área Bajo la Curva *ROC* (AUC-ROC): Mide la capacidad del modelo para distinguir entre clases, evaluando la relación entre verdaderos positivos y falsos positivos a diferentes umbrales.

Para cada algoritmo, se realizó una comparación de las métricas obtenidas entre los modelos generados con los distintos enfoques, con el objetivo de identificar el modelo más óptimo dentro de cada algoritmo.

Posteriormente, se llevó a cabo una comparación final entre los tres algoritmos, utilizando la totalidad de los datos, para determinar cuál ofrecía el mejor desempeño general.

3. Resultados y Discusión

3.1 Visualización de datos

Regresión:

A continuación, se presentan y analizan los resultados obtenidos al aplicar distintos modelos de aprendizaje automático sobre el conjunto de datos, para evaluar el desempeño de los modelos tanto con la totalidad de los datos, como con una selección de ellos.

Error absoluto máximo (M):

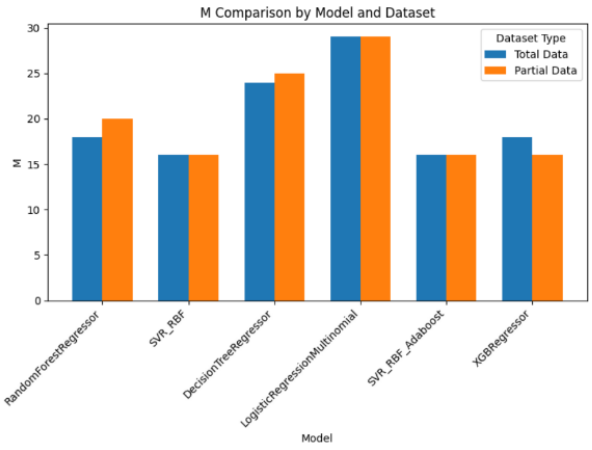

Fig. 3: Comparación de resultados obtenidos para la métrica M, por modelo y conjunto de datos.

En la figura 3, podemos visualizar que la mejor métrica la obtuvieron ambos modelos *SVR*, con un valor de 16.

El modelo *LogisticRegressionMultinomial* arrojó el valor más alto con 29.

El modelo *XGBRegressor* entrenado con la selección de atributos mejoró su error absoluto máximo en comparación con la totalidad de los datos

Error absoluto promedio (MAE):

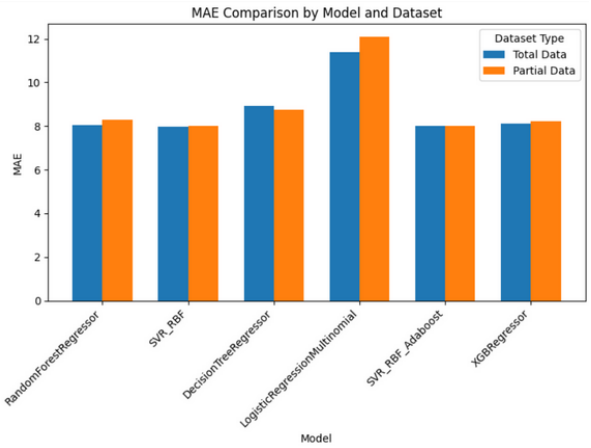

Fig. 4: Comparación de resultados obtenidos para la métrica MAE, por modelo y conjunto de datos.

En la figura 4, podemos visualizar que los valores para la métrica MAE se mantuvieron más parejos excepto por la clase *LogisticRegressionMultinomial* que tiene los valores más altos.

El modelo *SVR* entrenado con la totalidad de los datos obtuvo el mejor valor (7.98) seguido de su versión entrenada con los datos parciales (8.00).

El modelo *DecisionTreeRegressor* obtuvo una mejora en su entrenamiento con la parcialidad de los datos (24) en comparación con el entrenamiento con la totalidad de los datos (25).

Error cuadrático medio (MSE):

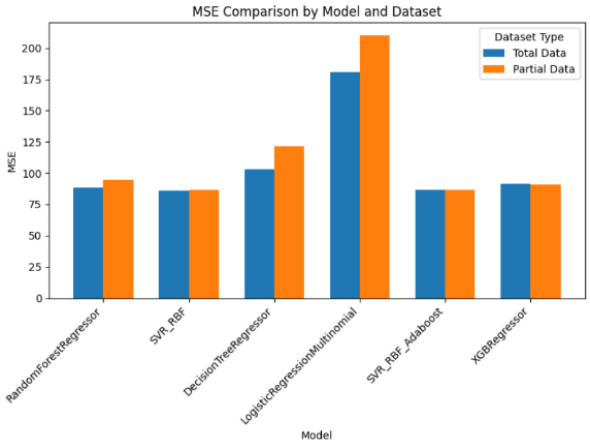

Fig. 5: Comparación de resultados obtenidos para la métrica MSE, por modelo y conjunto de datos.

En la figura 5, podemos observar que el modelo *SVR* entrenado con la totalidad de los datos obtuvo la mejor métrica (86.16) seguido de su versión entrenada con la parcialidad de los datos (86.41).

Para esta métrica no se registran mejoras en los modelos utilizando la parcialidad de los datos.

Raíz cuadrada del error cuadrático medio (RMSE):

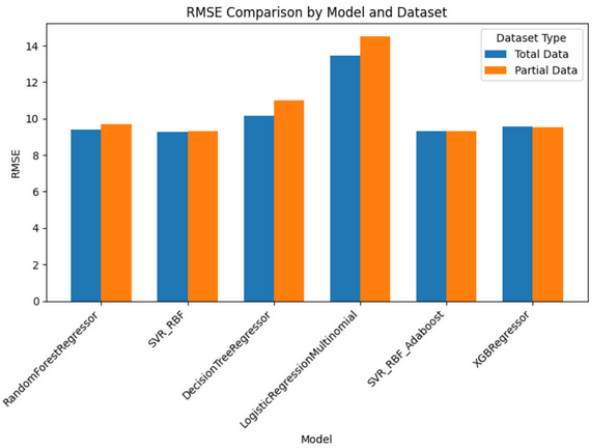

Fig. 6: Comparación de resultados obtenidos para la métrica MSE, por modelo y conjunto de datos.

En la figura 6, podemos visualizar que el modelo *SVR* entrenado con la parcialidad de los datos obtuvo la mejor métrica RMSE con el valor 9.28, seguido de su variante entrenada con la totalidad de los datos 9.3.

No se registran mejoras en los modelos entrenados con la parcialidad de los datos en comparación de su versión entrenada con la totalidad de los datos.

Clasificación:

A continuación, se presentarán los gráficos generados a partir de los modelos de clasificación utilizados.

Bosques Aleatorios:

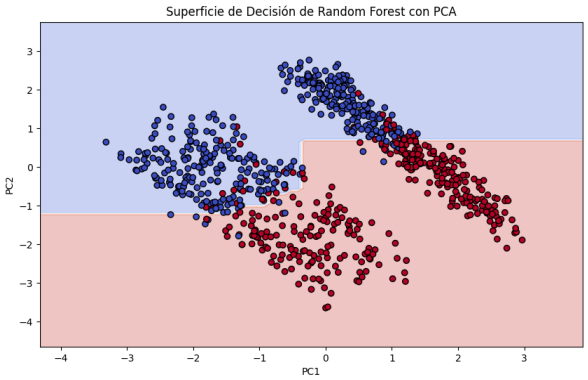

Fig. 7: Superficie de decisión de Bosques Aleatorios aplicando PCA.

En la figura 7 se observa la separación entre las dos clases: "con demencia" (representada por puntos rojos) y "sin demencia" (representada por puntos azules).

El modelo muestra un ajuste adecuado, y las métricas de precisión (0.96) y exactitud (0.93) condicen con las instancias clasificadas. Estas métricas indican un buen desempeño del modelo, destacándose su capacidad para identificar correctamente tanto las instancias positivas como las negativas, con una baja proporción de falsos positivos y falsos negativos.

Máquinas de Soporte Vectorial:

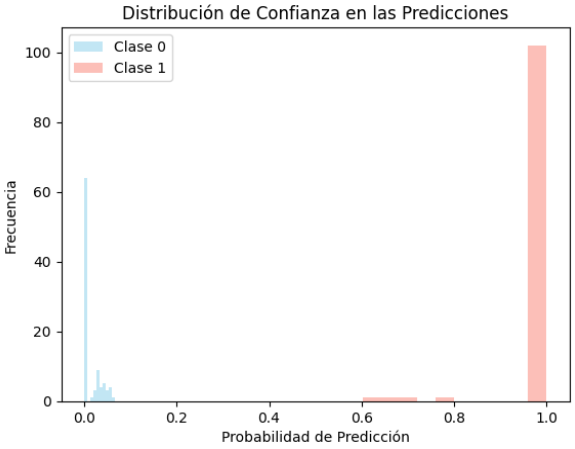

Fig. 8: Distribución de confianza en las predicciones para SVM con el total de características.

En la figura 8 se muestra la distribución de confianza del modelo SVM en las predicciones. Las barras indican la frecuencia de las probabilidades asignadas a las clases: Clase 0 ("sin demencia"), en azul, se agrupa hacia valores bajos, y Clase 1 ("con demencia"), en rojo, hacia valores altos.

La clara separación entre ambas distribuciones sugiere que el modelo clasifica con confianza, evidenciando su capacidad para distinguir correctamente entre las dos clases. Un solapamiento reducido respalda la efectividad del modelo en esta tarea.

Regresión Logística:

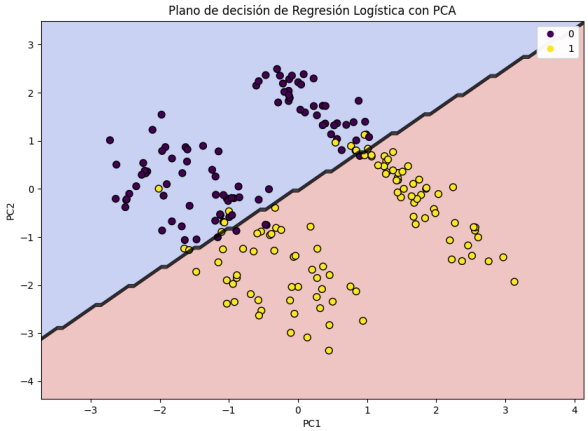

Fig. 9: Plano de decisión de Regresión Logística aplicando PCA.

La Figura 9 muestra la separación de las clases "con demencia" (amarillo) y "sin demencia" (morado), junto con el límite de decisión definido por el modelo de Regresión Logística. El modelo logra una separación razonable entre ambas clases; sin embargo, algunas instancias cercanas al límite reflejan cierta incertidumbre en la clasificación. Esto concuerda con la exactitud del modelo (0.92), que, aunque alta, no alcanza la perfección.

3.2 Resultados de los algoritmos

En la tabla 1 se puede visualizar el detalle del resultado de las métricas obtenido para cada modelo de regresión y su respectivo conjunto de datos.

Tabla 1. Resultado de métricas por modelo y conjunto de datos de entrenamiento, en verde se resalta el modelo con mejores métricas.

Modelo	Dataset	M	MAE	MSE	RMSE
RandomForestRegressor	Total data	18	8.04	88.31	9.4
RandomForestRegressor	Partial data	20	8.28	94.38	9.72
SVR	Total data	16	7.98	86.16	9.28
SVR	Partial data	16	8	86.41	9.3
DecisionTreeRegressor	Total data	24	8.93	103.18	10.16
DecisionTreeRegressor	Partial data	25	8.76	121.4	11.02
LogisticRegressionMultinomial	Total data	29	11.38	180.64	13.44
LogisticRegressionMultinomial	Partial data	29	12.07	209.97	14.49
SVR_Adaboost	Total data	16	8.01	86.44	9.3
SVR_Adaboost	Partial data	16	8	86.41	9.3
XGBRegressor	Total data	18	8.1	94.48	9.56
XGBRegressor	Partial data	16	8.21	90.62	9.52

En la tabla 2 se puede visualizar el detalle del resultado de las métricas obtenido para cada modelo de clasificación y su respectivo conjunto de datos.

Tabla 2. Resultado de métricas por modelo y conjunto de datos de entrenamiento, en verde se resalta el modelo con mejores métricas para cada algoritmo.

Modelo	Dataset	*Accuracy*	*Precision*	*Recall*	F1 *Score*	Curva *ROC*
Bosques Aleatorios	Total data	0.88	0.97	0.79	0.875	0.885
	PCA	0.93	0.96	0.905	0.875	0.885
	Partial data	1.00	1.00	1.00	1.00	1.00
SVM	Total data	0.985	1.00	0.97	0.985	0.985
	PCA	0.93	0.979	0.89	0.93	0.93
	Partial data	1.00	1.00	1.00	1.00	1.00
Regresión Logística	Total data	0.985	1.00	0.97	0.985	0.985
	PCA	0.92	0.95	0.89	0.92	0.92
	Partial data	0.99	1.00	0.98	0.99	0.99

4. Conclusiones

Regresión:

Resumen de los principales resultados:

El análisis comparativo de distintos modelos de aprendizaje automático aplicados al problema de regresión arrojó resultados significativos. El modelo *SVR* (*Support Vector Regressor*), tanto en su versión simple como en combinación con *Adaboost*, destacó consistentemente en todas las métricas evaluadas. En particular, presentó el menor error absoluto máximo (16), el menor error absoluto promedio (7.98) y el menor error cuadrático medio

(86.16) cuando se entrenó con la totalidad de los datos. Esto sugiere que el modelo *SVR* es particularmente efectivo para este conjunto de datos y problema específico.

El modelo *Logistic Regression Multinomial*, en contraste, presentó el desempeño más bajo en todas las métricas, con valores significativamente más altos de error absoluto máximo (29) y error cuadrático medio (180.64), lo que lo descarta como una opción viable para este análisis. Por otro lado, los modelos de ensamblado, como *Random Forest Regressor* y *XGBRegressor*, lograron resultados aceptables, pero sin superar al *SVR* en general.

El análisis sobre la relevancia de los atributos mostró que la reducción del conjunto de datos a los atributos más importantes, según el modelo Random Forest, no siempre mejora las métricas de desempeño. Sin embargo, en algunos casos, como el *XGBRegressor* y el *DecisionTreeRegressor*, esta selección demostró una mejora específica en el error absoluto máximo y error absoluto promedio, respectivamente.

Implicaciones teóricas y prácticas:

Los resultados destacan la importancia de seleccionar modelos que se ajusten adecuadamente a la naturaleza del problema y al conjunto de datos en cuestión. En este caso, la efectividad del *SVR* y su combinación con Adaboost sugiere que los datos presentan patrones que estos modelos pueden capturar de manera eficiente.

La comparación entre el uso de la totalidad de los atributos y la selección de los más importantes plantea una implicación clave: reducir dimensionalidad puede no siempre conducir a un mejor desempeño. Esto refuerza la necesidad de realizar análisis exploratorios previos para entender las relaciones subyacentes entre las variables antes de decidir sobre la reducción del conjunto de datos.

Conclusiones:

El modelo *SVR*, en sus variantes simple y con Adaboost, demostró ser el más robusto y eficiente para el problema de regresión planteado, destacando en todas las métricas de evaluación.

La selección de atributos más importantes no siempre mejora el desempeño del modelo, resaltando la importancia de evaluar empíricamente la reducción de dimensionalidad.

Los modelos ensamblados, como *Random Forest Regressor* y *XGBRegressor*, son opciones viables, aunque no alcanzaron el nivel de precisión del *SVR*.

Clasificación:

Resumen de los principales resultados:

Al analizar los resultados de los diferentes modelos de aprendizaje automático aplicados al problema de clasificación, se observó que la reducción de variables en general mejoró el rendimiento de los modelos dentro de un mismo algoritmo. Sin embargo, en el caso específico de la clasificación de casos de demencia, los modelos entrenados con 12 variables lograron un desempeño superior en comparación con aquellos que utilizaban solo 2 dimensiones obtenidas mediante PCA. Esto se debe a que un mayor número de variables permite al modelo capturar más información relevante sobre las instancias, lo que se traduce en una mayor precisión en la clasificación.

En los tres algoritmos evaluados, los modelos que utilizaron una selección de atributos mostraron el mejor rendimiento. Específicamente, se alcanzó una exactitud de 1.00 con Bosques Aleatorios y Máquinas de Soporte Vectorial, y de 0.99 con Regresión Logística. Por otro lado, los modelos que incluyeron la totalidad de las características destacaron en su precisión, lo que refleja una mayor proporción de verdaderos positivos entre las predicciones positivas, aunque esto no necesariamente implica una mejora en la identificación de todos los casos positivos reales.

Implicaciones teóricas y prácticas:

La reducción de dimensiones, aunque útil para disminuir el costo computacional y simplificar los modelos, debe evaluarse cuidadosamente según la tarea y el objetivo específico. Una reducción excesiva puede eliminar características relevantes que contribuyen a la correcta clasificación, lo que podría comprometer el rendimiento del modelo. Por lo tanto, es fundamental encontrar un equilibrio adecuado entre simplificar el modelo y preservar la información necesaria para lograr una clasificación precisa.

Conclusiones:

Al comparar los diferentes modelos de cada algoritmo, el más óptimo en todos los casos fue el modelo con selección de características. Además, dado que los puntajes de los modelos reducidos fueron perfectos, se realizó una comparación de métricas entre los modelos que utilizaron la totalidad de las características. Los resultados indicaron que tanto las Máquinas de Soporte Vectorial (SVM) como la Regresión Logística fueron los algoritmos más robustos para comparar entre pacientes con y sin demencia.

Conclusiones de las preguntas planteadas:

Mediante un mero análisis estadístico, se llegó a los siguientes resultados:

1. La demora promedio para obtener una resonancia magnética (RM) es similar entre los grupos: 29,6 días para pacientes sin demencia y 30,6 días para pacientes con demencia, lo que sugiere que no hay una diferencia significativa en la demora de la RM en relación con el estado de demencia.
2. Los pacientes con demencia tienden a obtener puntajes más bajos en pruebas cognitivas, lo cual indica un posible deterioro cognitivo. Además, existen ligeras diferencias en la prevalencia de afecciones como la diabetes y la hipertensión entre los grupos.
3. La prevalencia de demencia es bastante similar entre géneros, con 244 mujeres y 241 hombres en el grupo de demencia.
4. Una proporción significativa de pacientes con demencia (435 de 485) tiene la variante genética APO-E, mientras que un número menor (50 de 515) en el grupo sin demencia posee esta variante, lo que sugiere una fuerte correlación entre la presencia del APO-E y la demencia.
5. Los antecedentes familiares muestran una asociación notable: 255 pacientes con demencia informaron no tener antecedentes familiares, mientras que 230 sí los tenían. Sin embargo, un número considerable de pacientes con demencia tiene antecedentes familiares (290 de 515 sin demencia, 225 con demencia).
6. Un mayor número de exfumadores tiene demencia (252) en comparación con pacientes sin demencia (206), lo que podría indicar una relación entre el tabaquismo previo y el desarrollo de la condición.
7. Los niveles de actividad física ligera o sedentaria son prevalentes en ambos grupos, con una tasa ligeramente mayor de sedentarismo entre los pacientes con demencia.
8. Los pacientes con demencia reportan niveles similares de mala calidad del sueño y tipos de dieta en comparación con el grupo sin demencia, aunque las dietas mediterráneas son un poco más comunes entre los pacientes con demencia.
9. Afecciones como la diabetes, la hipertensión y las enfermedades cardíacas son comunes entre los pacientes con demencia, con una mayor incidencia de diabetes en el grupo de demencia (260) que en el de no demencia (253), lo cual sugiere que estas afecciones pueden influir en el desarrollo de la demencia.

Referencias

[1] EYH Tang, SL Harrison, L Errington, MF Gordon, PJ Visser, G Novak (2015/09/03). Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136181

[2] E. R. Mayeda, C. Shaw, “Invited Commentary: Algorithmic Dementia Classification—Promises and Challenges”, *AJE*, vol. 192, no. 4, pp. 4, Enero 2023.

[3] G. Livingston, A. Sommerlad, V. Orgeta, S. G. Costafreda, J. Huntley, D. Ames (2017/12/16). Dementia prevention, intervention, and care. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)31363-6/abstract?website=main%3Fpostid%3D128582%3Fmemberid&postid=128582&parentid=0&memberid=

[4] D. Newby, V. Orgeta, C. R. Marshall, I. Lourida, C. P. Albertyn, S. Tamburin, V. Raymont, M. Veldsman, I. Koychev (2023/10/14). Artificial intelligence for dementia prevention. Available: https://alz-journals.onlinelibrary.wiley.com/doi/full/10.1002/alz.13463

[5] Keystone Clinic & Surgery (2023/06/13). Dementia. Available: https://keystonemedical.com.sg/dementia/

[6] “Inteligencia Artificial y Aprendizaje Automático”, class notes for TAD-12, UNRaf, 2do cuatrimestre 2024.

[7] T. Amr, *Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits*. Birmingham: Packt Publishing Ltd., 2020

[8] Scikit-learn. Available: https://scikit-learn.org/stable/