Metodología para Ponderar la Importancia de Variables no Interpretables mediante Variables Interpretables.

Autor: **Pablo Verde, Lluís Llull, Pablo Martín**

Institución: **Universidad Politécnica de Madrid**

Fecha: January 8, 2025

Abstract

Esta investigación presenta un enfoque pionero para mejorar la interpretabilidad de los modelos de aprendizaje automático que dependen de variables no interpretables. Al crear vínculos entre las variables no interpretables y las interpretables, este marco permite a los usuarios comprender mejor y confiar en las predicciones y decisiones tomadas por estos modelos. Utilizando técnicas avanzadas como los valores SHAP y la importancia de las características, la metodología evalúa el impacto de las variables no interpretables y lo convierte en métricas interpretables mediante estrategias de ponderación basadas en correlaciones. El planteamiento propuesto se ha validado en múltiples conjuntos de datos, demostrando su capacidad para equilibrar la interpretabilidad y la precisión del modelo sin sacrificar el rendimiento. Este estudio pone de relieve la importancia de combinar la precisión técnica con la transparencia orientada al usuario, especialmente en áreas críticas como la sanidad y las finanzas. El trabajo futuro tiene como objetivo ampliar esta metodología a los modelos de aprendizaje profundo y otros campos de alta dimensión, allanando el camino para una adopción más amplia y una mayor confianza en los sistemas de IA.

Contents

List of Figures

List of Tables

Thanks

1 Introduction

1.1 SOTA

En campos críticos como el médico, la interpretabilidad y explicabilidad de los modelos de Inteligencia Artificial son esenciales para generar confianza [1] [2]. Los modelos "black box", como las redes neuronales, ofrecen predicciones sin que se comprenda fácilmente su funcionamiento, lo que puede generar incertidumbre y desconfianza, especialmente en contextos donde las decisiones tienen un impacto significativo, como en diagnósticos médicos o decisiones judiciales. Por otro lado, los modelos "glass box" son inherentemente más transparentes, lo que facilita la comprensión de su comportamiento, pero a menudo con el costo de tener un rendimiento inferior en tareas complejas [3].

La interpretabilidad se refiere a la facilidad con que un modelo puede entenderse directamente por los humanos, mientras que la explicabilidad implica las técnicas que permiten comprender los resultados y la lógica detrás de las decisiones del modelo. Estas técnicas no solo ayudan a los desarrolladores a verificar que los modelos se comportan de la manera esperada, sino que también permiten a los usuarios finales, como médicos o jueces, confiar en las decisiones automatizadas.

Técnicas como SHAP [4]y feature importance [5] en árboles de decisión permiten descomponer las contribuciones de las variables a las predicciones y brindan una forma de entender cómo cada característica influye en la salida del modelo, lo cual es fundamental para garantizar la transparencia. Sin embargo, estas técnicas tienen limitaciones, especialmente en modelos complejos y en contextos donde las variables no son directamente interpretables por el usuario final. En modelos como redes neuronales profundas, los resultados de estas técnicas pueden ser difíciles de interpretar si las variables de entrada no tienen una relación clara con conceptos humanos.

Además, se utilizan técnicas estadísticas para analizar correlaciones entre variables [6]. Sin embargo, en modelos de alta dimensionalidad, donde las variables pueden estar altamente correlacionadas o ser abstractas (como en el caso de las representaciones vectoriales de palabras en NLP o las características extraídas de imágenes), la dificultad para interpretar el modelo aumenta.

1.2 Motivación

El rápido avance en el uso de modelos de IA en diversos campos plantea un gran desafío: lograr interpretabilidad y confianza en un entorno donde surgen decisiones que afectan a aspectos importantes, como la salud, la seguridad o el negocio. Si bien las variables interpretables permiten a los usuarios finales comprender y validar las predicciones,

los modelos a menudo se basan en variables no interpretables, como índices compuestos multidimensionales, coeficientes de modelos avanzados, indicadores de simulación o modelado, entre otros. Aunque estas variables son esenciales para el desempeño del modelo, son difíciles de entender y, por lo tanto, de aceptar.

La principal motivación de este trabajo radica en la necesidad de cerrar la brecha entre la precisión técnica y la interpretabilidad. Se propone desarrollar un método para vincular variables no interpretables con variables interpretables, creando una perspectiva más clara sobre cómo estas últimas pueden usarse como puente para interpretar los componentes menos comprensibles del modelo.

Esta necesidad se vuelve importante en aplicaciones como la medicina, donde un modelo muy preciso, pero vago puede ser rechazado porque no justifica claramente sus decisiones, o en campos como las finanzas, donde se necesitan explicaciones fáciles de entender para aplicar predicciones. Además, este enfoque no sólo busca mejorar la transparencia, sino que también explora nuevas estrategias para mejorar el uso de variables en modelos multidimensionales, mejorando el poder explicativo sin perjudicar al desempeño.

En definitiva, este trabajo surge del reto de crear metodologías relacionadas con la Inteligencia Artificial que no sólo sean eficaces, sino también comprensibles, que permitan integrarse en los procesos de toma de decisiones de forma responsable y fiable.

2 Cuerpo

2.1 Metodología

Primero de todo, es necesario definir los siguientes dos conceptos:

- **Datos interpretables:** Conjunto de datos los cuales suponemos que el usuario que va a utilizar el modelo entiende.
- **Datos no interpretables:** Conjunto de datos los cuales suponemos que el usuario que va a utilizar el modelo **NO** entiende.

Vamos a diferenciar el algoritmo en dos diferentes algoritmos que se basan en la misma idea pero con funcionalidades diferentes. Hay que tener en cuenta que varios algoritmos tienen partes que son iguales y en general son muy similares. Únicamente cambia la forma en la que obtenemos la contribución de cada variable al modelo.

- **Algoritmo de importancia:** versión que utiliza la *feature importance* para posteriormente ponderar estos valores en función de las correlaciones.
- **Algoritmo explicativo:** esta versión utiliza métodos explicativos como *SHAP* para posteriormente ponderar estos valores en función de las correlaciones.

Para el funcionamiento de esta metodología vamos a suponer que ya disponemos de un modelo, en este modelo podemos diferenciar dos casos:

- **Caso 1:** Modelo del cual se puede obtener *feature importance* - Como por ejemplo modelos basados en árboles (*Random Forest* o *XGBoost*) o modelos lineales (*Linear Regression* o *Logistic Regression*). En este caso podemos usar las dos opciones del algoritmo, tanto el **Algoritmo de importancia** o el **Algoritmo explicativo**.
- **Caso 2:** Modelo del cual no podemos obtener *feature importance* - En estos casos, con modelos como *SVM*, *KNN* o *Redes Neuronales* tan solo podremos realizar la versión de **Algoritmo explicativo**.

En ambos casos vamos a disponer del dataset usado para entrenar el modelo, vamos a suponer los siguientes casos:

- Hay un total de N features diferentes
- Hay un total de E *datos interpretables*. S es variable y arbitrario, es decir, nosotros podemos decidir que número(E) de features vamos a considerar no interpretables. Es más, si bien es cierto que el objetivo es que en el conjunto de datos que pertenecen

a E sean estrictamente todos los datos no interpretables, se pueden incluir datos interpretables en el conjunto E. Es importante añadir, que se supone que como menor sea E, más fiel será la representación obtenida por la metodología a la realidad. Aun así, se va a experimentar con este conjunto E para sacar conclusiones.

- Hay un total de S *datos interpretables*, donde $S = N - E$

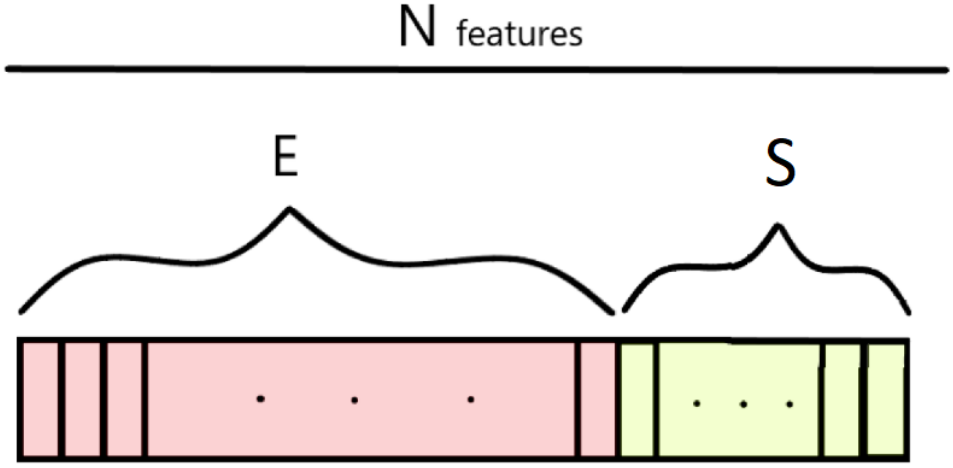

Figure 1: Esquema de datos input
Own elaboration

La metodología se compone de las siguientes fases.

- Selección y entrenamiento del modelo para la metodología.
- Clasificar los N features en los conjuntos E y S
- Aplicar uno de los dos algoritmos. *Algoritmo de correlación* o *Algoritmo de explicación*
- Evaluación de los resultados obtenidos

2.2 Explicación del algoritmo

2.2.1 Sistemas de Contribución

FEATURE IMPORTANCE

El *Feature Importance* es una técnica que evalúa la relevancia de cada variable en el modelo de aprendizaje automático, proporcionando una medida cuantitativa de cuánto

contribuye cada característica al rendimiento o precisión del modelo. Este método es compatible tanto con modelos interpretables (e.g., árboles de decisión) como con modelos más complejos (e.g., Random Forest o XGBoost)[5].

Definición Formal

Existen varios enfoques para calcular la importancia de características:

1. Basada en la Reducción de Impureza

En modelos basados en árboles, la importancia de una característica i se calcula como la reducción acumulada de la impureza (ΔI) en todos los nodos donde participa la característica:

$$FI(i) = \sum_{t \in T} \Delta I(t) \cdot \mathbb{1}\{\text{feature } i \text{ en } t\},$$

donde:

- T: conjunto de nodos del árbol.
- $\Delta I(t)$: reducción de impureza (e.g., Gini o entropía) en el nodo t.
- $\mathbb{1}\{\cdot\}$: indica si la característica i se usó en el nodo t.

2. Basada en Permutaciones

Este enfoque evalúa el impacto de permutar los valores de una característica en el conjunto de datos de prueba. La importancia se mide como la diferencia en la métrica de desempeño:

$$FI(i) = \text{score\_original} - \text{score\_permutado},$$

donde:

- score_original: precisión del modelo antes de la permutación.
- score_permutado: precisión después de permutar la característica i.

Coeficientes en Modelos Lineales

En modelos lineales, los coeficientes asociados a las características proporcionan una medida directa de su importancia. Estos valores pueden normalizarse para comparar variables con diferentes escalas.

SHAP

SHAP utiliza los valores de Shapley de la teoría de juegos para asignar importancias justas y consistentes a las variables que contribuyen a una predicción específica. Este enfoque evalúa el impacto de cada variable considerando todas las posibles combinaciones de las restantes[4].

El valor de Shapley para una variable i en un punto x se define como:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right],$$

donde:

- N: conjunto de todas las variables.
- S: subconjunto de variables excluyendo i.
- $f(S)$: predicción del modelo considerando únicamente las variables en S.

Los valores de Shapley poseen las siguientes propiedades deseables:

- **Consistencia:** Si la contribución marginal de una variable i aumenta en cualquier subconjunto S, el valor de ϕ_i también aumenta.
- **Eficiencia:** La suma de los valores de Shapley de todas las variables equivale al cambio total en la predicción.

El modelo explicativo de SHAP puede representarse como:

$$g(x) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i,$$

donde:

- ϕ_0: valor base (predicción promedio del modelo en el conjunto de referencia).
- ϕ_i: contribución de la variable i a la predicción.

2.2.2 Algoritmo

1. **Paso 1 -** Obtención de la 'contribución' para todos los valores, tanto los valores interpretables (S) como los no interpretables (E). Según se ha explicado en el apartado anterior. Nombraremos a esta contribución como $con(x)$
2. **Paso 2 -** Generación de la Matriz de correlaciones

   El segundo paso será generar la **Matriz de correlación** de E enfrente de S. Se van a pasar todas las correlaciones a valor absoluto. La tabla de correlaciones obtenida tendrá la siguiente forma:

	S1	S2	...	Sm
E1	c_{11}	c_{12}	...	c_{1m}
E2	c_{21}	c_{22}	...	c_{2m}
...	...	...	...	...
En	c_{n1}	c_{n2}	...	c_{nm}

Table 1: Example of correlation matrix obtained

Se nombrarán los valores obtenidos como c_{ij} donde i será el número del valor no interpretable (E) y j el del valor interpretable (S).

3. **Paso 3 -** Obtención de valores auxiliares s' En el siguiente pasó se generarán unos valores auxilares para todo $s \in S$, que los nombraremos como s'. Estos valores se calcularán de la siguiente forma:

$$\forall s_j \in S, s'_j = \sum_{e_i \in E}^{n} c_{ij} * con(e_i)$$

Es decir, s' será el sumatorio de los productos de los valores de contribución de las variables no interpretables con la correlación de dicha variable no interpretable con la variable interpretable que se está 'evaluando'. En resumen, **s'** será una nueva forma de aproximar la importancia de las variables representadas únicamente con las variables interpretables, ponderando todos los valores de $e \in E$.

4. **Paso 4- Combinación de s y s'** Para ser más precisos para determinar la importancia de $s \in S$ será necesario, combinar los valores obtenidos s' con los valores originales s. Nombraremos a esta nueva aproximación definitiva como S^*. Se han propuesto diferentes formas para obtener S^*, que se resumen en la siguiente tabla:

Método	**Fórmula**	**Explicación**
Suma simple	$s^* = s' + s$	Suma simple de los dos valores.
Suma ponderada	$s^* = \alpha \cdot s' + (1 - \alpha) \cdot s$	Consiste en realizar una suma ponderada, es decir, que cada uno de los valores tiene un peso diferente en la operación.
Multiplicación	$s^* = s' \cdot s$	Consiste en multiplicar los valores.
Multiplicación Aumentada	$s^* = s' \cdot (1 + s)$	Consiste en primero sumar 1 a todos los valores de s prima y posteriormente multiplicar los valores. Así aseguramos que el resultado nunca sea menor que el propio resultado original.
Escalar valores	$s^* = scale(s') \, op \, scale(s)$	Se basa en realizar cualquiera de las operaciones comentadas anteriormente pero realizando un escalado previo de los valores.

Table 2: Comparación de estrategias

5. **Output**

Se obtendrá S^*, es decir, una lista de tamaño S con la nueva importancia ponderada de cada valor $s \in S$.

3 Experimentación

3.1 Experimento 1 - Parametrización y depuración del algorítmo

Se va a utilizar este experimento para depurar el algoritmo, nos servirá para seleccionar el mejor método de combinación y para terminar de ajustar el funcionamiento del algoritmo.

Para este algoritmo se va a utilizar el dataset de predicción de calidad del vino. [7]. Un dataset con 1600 muestras de vino tinto y blanco del norte de Portugal. El objetivo es modelar la calidad del vino en función de pruebas fisicoquímicas. Incluye las siguientes variables:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality (target, score entre 0 i 10)

Para este experimento vamos a utilizar dos modelos, se exponen las métricas en la siguiente tabla, aun así, estas métricas se exponen tan solo para demostrar la validez de dichos modelos y no van a repercutir directamente en la metodología:

Model	**RS2**	**MAE**	**MSE**
Linear Regression	0.40	0.5	0.39
Random Forest	0.55	0.42	0.39

Table 3: Modelos experimento 1

Se va a probar los diferentes sistemas de combinación comentando sus pros y contras hasta encontrar el sistema óptimo para la metodología. Se va a utilizar la variable *total sulfur dioxide* como la variable no interpretable:

Suma

En la suma existe el problema que es muy vulnerable a las magnitudes de los dos valores. Por ejemplo, en el siguiente escenario, donde *S prima* es prácticamente inexistente al compararlo con los valores de S original. Este método en la mayoría de casos no es útil.

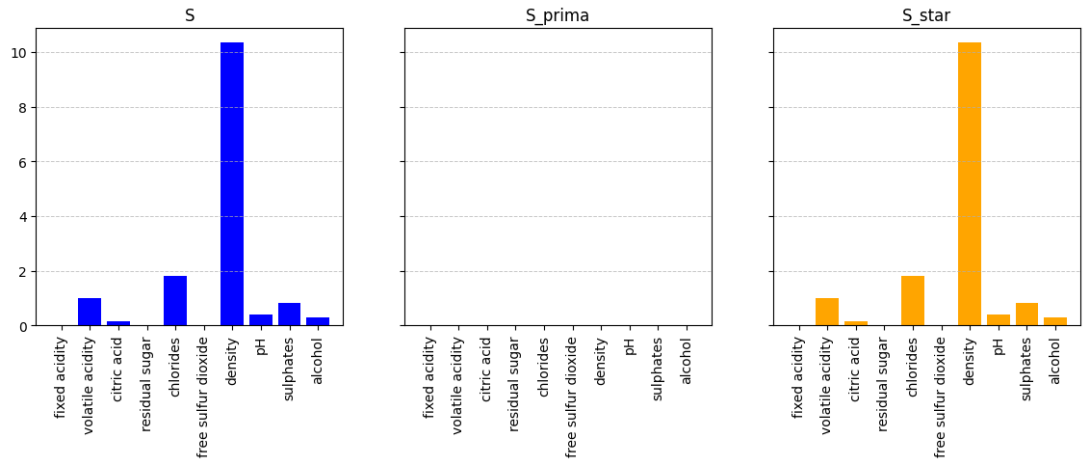

Figure 2: Resultados suma
Own elaboration

Aquí vemos en más detalle los resultados de *S prima*, pero se ha observado como no influyen a S^*

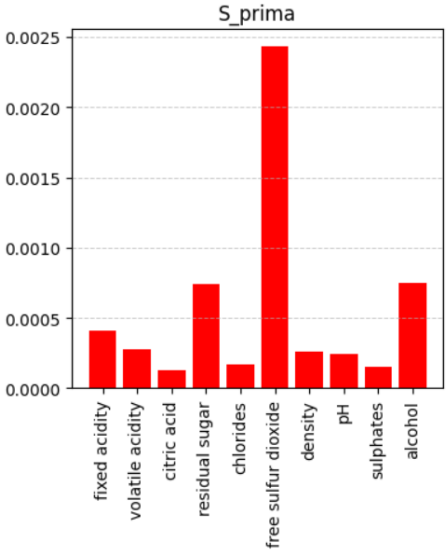

Figure 3: S prima
Own elaboration

Operaciones ponderadas

Con este approach evitamos el problema que teníamos anteriormente de las diferencias de magnitud. Aun así, el hecho de añadir un parámetro α añade complejidad al algoritmo. Además, el hecho de no seleccionar un parámetro adecuado podría hacer que el algoritmo no sea funcional. Todo esto se suma al hecho de que el parámetro α puede ser completamente distinto dependiendo de cada experimento, dataset, modelo, etc. Por estos motivos este método también se descarta.

Suma escalada

Con este approach podemos evitar el problema de las magnitudes comentado previamente en la suma. Aun así, aparecen una serie de escenarios que no nos resultan útiles para nuestra metodología. No solo eso, sino que este método puede causar malinterpretaciones y resultados erróneos.

Por ejemplo exponemos el siguiente caso, dado el modelo de *Linear Regresion*, obtenemos los siguientes valores de *feaeture importance*

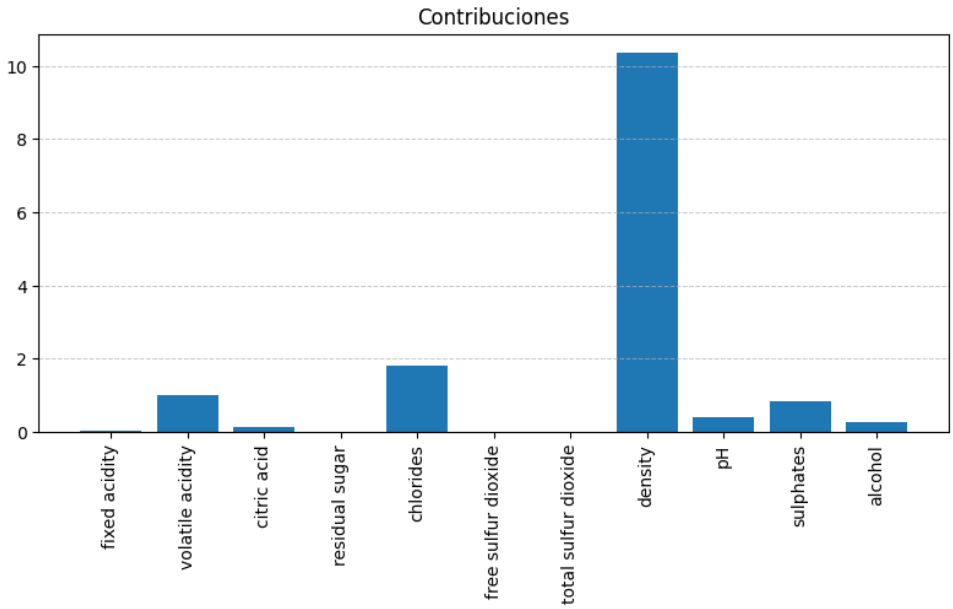

Figure 4: Correlaciones modelo de LR
Own elaboration

Destaca que para este modelo, las variables *total sulfur dioxide y free sulfur dioxide* no disponen de ninguna importancia. Aun así, en la siguiente tabla vemos como el resultado de S^* muestra que la variable *free sulfur dioxide* es muy importante, cuando realmente esto no es cierto.

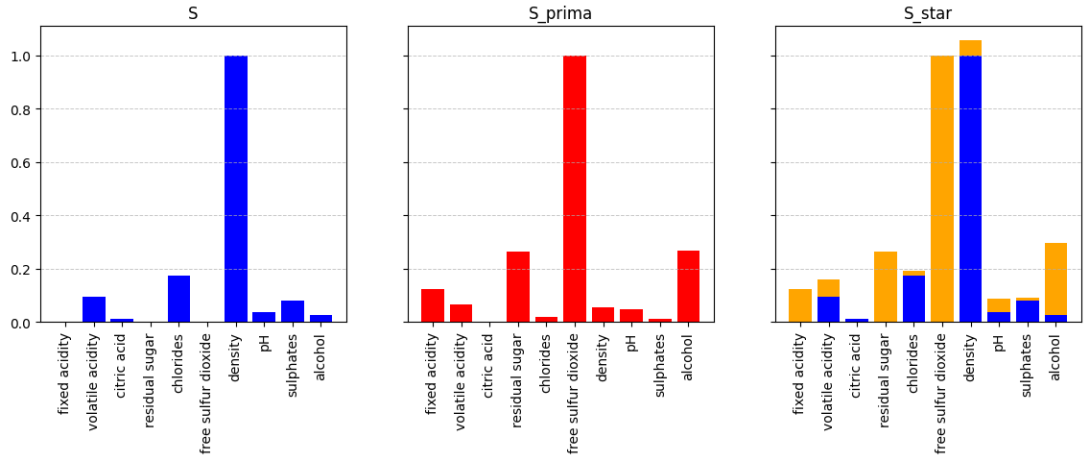

Figure 5: Resultados suma escalando
Own elaboration

Esto sucede ya que en *S prima free sulfur dioxide* adquiere un valor muy alto, ya que tiene una gran correlación con *total sulfur dioxide*, hecho lógico al ser las dos variables relacionadas con el dióxido de sulfuro. Al realizar la suma, no estamos sumando las mismas unidades, por una parte, tenemos la importancia directa de la variable y, por otro lado, la importancia correlacionada con la variable no interpretable. Es decir, esta *S prima* nos debe servir para ponderar la S original, realizar una suma directamente, altera los resultados y no se obtiene el resultado que se busca. Por este mismo motivo, este sistema también es descartado.

Multiplicación

Como hemos comentado en el último apartado, *S prima* busca ponderar el resultado de *S*, la multiplicación nos puede servir para conseguir esto. Sin embargo, en *S prima* obtenemos por norma general valores muy próximos a 0 (tanto escalando como sin escalar), esto provoca que los resultados finales obtenidos en (S^*) sean prácticamente inexistentes, indicando que las variables no tienen importancia, cuando esto no es cierto. Aquí vemos dos casos que ejemplifican el hecho comentado.

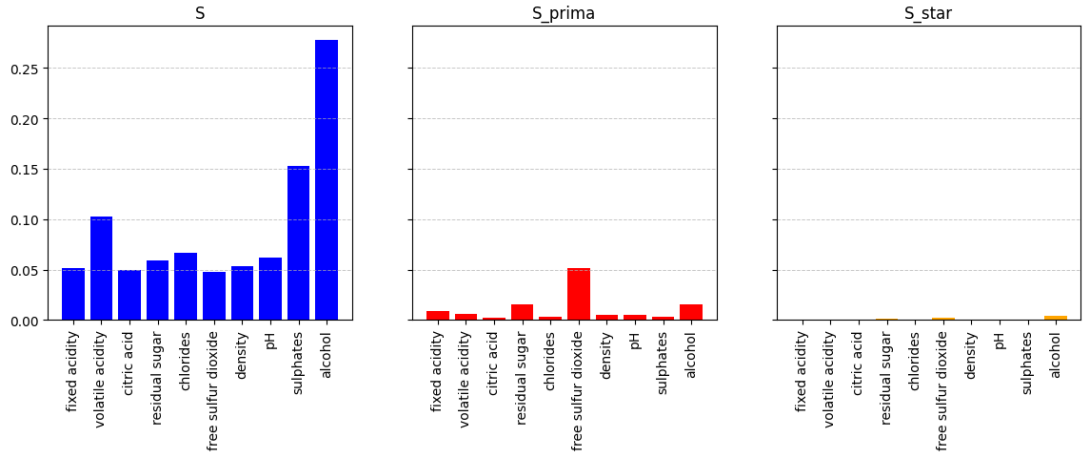

Figure 6: Resultados multiplicación
Own elaboration

A priori, parece que escalando se obtienen mejores resultados, no solo porque nos permite representar y comparar de una forma más sencilla, sino también, porque de esta forma no creamos modificaciones tan exageradas al multiplicar.

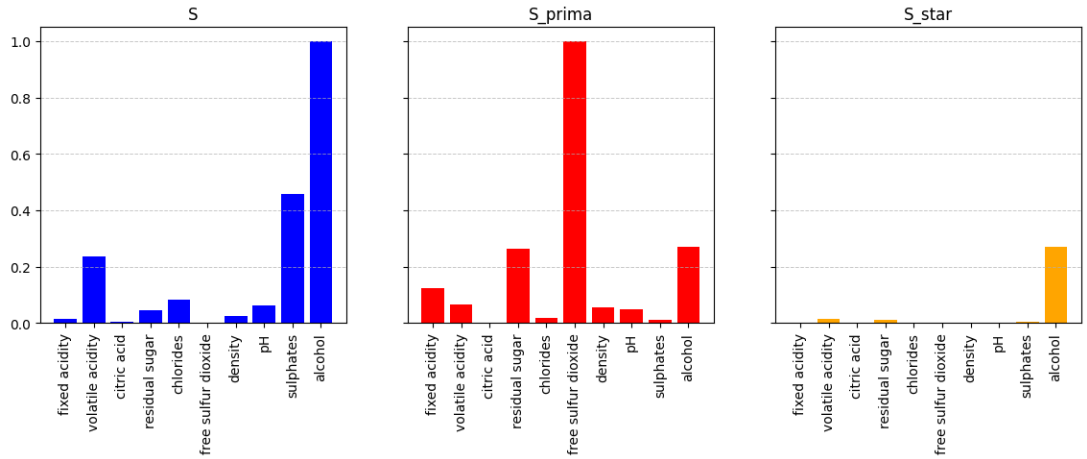

Figure 7: Resultados multiplicación escalada
Own elaboration

De todas formas, al escalar aparece un problema. Queda muy bien ejemplificado en el siguiente caso, si volvemos a fijarnos en la figura 4 vemos como la variable *fixed acidity* no tiene prácticamente peso al definir la importancia de cada variable, aun así, en la siguiente figura vemos como la variable de *density* aumenta considerablemente, cuando realmente esto no debería ser así, ya que *fixed acidity* no debería ponderar importancia a las otras variables. Esto es causado por el motivo de escalar la *S prima*, si escalamos este conjunto tenemos el problema de que la variable más grande (por pequeña que sea) se le otorga mucha importancia (representado en naranja la importancia ponderada). Por eso se decide tan solo escalar la *S original*, vemos aquí los nuevos resultados:

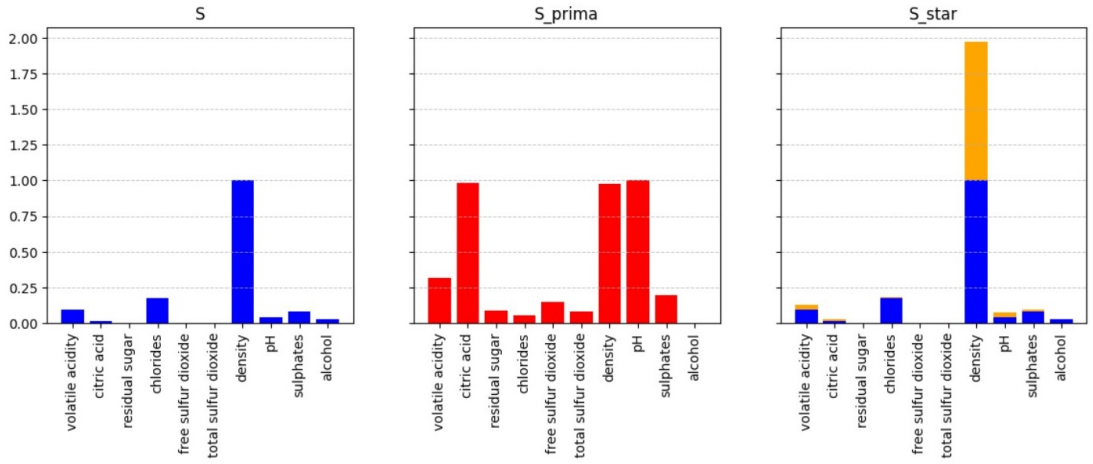

Figure 8: Problema multiplicación escalada
Own elaboration

Por este motivo, se ha decidido tan solo escalar la *S original* y no escalar *S prima*, así conseguimos evitar este problema

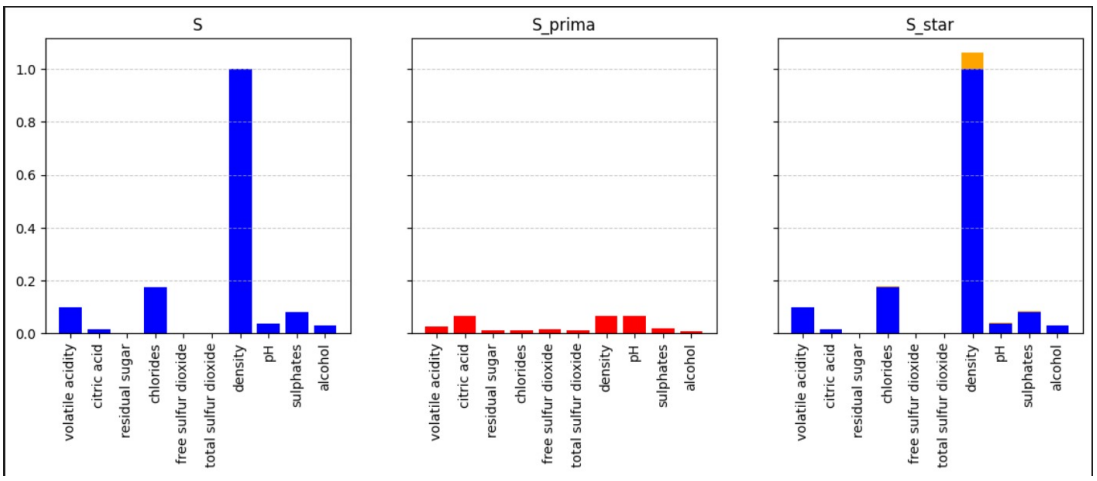

Figure 9: Problema solucionado multiplicación escalada
Own elaboration

Se puede considerar que la filosofía de la ponderación mediante una multiplicación puede servirnos, sin embargo, hay que solucionar este problema comentado de la proximidad a 0 de los valores.

Multiplicación aumentada

Finalmente, para solucionar el problema anterior (valores demasiado próximos al 0), se usa este método. Vamos a usar directamente el sistema de tan solo escalar la *S original*, por los motivos expuestos anteriormente.

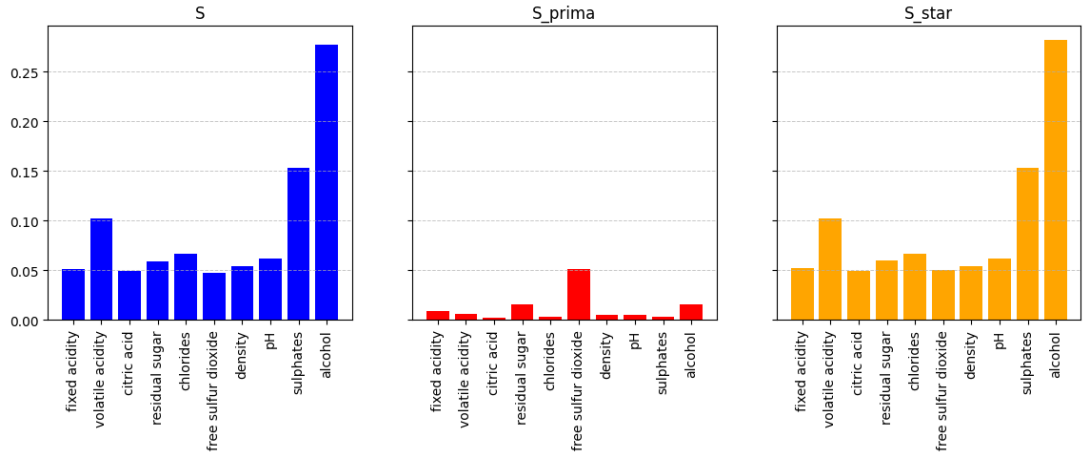

Figure 10: Resultados multiplicación aumentada
Own elaboration

En las siguientes figuras vemos los resultados de la suma aumentada tanto para el modelo de *Random Forest* como *Linear Regression*. Vemos que con este sistema evitamos todos los problemas comentados anteriormente y conseguimos realmente el objetivo que se busca para el sistema de contribución, que se trata de representar la importancia de la variable no interpretable en función de las otras variables representables

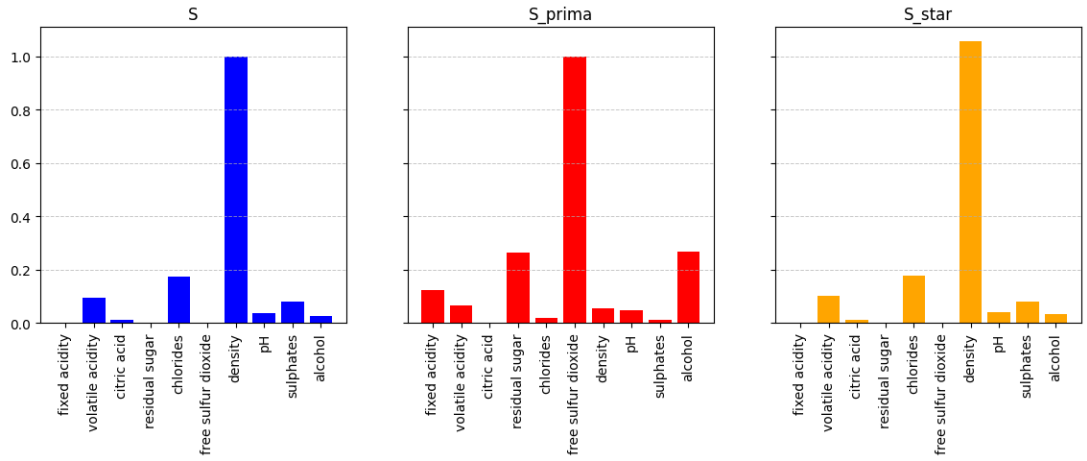

Figure 11: Resultados multiplicación aumentada escalando LR
Own elaboration

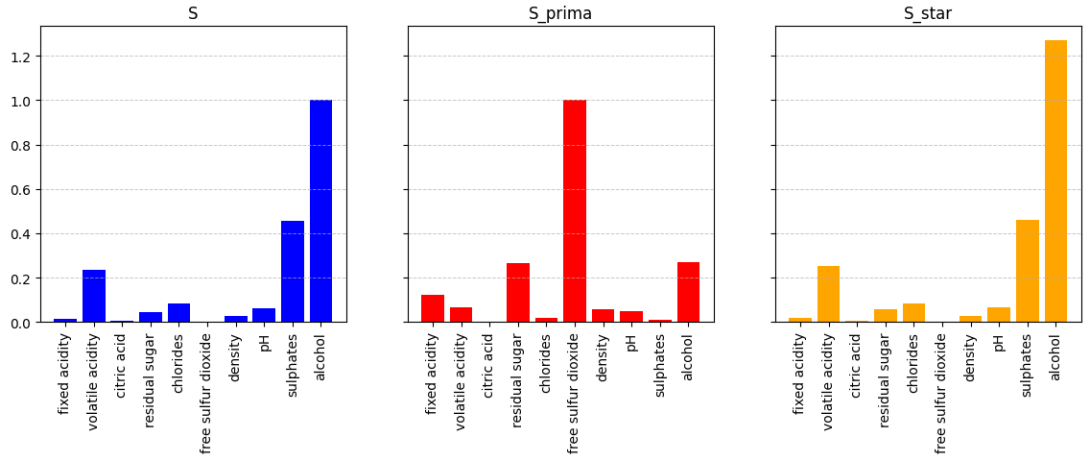

Figure 12: Resultados multiplicación aumentada escalando RF
Own elaboration

En definitiva, vemos como este método de combinación es con el que obtenemos mejores resultados y la metodología resulta más realista.

3.2 Experimento 2

En el segundo experimento, emplearemos diferentes sistemas de contribución para averiguar cuál ofrece una mejor representación de la importancia de las variables mencionadas en la sección 2.2.1 2.2.1. En particular, utilizaremos el feature importance y los valores *SHAP*.

Para ello, se van a emplear dos datasets diferentes: el dataset de vinos utilizado en el anterior experimento 3.1 y el dataset de diabetes. Este ultimo contiene datos de mujeres de una región de la India con y sin diabetes. El objetivo de este dataset es predecir si un paciente tiene diabetes, basándose en ciertas mediciones diagnósticas.

El dataset cuenta con las siguientes variables:

- Pregnancies: Número de embarazos
- Glucose: Nivel de glucosa en sangre
- BloodPressure: Presión sanguínea
- SkinThickness: El grosor de la pie
- Insulin: Nivel de insulina en sangre
- BMI: Índice de masa corporal
- DiabetesPedigreeFunction: Función que determina el nivel de riesgo de tener diabetes de tipo 2 basado en el historial familiar

- Age: Edad
- Outcome: Resultado a predecir (1 tiene diabetes, 0 no tiene diabetes)

Para este experimento se va a utilizar el método de combinación de multiplicación aumentada expuesto en el apartado 3.1 2, ya que este es el mejor método encontrado.

En este experimento se va a evaluar el uso de los sistemas de contribución en varios escenarios, siendo estos en el que una variable no interpretable tiene una alta contribución evidente en el resultado final, cuando una variable no interpretable tiene una baja contribución evidente en el resultado final, y cuando existen múltiples variables no interpretables que tienen una alta contribución en el resultado final.

Además, para el dataset de los vinos se va a usar un modelo de regresión de RandomForestRegressor, mientras que, para el dataset de diabetes, se usará un modelo de clasificación de RandomForestClassifier. Tanto las metricas obtenidas con estos modelos, como los valores SHAP y feature importance obtenidos sobre estos datasets se exponen a continuacion:

Model	RS2	MAE	MSE
Random Forest	0.55	0.42	0.39

Table 4: Modelo experimento 2 Dataset de vino

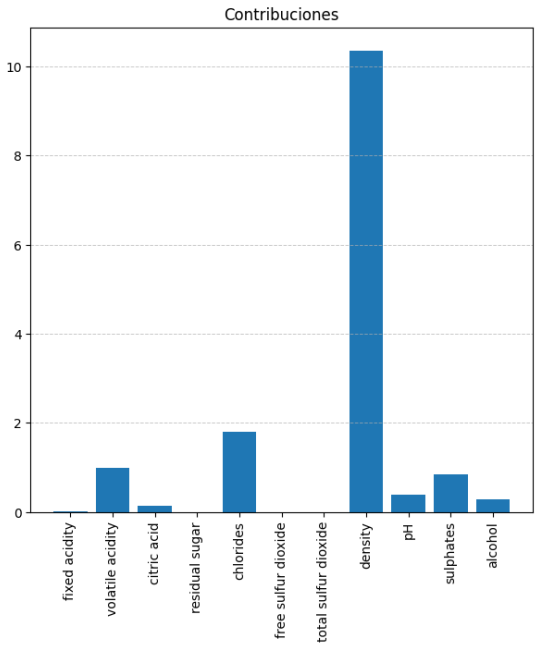

Figure 13: Valores de feature importance obtenidos en el dataset de vinos
Own elaboration

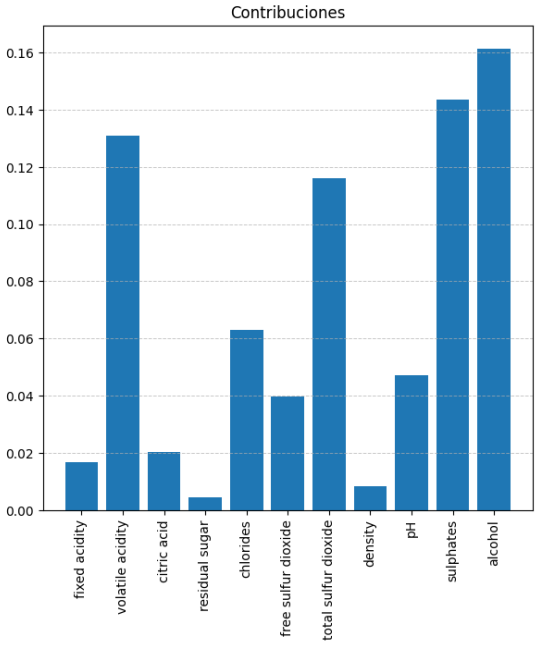

Figure 14: Valores SHAP obtenidos en el dataset de vinos
Own elaboration

Class	**precision**	**recall**	**f1-score**
0	0.80	0.83	0.81
1	0.67	0.62	0.64

Table 5: Modelo experimento 2 Dataset de diabetes

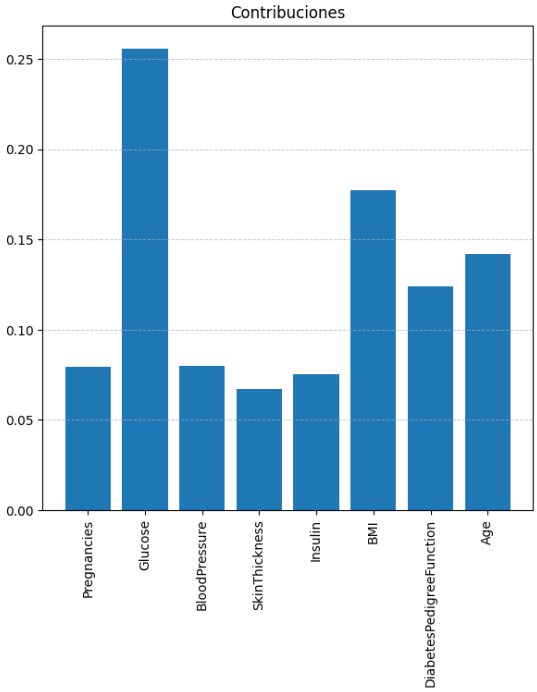

Figure 15: Valores de feature importance obtenidos en el dataset de diabetes
Own elaboration

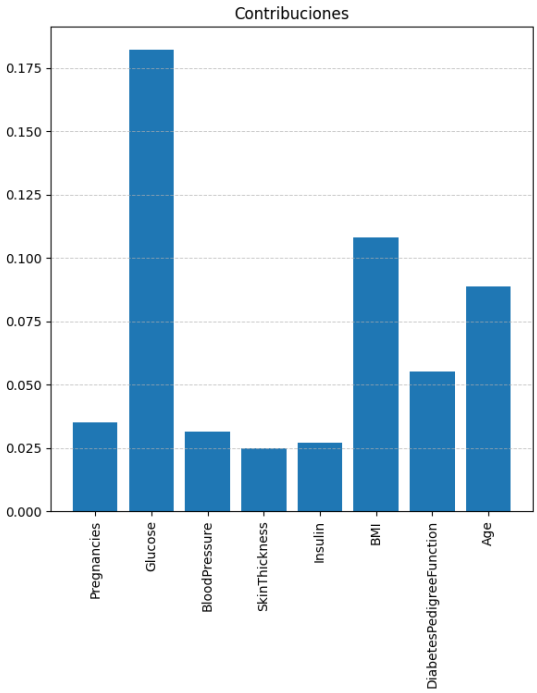

Figure 16: Valores SHAP obtenidos en el dataset de diabetes
Own elaboration

3.2.1 Variable no interpretable tiene una alta contribución evidente en el resultado final

Dataset de vinos

En los vinos, la densidad influye bastante en su calidad, debido a que es una medida la fermentación del vino, así como de los componentes disueltos en este como azucares o ácidos. Por ello la variable *density* debería ser bastante importante para predecir la calidad del vino.

Utilizando feature importance

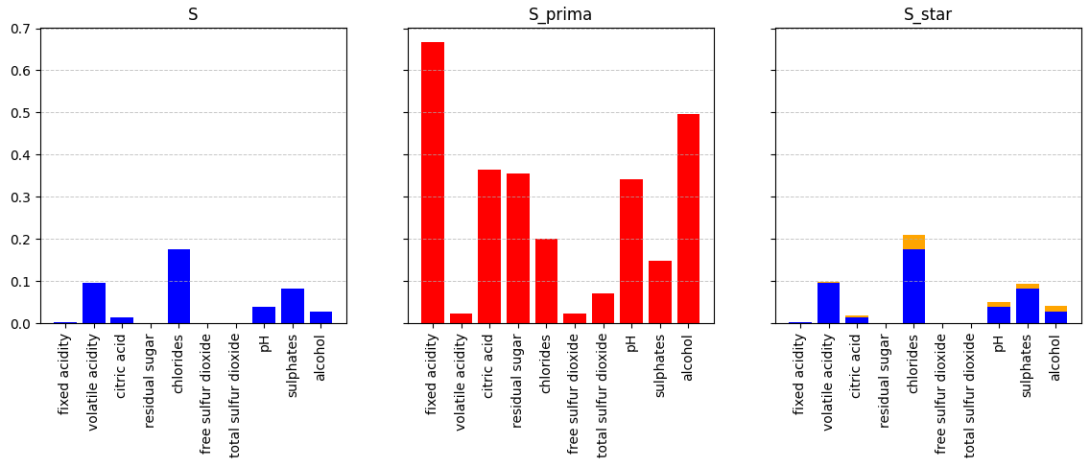

Figure 17: Resultados de la calidad del vino usando feature importance para la variable no interpretable densidad

Own elaboration

Utilizando valores SHAP

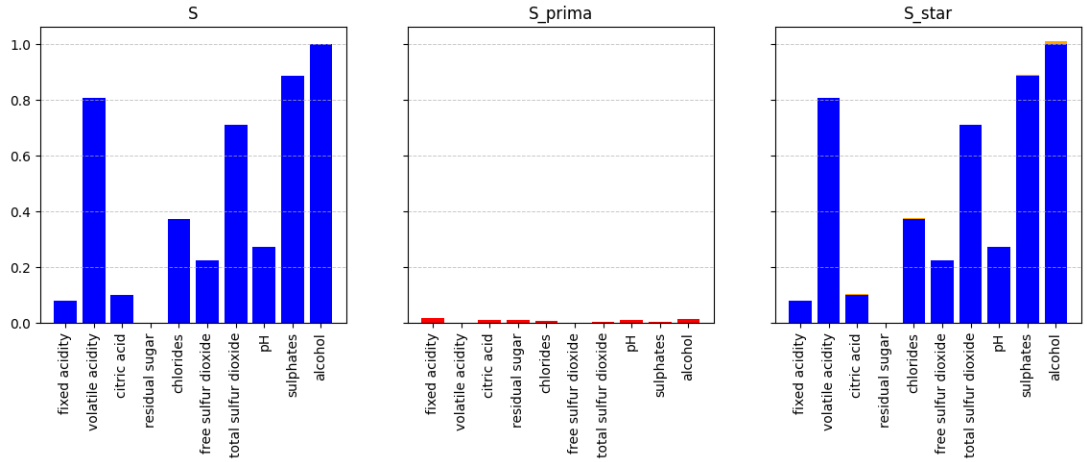

Figure 18: Resultados de la calidad del vino usando valores shap para la variable no interpretable densidad

Own elaboration

La importancia de las variables que asigna el feature importance y los valores SHAP varian bastante como puede obvervarse en estas graficas 13 14. Desde cierto punto de vista ambos pueden tener la razon, ya que el modelo puede estar aprendiendo a predecir los valores en base a la densidad del vino y los cloritos, o bien, puede estar utilizando el resto de variables que afectan a la densidad del vino, como lo son el alcohol o la acidez.

Si comparamos las graficas del feature importance con la de los valores SHAP, observamos que los valores de feature importance de *S prima* son mayores, haciendo que el incremento de valores de *S star* en comparacion a *S* sea mayor.

Dataset de diabetes

La cantidad de glucosa en sangre influye bastante para determinar si una persona tiene o no diabetes, debido a que genera un problema con la regulación de la glucosa, lo que causa niveles anormalmente altos en sangre. Por ello la variable *Glucose* debería ser bastante importante para predecir si un paciente tiene o no diabetes.

Utilizando feature importance

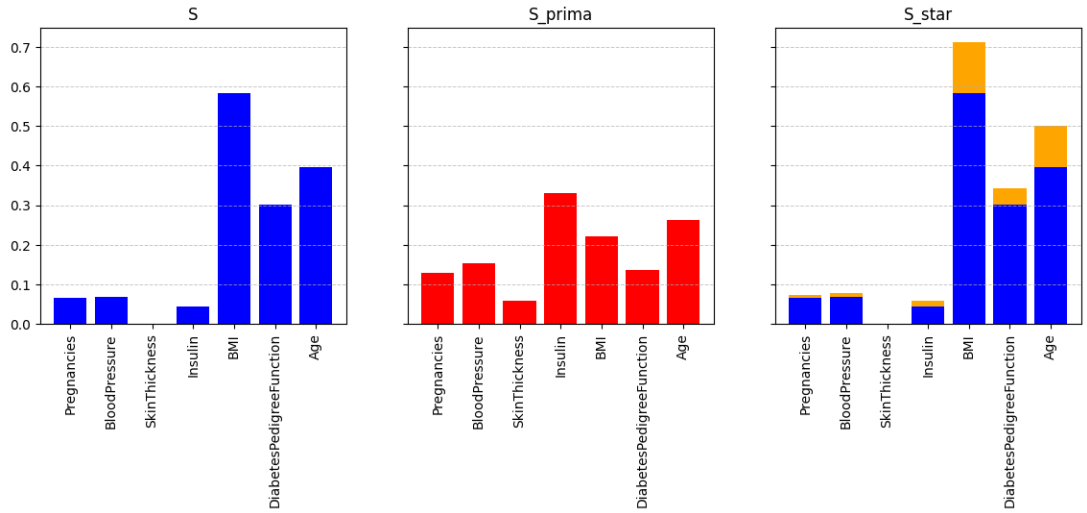

Figure 19: Resultados la predicción de diabetes usando feature importance para la variable no interpretable Glucose

Own elaboration

Utilizando valores SHAP

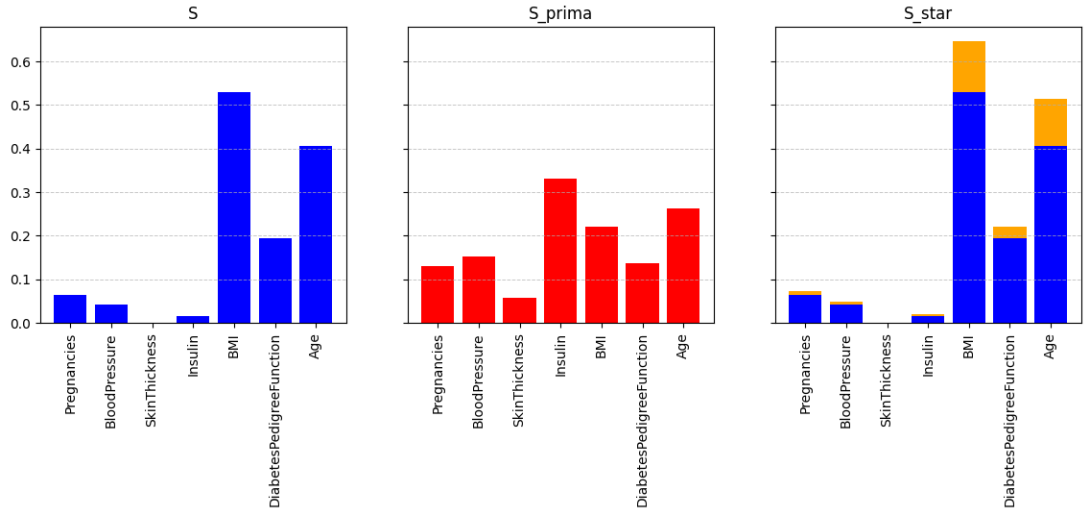

Figure 20: Resultados de la predicción de diabetes usando valores shap para la variable no interpretable Glucose

Own elaboration

En este dataset obtenemos valores similares de los cuales podemos sacar conclusiones mas elaboradas. De esta forma, los valores *S star* incrementan bastante al utilzar una variable no interpretable importante, como es el caso de la glucosa (Glucose). La pequeña diferencia entre estos 2 métodos, es que segun el feature importance el modelo le esta dando mas importancia al BMI y a la variable DiabetesPredigreeFunction, mientras que los valores SHAP le estan dando menos. Esto tiene bastante sentido, ya que las contribuciones originales de estos valores siguen la misma tendencia 15 16.

3.2.2 Variable no interpretable tiene una baja contribución evidente en el resultado final

Dataset de vinos

En los vinos, la cantidad de azúcar residual no suelen afectar mucho a la calidad de los vinos en comparación con otros factores. Por ello la variable *residual sugar* no debería ser muy importante para predecir la calidad del vino. Esto tambien puede observarse en los valores de feature importance y SHAP que obtiene la variable originalmente 13 14, donde tiene un resultado muy bajo. Por ello vamos a utilizar esta variable como variable no interpretable para esta prueba.

Utilizando feature importance

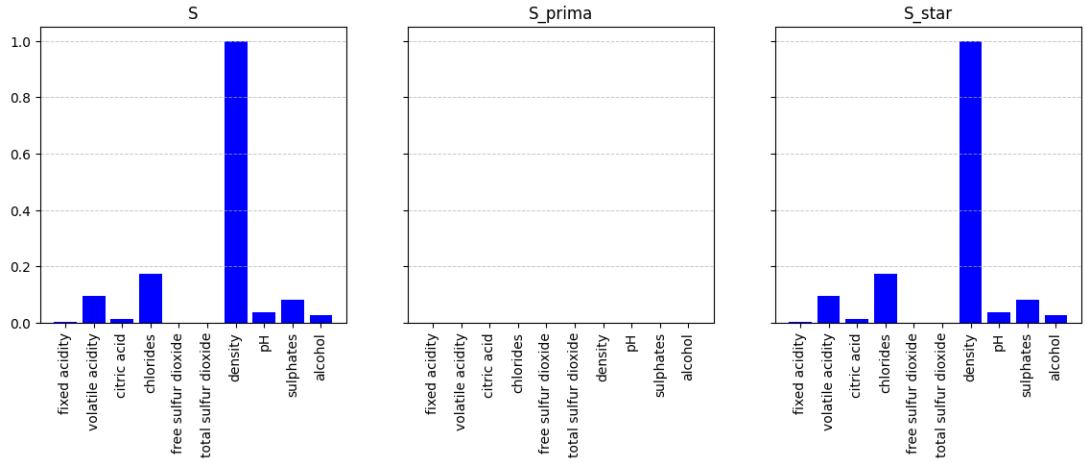

Figure 21: Resultados de la calidad del vino usando feature importance para la variable no interpretable residual sugar

Own elaboration

Utilizando valores SHAP

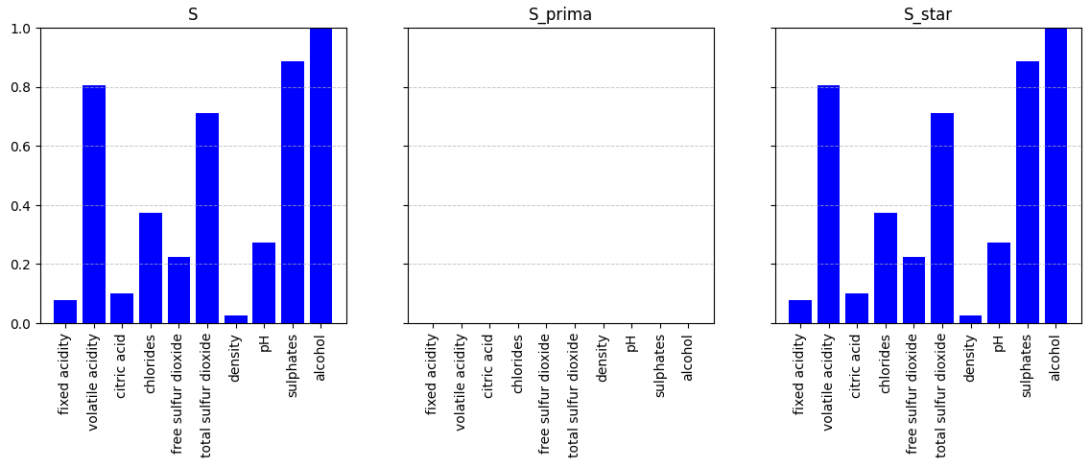

Figure 22: Resultados de la calidad del vino usando valores shap para la variable no interpretable residual sugar

Own elaboration

Al tener una variable no interpretable que no tiene sentido que el modelo le de una gran importancia, se puede observar como tanto el uso de feature importance, como el uso de los valores SHAP, apenas incrementan la importancia final de las variables. Esto es bastante positivo, ya que indica que nuestro modelo no le esta asignando una mayor importancia a las variables interpretables, cuando no se puede interpretar una variable que apenas tiene importancia para la contribución final.

3.2.3 Múltiples variables no interpretables que tienen una alta contribución en el resultado final

Dataset de vinos

En el dataset de los vinos, vamos a tomar 4 variables que vamos a suponer que no son interpretables. Esto nos permitirá simular un caso de uso real donde existan multiples variables que no se puedan interpretar. De esta forma obtenemos 4 variables, las cuales son: density, fixed acidity, residual sugar, total sulfur dioxide.

Utilizando feature importance

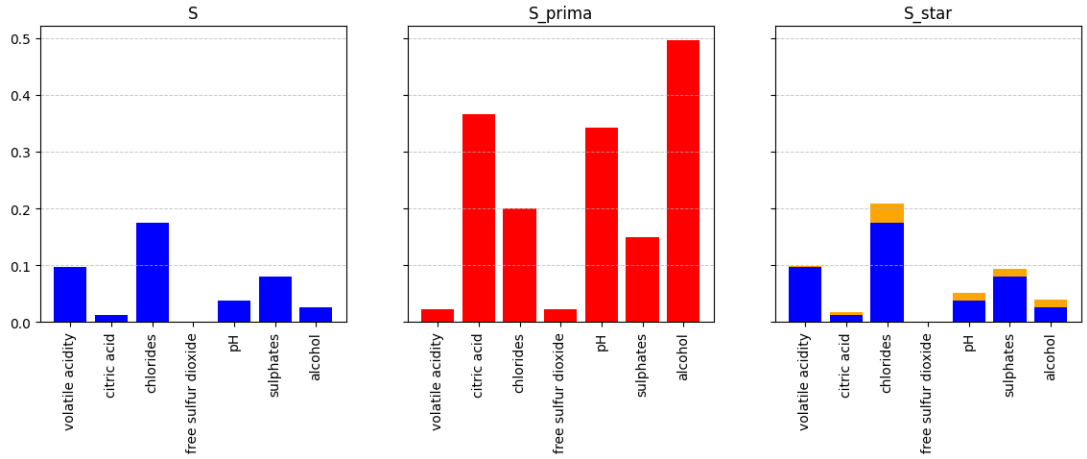

Figure 23: Predicción de la calidad del vino usando feature importance y tomando como variables no interpretables density, fixed acidity, residual sugar, total sulfur dioxide
Own elaboration

Utilizando valores SHAP

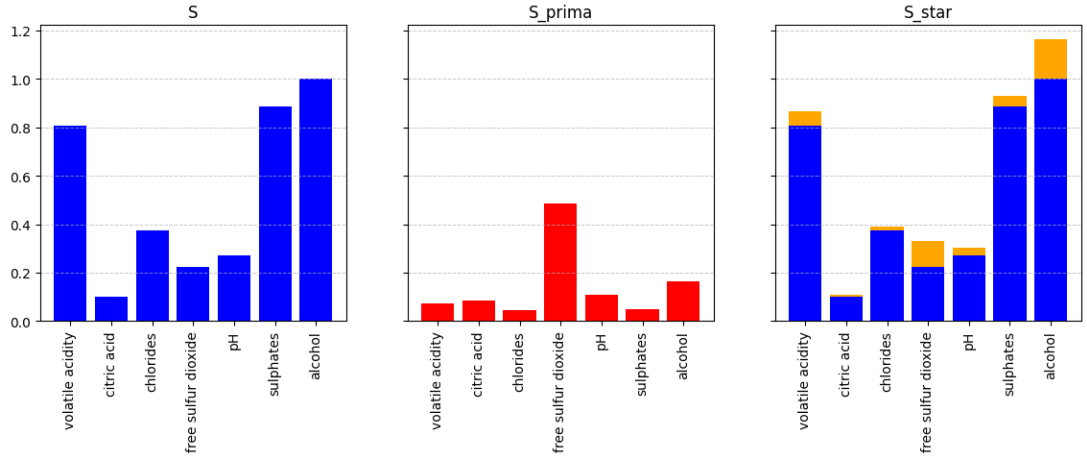

Figure 24: Predicción de la calidad del vino usando valores SHAP y tomando como variables no interpretables density, fixed acidity, residual sugar, total sulfur dioxide

Se puede observar como tanto utilizando el feature importace, como utilizando los valores SHAP, los valores de *S star* incrementan en base a la correlación que tienen con las variables no interpretables.

Este aumento es menor en el caso del feature importance, debido a que la diferencia entre la contribucion que aporta la variable density con respecto a las otras variables es muy grande 13. Esto no sucede en el caso de los valores SHAP, debido a que obtiene unos valores de contribución mucho mas homogeneos 14.

Dataset de diabetes

Asi como hemos hecho en el anterior dataset, en este vamos a aplicar nuestro algoritmo utilizando 3 variables que suponemos que no son interpretables. De esta forma usamos 3 variables, las cuales son: Glucose, Insulin, SkinThickness.

Utilizando feature importance

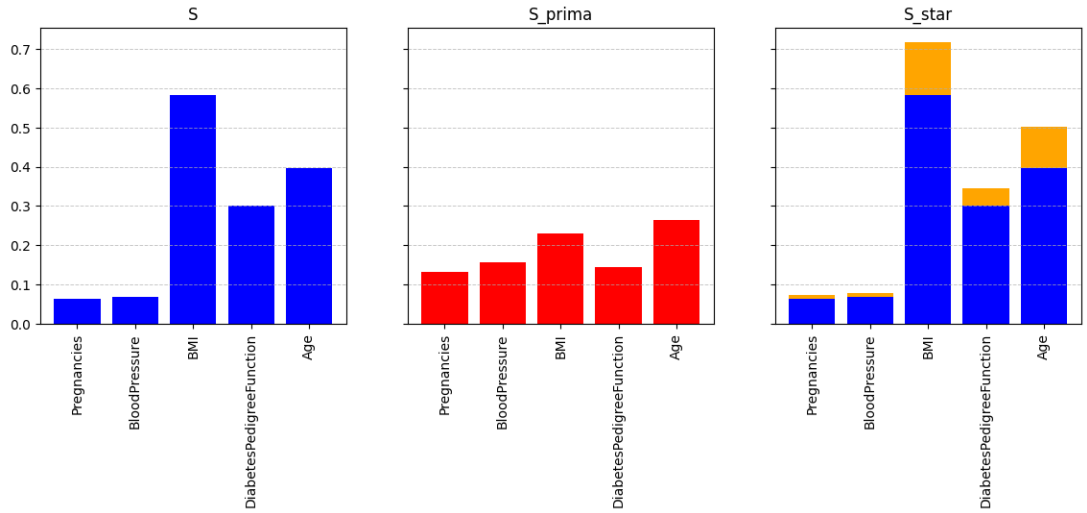

Figure 25: Resultados la predicción de diabetes usando feature importance y tomando como variables no interpretables Glucose, Insulin, SkinThickness

Own elaboration

Utilizando valores SHAP

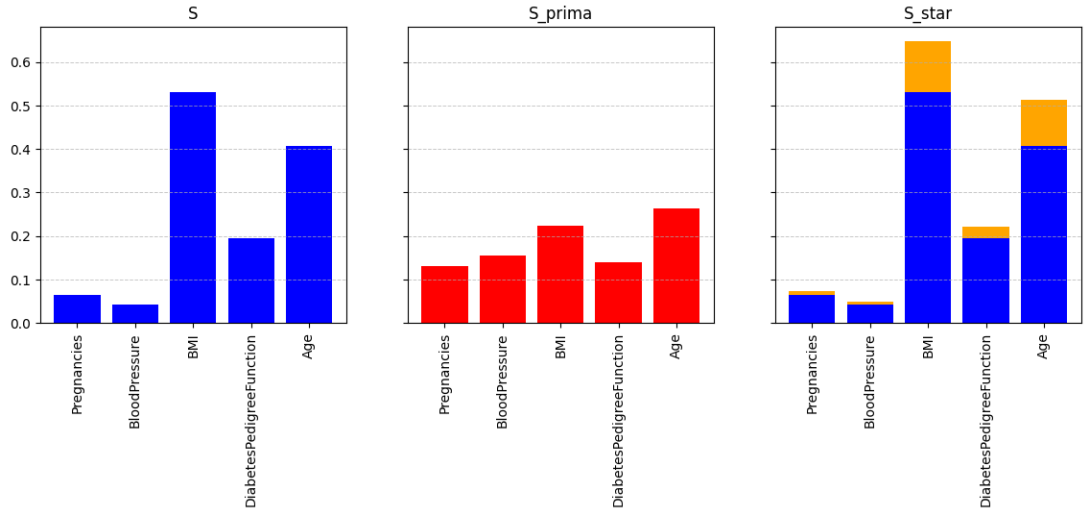

Figure 26: Resultados la predicción de diabetes usando valores SHAP y tomando como variables no interpretables Glucose, Insulin, SkinThickness
Own elaboration

En este dataset observamos como, al utilizar multiples variables, los valores de *S star* incrementan acorde a los valores de las variables no interpretables y su correlación con las variables interpretables.

Nuestra metodología funciona bien tanto usando feature importance, como utilizando SHAP values. Sin embargo su uso dependerá del modelo en concreto, pudiento utilizar los valores SHAP en cualquier modelo, mientras que unicamente se podrá utilizar feature importance en ciertos modelos.

4 Conclusiones

4.1 Conclusiones

El presente trabajo introduce una metodología innovadora para mejorar la interpretabilidad de modelos de aprendizaje automático que emplean variables no interpretables, permitiendo a los usuarios comprender mejor las decisiones y predicciones generadas por dichos modelos. Este objetivo se ha logrado mediante la vinculación de estas variables no interpretables con variables interpretables a través de algoritmos que consideran tanto la importancia de las características como sus correlaciones.

Los resultados obtenidos de la experimentación realizada han demostrado que es posible generar representaciones más interpretables al combinar información de variables no interpretables con datos interpretables. En particular, el enfoque de multiplicación aumentada escalada se destacó como la mejor estrategia para ponderar estas variables, logrando un equilibrio óptimo entre transparencia y precisión. Este método evitó problemas comunes observados en estrategias como la suma simple o multiplicación sin escalado, ofreciendo resultados consistentes y comprensibles.

La experimentación también nos ha dejado las conclusiones de que:

1. **Uso limitado de variables no interpretables:**
   Incluir un exceso de variables no interpretables puede complicar la comprensión del modelo sin aportar beneficios significativos en la interpretabilidad. Es preferible priorizar las variables cuya contribución pueda ser entendida y evaluada por los usuarios finales.

2. **La correlación guía la ponderación:**
   Cuando existe una correlación clara entre variables interpretables y no interpretables, la ponderación basada en dicha correlación resulta efectiva. Este enfoque permite que las variables interpretables actúen como un puente comprensible para explicar el impacto de las variables abstractas. No obstante, cabe destacar la importancia de que dichas variables contribuyan a la predicción final del modelo.

3. **Falta de importancia inicial en variables interpretables:**
   Se ha observado que si una variable interpretada inicialmente como no importante no presenta relevancia al comienzo del análisis, tampoco adquiere relevancia significativa tras el proceso de ponderación (S^*).

A pesar de los logros, es importante destacar algunas limitaciones que surgieron durante el desarrollo. La metodología se ha probado en pocos casos, y su aplicación a modelos más complejos como redes neuronales profundas, así como a otros dominios, sigue siendo una tarea pendiente. Además, aunque se usaron herramientas de explicabilidad como SHAP, su efectividad podría variar según el contexto del modelo y los datos utilizados.

En conclusión, esta investigación representa un avance significativo hacia la mejora de la interpretabilidad en modelos de IA, especialmente en aquellos que emplean variables complejas o abstractas. La metodología desarrollada no solo promueve un mayor entendimiento y confianza en los usuarios finales, sino que también abre nuevas líneas de investigación para adaptar y generalizar este enfoque a diferentes tipos de modelos y aplicaciones.

4.2 Trabajos Futuros

En este proyecto, hemos desarrollado y probado una metodología para mejorar la interpretabilidad de las variables no interpretables al vincularlas con variables interpretables. Sin embargo, existen varias áreas que podrían explorarse más a fondo:

1. **Inclusión de Técnicas Adicionales de Explicabilidad**
   Aunque utilizamos SHAP y feature importance para modelos basados en árboles, así como otros métodos correlacionales, no incorporamos herramientas como LIME o avances más recientes en métodos de explicabilidad que son menos famosos pero prometedores para ciertos modelos. Explorar como nuevas técnicas funcionan con nuestra metodologia.

2. **Expansión a Clases de Modelos Más Amplias**
   La metodología se probó con modelos que permiten extraer directamente la importancia de las características (por ejemplo, modelos basados en Machine Learning). Para modelos basados en Deep Learning, que son inherentemente más opacos, deberían evaluarse métodos avanzados de explicabilidad (por ejemplo, Integrated Gradients, Layer-wise Relevance Propagation) para determinar cuán efectivamente pueden mapear variables no interpretables a variables interpretables.

3. **Generalización a Diferentes Dominios**
   La metodología actual solo se evaluó en el contexto específico de dos datasets. Probar en diferentes campos, como finanzas, ciencias ambientales o procesamiento de lenguaje natural, podría validar su versatilidad y destacar áreas de mejora.

4. **Optimización del algoritmo**
   Seguir depurando el algoritmo y probar diferentes tecnicas y optimizaciones para evitar sesgos y conseguir que sea menos vunerables a diferentes conjuntos de datos.

Al abordar estas áreas, este trabajo puede evolucionar hacia una herramienta más integral y flexible, pero debido a la falta de tiempo y la gran cantidad de trabajo que queda por realizar en el area, se ha introducido en este apartado.

Bibliography

[1] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58 (June 2020), pp. 82–115. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103 (visited on 10/09/2024).

[2] Rudresh Dwivedi et al. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions". In: *ACM Comput. Surv.* 55.9 (Jan. 2023), 194:1–194:33. ISSN: 0360-0300. DOI: 10.1145/3561048. URL: https://dl.acm.org/doi/10.1145/3561048 (visited on 10/09/2024).

[3] Arun Rai. "Explainable AI: from black box to glass box". en. In: *Journal of the Academy of Marketing Science* 48.1 (Jan. 2020), pp. 137–141. ISSN: 1552-7824. DOI: 10.1007/s11747-019-00710-5. URL: https://doi.org/10.1007/s11747-019-00710-5 (visited on 10/09/2024).

[4] Dylan Slack et al. "Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods". In: *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. ACM, 2020, pp. 180–186. DOI: 10.1145/3375627.3375830. URL: https://dl.acm.org/doi/abs/10.1145/3375627.3375830.

[5] Scott M. Lundberg and Su-In Lee. "Feature Importance Analysis: From Trees to Additive Models". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Springer, 2019, pp. 303–330. DOI: 10.1007/978-3-030-10925-7_40. URL: https://link.springer.com/chapter/10.1007/978-3-030-10925-7_40.

[6] *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics - ScienceDirect*. URL: https://www.sciencedirect.com/science/article/pii/S0167947320301341?casa_token=U8p-B56SRaMAAAAA:9ruhy93bfQ3RuJlr_NblLuG6-da7By2dCvbbB2sj3ZLdpH9KOAQL4CEhe3aBExFkuN_8eH_BtXM (visited on 10/09/2024).

[7] A. Cerdeira Paulo Cortez. *Wine Quality*. 2009. DOI: 10.24432/C56S3T. URL: https://archive.ics.uci.edu/dataset/186 (visited on 12/14/2024).