UNIVERSIDAD
DE MURCIA

Informe de prácticas

Análisis y Aplicaciones de PDP e ICE para la Interpretabilidad de Modelos de Machine Learning

Autor: Juan Carlos Valera López
Profesor: José Manuel Juárez Herrero

Fecha: 25 de mayo 2025

ÍNDICE DE CONTENIDOS

TABLA DE ILUSTRACIONES

1. Introducción

En la actualidad, los modelos de aprendizaje automático (Machine Learning, ML) han demostrado una gran capacidad para realizar predicciones precisas en una amplia gama de dominios. Sin embargo, este avance predictivo tiene como efecto colateral un aumento en la opacidad del proceso que han seguido, especialmente en los modelos más potentes como las redes neuronales o los ensambles de árboles, actuando como "caja negra" y dificultando la comprensión humana. Esta opacidad inherente presenta desafíos significativos, especialmente para su implementación es aplicaciones críticas, donde la comprensión del proceso de toma de decisiones es crucial para la confianza, la seguridad y la depuración.

Es en consecuencia de estos desafíos de los modelos de caja negra que ha surgido el campo de la Inteligencia Artificial Explicable (XAI, por sus siglas en inglés), cuyo objetivo principal es desarrollar técnicas y métodos que nos permitan a los humanos entender y confiar en los sistemas de IA. La interpretabilidad hace referencia a la capacidad de mapear conceptos abstractos de los modelos a una forma comprensible para los humanos, mientras que la explicabilidad es un término más fuerte que requiere interpretabilidad junto con un contexto adicional.

Este estudio se enfoca en dos técnicas prominentes de XAI: los Gráficos de Dependencia Parcial (PDP) y los gráficos de Expectativa Condicional Individual (ICE). Los objetivos principales de este trabajo son:

1. Comprender en profundidad los fundamentos teóricos, el cálculo y ls interpretación de los g´raficos PDP e ICE, incluyendo su variante c-ICE.
2. Aplicar estas técnicas para analizar e interpretar el comportamiento de diferentes modelos de Machine Learning mediante su implementación en dos entornos diferentes: Orange Data Mining y Python.
3. Comparar y analizar los distintos gráficos obtenidos y las conclusiones que podemos extraer de estos.
4. Evaluar la utilidad y las limitaciones de estas técnicas en un contexto práctico.

2. Fundamentos teóricos: PDP e ICE

La interpretabilidad de los modelos de ML puede abordarse desde perspectivas globales, que buscan entender el comportamiento general del modelo, o locales, que se enfocan en explicar predicciones individuales. Aunque ambos tratan de representar las relaciones funcionales entre las variables de entrada y las predicciones de un modelo de ML, los PDP ofrecen una visión global, mientras que los ICE proporcionan una perspectiva más local y granular.

2.1. Gráficos de Dependencia Parcial (PDP)

2.1.1. Definición y propósito

Un Gráfico de Dependencia Parcial (PDP, Partial Dependence Plot) muestra el efecto marginal que una o dos características tienen sobre el resultado predicho por un modelo de ML. Su propósito es ayudar a visualizar si la relación entre la variable objetivo y una característica es lineal, monotónica o más compleja. Los PDP son considerados un método global porque tienen en cuenta todas las instancias del conjunto de datos para ofrecer una declaración sobre la relación global de una característica con el resultado predicho.

La intuición detrás de un PDP es tratar de aislar el efecto de la característica (o características) de interés. Para ello, se promedian los efectos de todas las demás características del modelo, conocidas como características complementarias. Imaginemos que queremos entender cómo la temperatura afecta a la predicción del alquiler de bicicletas. Un PDP para la temperatura nos mostraría cómo cambia la predicción promedio del número de bicicletas alquiladas a medida que varía la temperatura, manteniendo el efecto de todas las demás variables (como la humedad, el día de la semana, etc.) "promediado". Esto nos permite observar la tendencia principal de la relación entre la temperatura y el alquiler de bicicletas, según lo aprendido por el modelo.

Un aspecto interesante de los PDP es que su cálculo puede interpretarse, con ciertas precauciones, desde una perspectiva causal. Al variar sistemáticamente una característica y observar los cambios en las predicciones, se está simulando una intervención sobre esa característica para medir su impacto en el resultado. Sin embargo, esta interpretación causal es más directa y fiable cuando la característica de interés no está correlacionada con otras características del modelo. Si existe correlación, el proceso de variar una característica de forma aislada puede llevar a la creación de puntos de datos sintéticos que son poco probables o incluso imposibles en el mundo real. Por ejemplo, si "edad" y "años de experiencia" están altamente correlacionados, un PDP que varíe la "edad" manteniendo fijos los "años de experiencia" podría evaluar combinaciones poco realistas (p.ej., una persona muy joven con muchos años de experiencia). Por lo tanto, es crucial considerar las correlaciones entre características antes de extraer conclusiones causales fuertes basadas únicamente en PDPs.

2.1.2. Computación

Los Gráficos de Dependencia Parcial funcionan marginalizando la salida del modelo de ML sobre la distribución de las características que no son de interés. De esta manera, la función resultante muestra la relación entre las características de interés y el resultado predicho.

Formalmente, para un modelo f, la función de dependencia para un conjunto de características de interés, x_s, se define como la expectativa de la predicción del modelo sobre la distribución de las características complementarias, x_c:

$$pd_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} f(x_s, x_c^i)$$

Donde:

- n : número de instancias.
- x_s: Conjunto de características de interés (1 o 2 para poder visualizarlas).
- x_c^i : Valores de las características complementarias para una instancia i.

Menos formalmente, el proceso para calcular un PDP para una característica es el siguiente:

1. Seleccionar una característica de interés.
2. Definir un rango de valores mínimo y máximo equiespaciados.
3. Para cada instancia, calculas las predicciones que obtenemos en el modelo con cada valor del rango definido, manteniendo constantes el resto de características.
4. Promediar las predicciones para cada valor del rango definido.
5. Graficar los valores promedios de las predicciones.

Este proceso de generar "muestras falsas", nos permite estudiar la tendencia del modelo, pero también introduce una de las limitaciones más importantes del método, la cual trataremos más adelante.

2.1.3. Interpretación

La interpretación de los PDP varía ligeramente dependiendo de si la característica de interés es numérica o categórica, y del tipo de modelo de ML usado.

Características Numéricas

Para una característica numérica, el PDP se visualiza típicamente como un gráfico de líneas, donde en el eje x se representan los diferentes valores de la característica numérica que se está analizando, mientras que en el eje y se representa el promedio del resultado predicho por el modelo.

Una pendiente ascendiente sugiere que, en promedio, a medida que aumenta el valor de la característica, también lo hace la predicción, mientras que una pendiente descendente indica lo contrario. Cuanto más pronunciada sea la pendiente, más fuerte será la influencia de la característica en la predicción.

Si la línea no es recta, sino que presenta curvas, valles o picos, indica que el efecto de la característica sobre la predicción no es constante, lo que puede ser un factor que revele correlación/interacción con otras características complementarias.

Podemos observar un ejemplo en la siguiente imagen. Se trata de un gráfico PDP que estudia la relación que establece un modelo de Random Forest entre la cantidad de alcohol que tiene un vino blanco y la calidad de este.

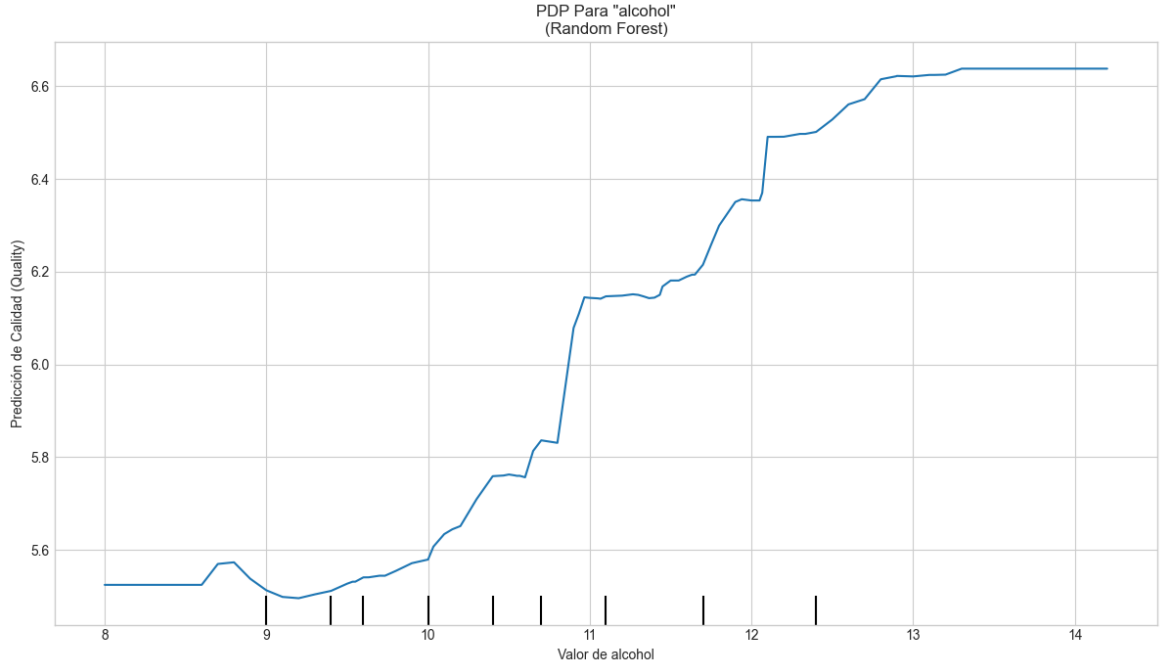

Ilustración 1:PDP-RF-quality of white wine/alcohol

Como ya hemos comentado, el modelo usado en la predicción también afecta a su visualización, pudiendo observar líneas "suaves" en la variación del promedio de la predicción por ser el anterior un modelo de Random Forest. Comparémoslo con el producido por un árbol de decisiones para la misma característica y mismo DataSet:

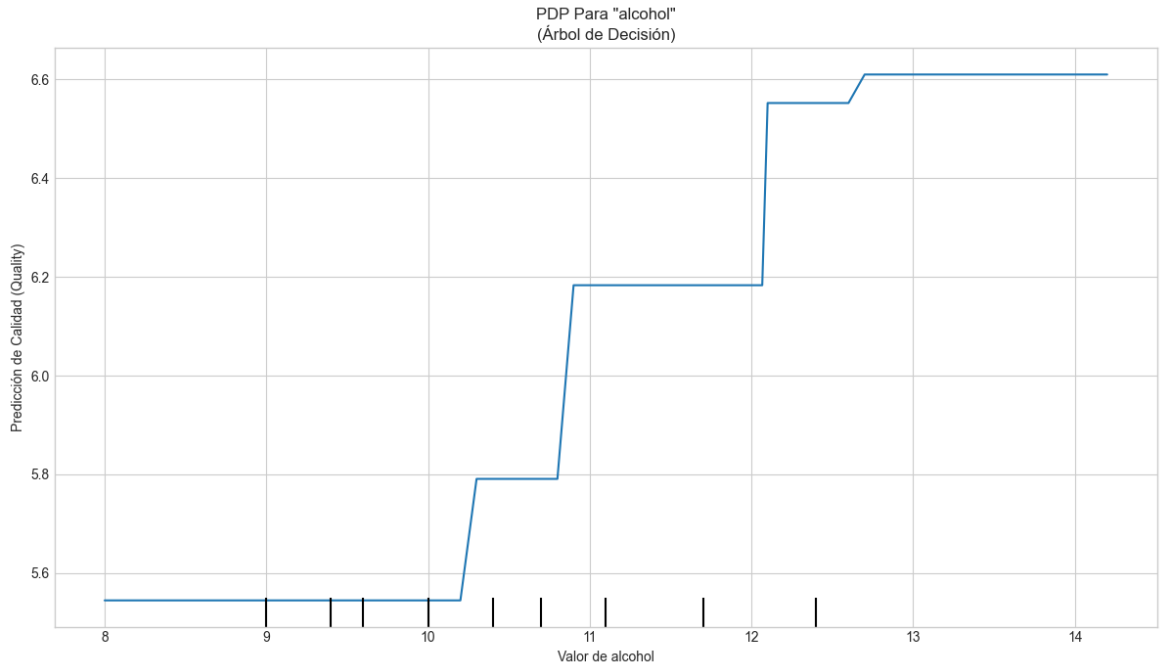

Ilustración 2:PDP-DT-quality of white wine/alcohol

Es fácil comprobar que, aunque ambos han establecido la relación "a mayor nivel de alcohol, mayor calidad", el PDP del árbol de decisión es mucho más escalonado.

Características Categóricas

El PDP muestra la predicción promedio para cada categoría, lo que usualmente se suele visualizar como un gráfico de barras, donde cada barra representa una categoría de las características y la altura de la barra indica la predicción promedio para esa categoría.

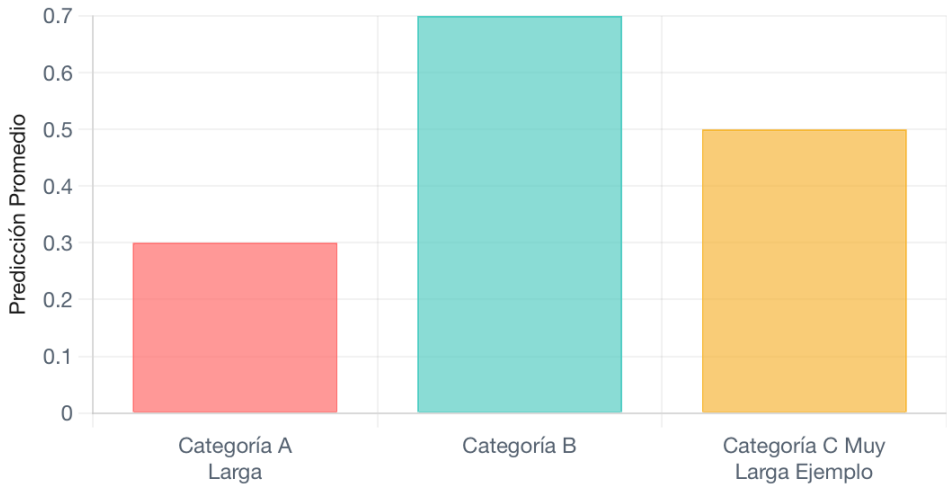

Ilustración 3: Ejemplo PDP variable categórica

La *Ilustración 3* muestra un ejemplo de cómo podría ser un PDP para una característica categórica.

Dimensionalidad

Los gráficos mostrados hasta este momento relacionan la predicción del modelo con una única característica de interés, pero esto puede ampliarse, como máximo, hasta dos, debido la conocida maldición de la dimensionalidad, generando en este caso un mapa de contorno/calor. En este tipo de gráficos relacionamos las variables de interés en los ejes x, e y, mientras que el color refleja la predicción. En este caso, cuanto más oscuro sea el color, mayor es el valor de la predicción.

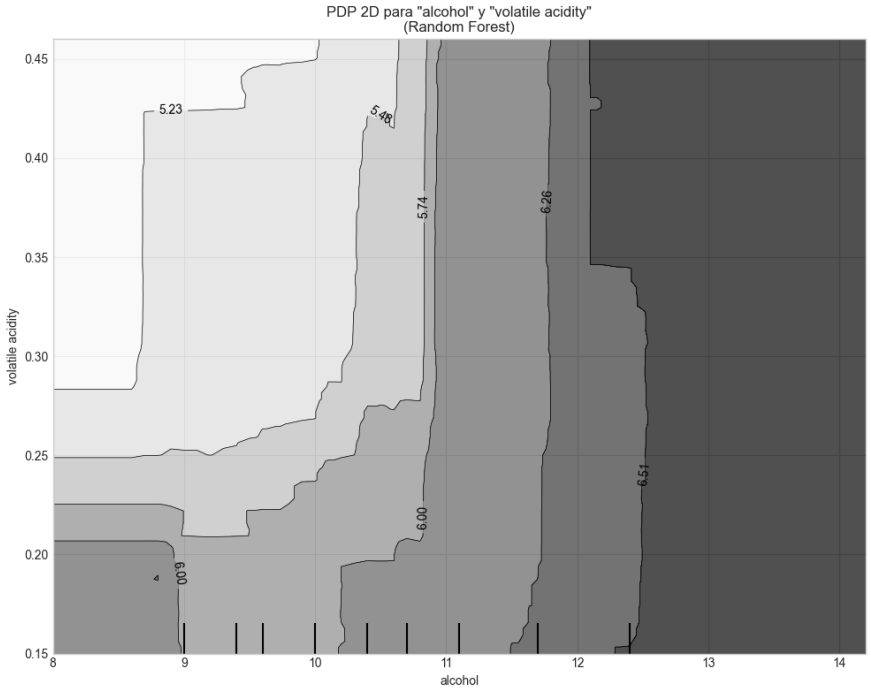

Ilustración 4:Mapa de calor PDP

2.1.4. Ventajas y Limitaciones

Como hemos comentado, los Gráficos de Dependencia Parcial son herramientas valiosas, pero es fundamental conocer tanto sus fortalezas como sus debilidades para realizar una interpretación adecuada.

Ventajas:

- **Interpretación Intuitiva:** Los PDP son generalmente fáciles de entender. Muestran de manera clara cómo cambia la predicción promedio del modelo a medida que varía el valor de una característica, lo cual es accesible incluso para audiencias no técnicas.

- **Visión Global:** Proporcionan una comprensión global del comportamiento del modelo con respecto a una o dos características, resumiendo la tendencia general a través de todo el conjunto de datos.

- **Agnósticos al Modelo:** Una gran ventaja es que los PDP pueden utilizarse para interpretar cualquier modelo de machine learning, desde los más simples hasta las "cajas negras" más complejas como redes neuronales o ensambles.

- **Interpretación Causal (Condicional):** Bajo la estricta condición de que la característica de interés no esté correlacionada con otras características, los PDP pueden ofrecer indicios sobre una relación causal entre la característica y la predicción.

Limitaciones:

- **Suposición de Independencia de Características:** Esta es la limitación más crítica. Los PDP asumen que la(s) característica(s) de interés no están correlacionadas con las otras características del modelo. Si existe correlación, el proceso de marginalización promedia las predicciones sobre instancias sintéticas que pueden ser poco realistas o incluso imposibles en la distribución de datos real, llevando a interpretaciones potencialmente engañosas.

- **Oculta Efectos Heterogéneos:** Al mostrar solo el efecto promedio, los PDP pueden ocultar relaciones heterogéneas, es decir, situaciones en las que una característica afecta a diferentes instancias o subgrupos de datos de manera distinta. Por ejemplo, una característica podría aumentar la predicción para un grupo y disminuirla para otro, resultando en un PDP plano que sugiere, erróneamente, que no hay efecto. Es aquí donde los gráficos ICE se vuelven indispensables.

- **Máximo de Dos Características:** Debido a las limitaciones de la percepción humana y la representación bidimensional, los PDP solo pueden visualizar de manera significativa el efecto de una o, como máximo, dos características a la vez, generando gráficos 2D. Esto se relaciona con la "maldición de la dimensionalidad" en la interpretación: aunque los modelos pueden manejar datos de alta dimensión, nuestras herramientas de interpretación visual a menudo se limitan a proyecciones de baja dimensión. Esto significa que los PDP solo ofrecen una visión parcial de la función aprendida por el modelo, y las interacciones complejas que involucran tres o más características no serán capturadas.

2.2. Gráficos de Expectativa Condicional Individual (ICE)

2.2.1. Definición y propósito

Los gráficos de Expectativa Condicional Individual (ICE, Individual Conditional Expectation) son una técnica de visualización que desagrega los promedios globales de los PDPs para mostrar cómo la predicción de un modelo cambia para cada instancia individual a medida que varía el valor de una característica específica, mientras todas las demás características de esa instancia se mantienen constantes. En esencia, un gráfico ICE traza una línea por cada observación del conjunto de datos, representando su propia curva de predicción condicionada.

El propósito principal de los gráficos ICE es revelar la heterogeneidad en las predicciones del modelo que podría quedar oculta por el efecto promedio que muestran los PDPs. Si el efecto de una característica sobre la predicción es diferente para distintas instancias (debido a interacciones con otras características o subgrupos), los gráficos ICE lo harán visible al mostrar líneas con diferentes formas, pendientes o niveles.

Son útiles para la depuración de modelos, conllevando una comprensión más granular, ya que cada línea está condicionada a la combinación particular de los otros valores de características de esa instancia.

2.2.2. Computación

La computación de los gráficos ICE es similar en forma al de los PDPs, omitiendo el proceso del cálculo del promedio. En cuanto a carga de cálculo, el proceso para gráficos ICE es similar a PDP, sin embargo, a la hora de graficar, el coste computacional es considerablemente mayor en el caso de gráficos ICE, al tener que representar cada una de las instancias, en lugar de una única representación para el promedio como ocurre con PDPs.

$$ice_s(x_s) = f(x_s, x_c^i)$$

Donde:

- x_s: Conjunto de valores de la característica de interés.
- x_c^i : Valores de las características complementarias para una instancia i.

2.2.3. Interpretación

Como ocurre con los PDPs, la visualización de los gráficos ICE puede variar en función del modelo; sin embargo, estos no están pensados para variables categóricas, sino que únicamente para características numéricas.

La característica más distintiva de un gráfico ICE es que cada línea trazada representa una única instancia del conjunto de datos. Los gráficos ICE suelen incluir una representación del PDP, que permiten compararlos directamente con estos últimos. Continuando los ejemplos del DataSet de calidad de vino blanco de la sección de PDP:

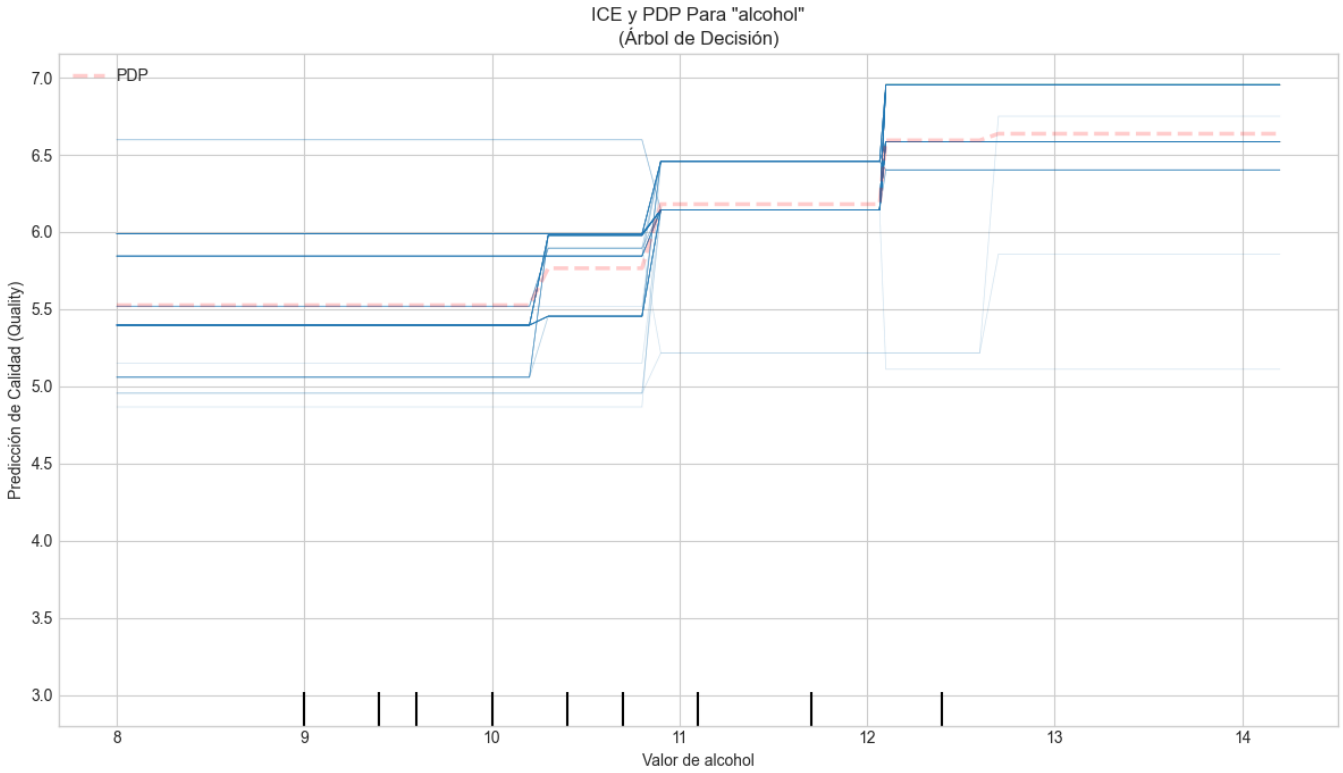

Ilustración 5: ICE&PDP-DT-WineQuality-Alcohol

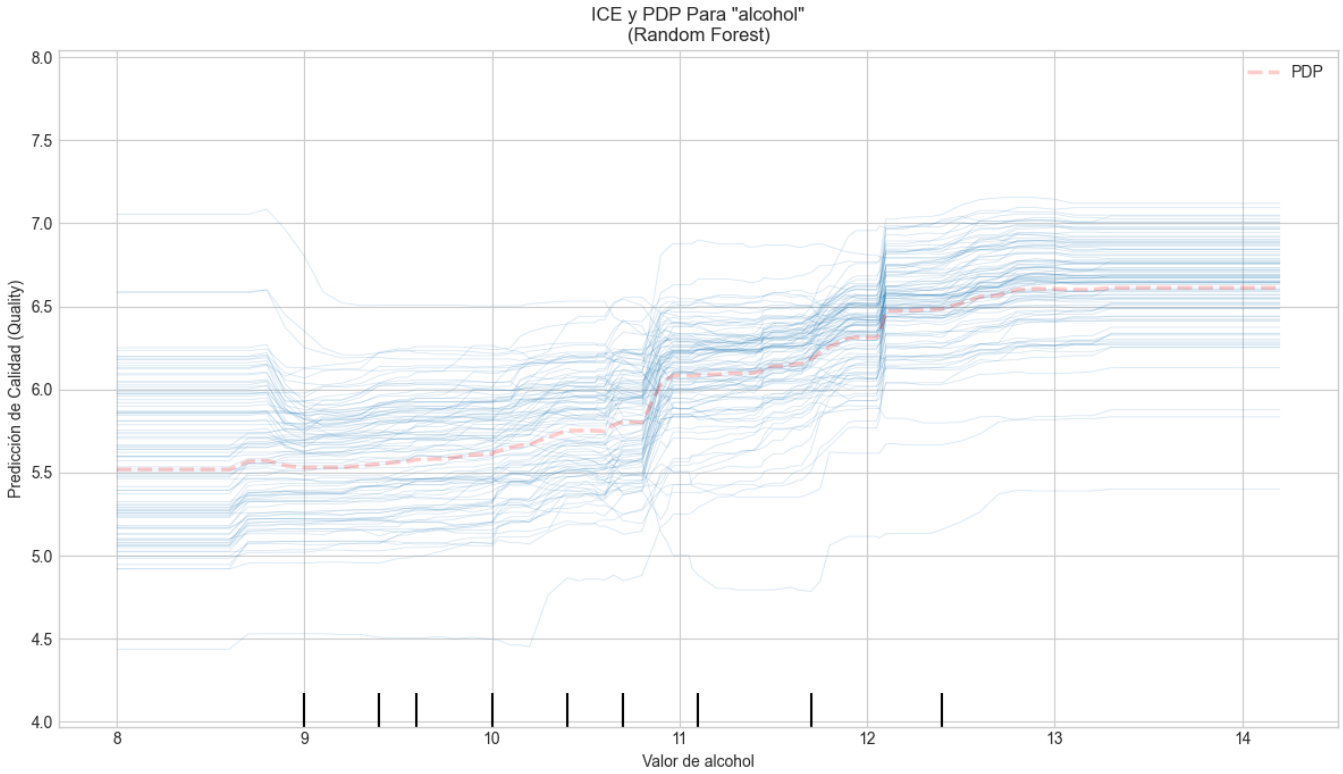

Ilustración 6:ICE&PDP-RF-WineQuality-Alcohol

En las ilustraciones 5 y 6, podemos observar la representación conjunta de PDP y gráficos ICE, siendo el PDP la línea discontinua roja. Con esta representación, es más sencillo encontrar ciertos patrones que han podido ser ocultados/aplanados por el PDP, por ejemplo, podemos encontrar en este caso concreto que hay ciertas instancias en las que, a pesar de aumentar la cantidad de alcohol, la calidad de este no aumenta, sino que disminuye.

De forma similar a como ocurría en el caso de PDP, la forma de las líneas está condicionada por el modelo sobre el que se apliquen, encontrando una representación más escalonada en el árbol de decisiones, mientas que unas líneas más "suaves" en el caso del Random Forest.

2.2.4. Ventajas y limitaciones

Los gráficos ICE ofrecen una visión más detallada que los PDP, pero también conllevan sus propias ventajas y limitaciones.

Ventajas:

- **Intuitivos de Entender (a nivel individual):** El concepto de una línea que representa las predicciones para una única instancia mientras se varía una característica es bastante directo y fácil de comprender.

- **Descubren Relaciones Heterogéneas e Interacciones:** Esta es la principal ventaja sobre los PDP. Los ICE pueden revelar si el efecto de una característica varía entre diferentes instancias o subgrupos, lo cual es crucial para entender las interacciones del modelo.

- **Visión Más Granular:** Permiten una exploración más profunda del comportamiento del modelo a nivel de instancia individual, lo que puede ser útil para la depuración del modelo o para entender predicciones específicas.

Limitaciones:

- **Densidad Visual (Overcrowding):** Cuando se trazan muchas curvas ICE (una para cada instancia en un conjunto de datos grande), el gráfico puede volverse extremadamente denso y difícil de interpretar, haciendo imposible discernir patrones individuales o la tendencia general. Para mitigar este problema, se suele seleccionar una submuestra y/o aplicar transparencia en las líneas, lo que proporciona un poco de claridad.

- **Visualización Significativa de una Sola Característica:** Los gráficos ICE solo pueden mostrar de manera efectiva el efecto de una característica a la vez, y no de dos, como podía mostrar un PDP. Intentar visualizar el efecto de dos características con ICE requeriría graficar múltiples superficies superpuestas, lo que resultaría ininteligible.

- **Problema de Correlación:** Al igual que los PDP, los gráficos ICE pueden sufrir si la característica de interés está altamente correlacionada con otras características. El proceso de variar una característica mientras se mantienen las otras fijas puede generar puntos a lo largo de las líneas ICE que representan combinaciones de características poco realistas o inválidas según la distribución conjunta real de los datos.

- **Dificultad para Ver el Efecto Promedio:** Con muchas líneas individuales, puede ser difícil discernir la tendencia promedio (es decir, el PDP). Aunque una solución simple, que ya hemos aplicado en los ejemplos anteriores, es superponer la línea del PDP en el gráfico ICE.

- **Dificultad para distinguir los efectos individuales:** Cuando los puntos iniciales de las predicciones son muy variados, cuesta ver de forma unificada las tendencias. Una solución para esto es utilizar gráficos ICE centrados.

2.2.5. Gráficos ICE Centrados (c-ICE)

Los gráficos ICE Centrados (c-ICE) son una variante de los gráficos ICE diseñada para abordar el problema de que las curvas individuales pueden comenzar en diferentes niveles de predicción, lo que dificulta la comparación de sus formas y la detección de efectos heterogéneos.

Un gráfico c-ICE se obtiene al anclar todas las curvas ICE a un punto de referencia común en el eje y, generalmente en el valor cero. Esto se logra restando de cada punto de una curva ICE la predicción del modelo para esa misma instancia en un valor fijo de la característica de interés. Un punto de anclaje común es el valor mínimo observado de la característica que se está analizando.

La fórmula para una curva c-ICE para la instancia i es:

$$cice_s(x_s) = f\left(x_s, x_c^i\right) - f(x_s^*, x_c^i)$$

Donde:

- x_s: Conjunto de valores de la característica de interés.
- x_c^i : Valores de las características complementarias para una instancia i.
- x_s^*: Valor de anclaje de la característica de interés x_s.

El propósito de centrar las curvas ICE es facilitar la comparación directa de cómo la predicción cambia para diferentes instancias a medida que varía la característica de interés, eliminando las diferencias en los niveles de predicción base. Esto ayuda a observar con mayor claridad si algunas observaciones tienen un patrón de respuesta divergente, incluso cuando las curvas ICE originales están muy superpuestas.

Continuando con los ejemplos anteriores para apreciar las diferencias, las ilustraciones 7 y 8 mostrarán un ejemplo de gráfico de expectativa condicional individual centrado para los modelos ya comentados: un árbol de decisiones y un Random Forest.

En estas ilustraciones es fácil comprobar cómo, al normalizar el punto de partida, cualquier heterogeneidad en las trayectorias de las curvas es mucho más apreciable. Esto proporciona una señal visual más directa de la presencia de interacciones que los gráficos ICE estándar, especialmente cuando se visualizan muchas líneas.

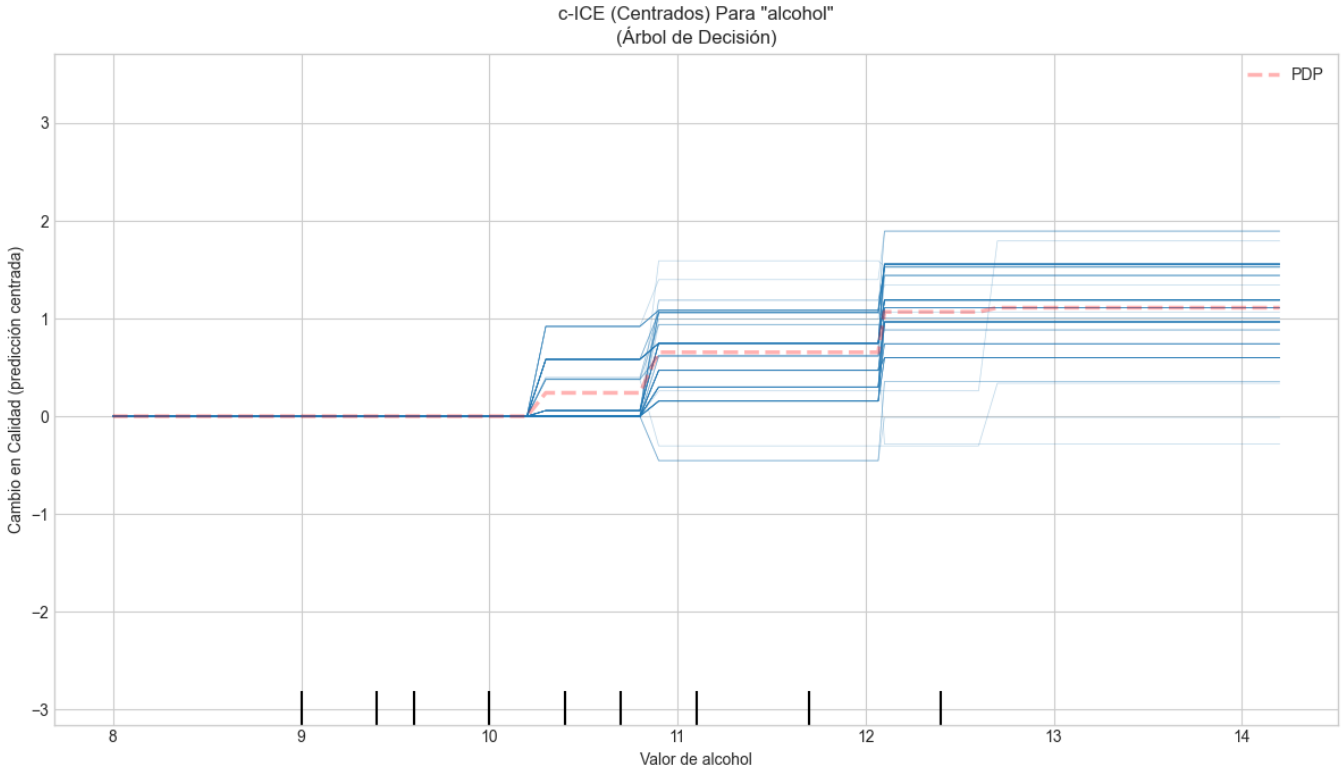

Ilustración 7:c-ICE&PDP-DT-WineQuality-Alcohol

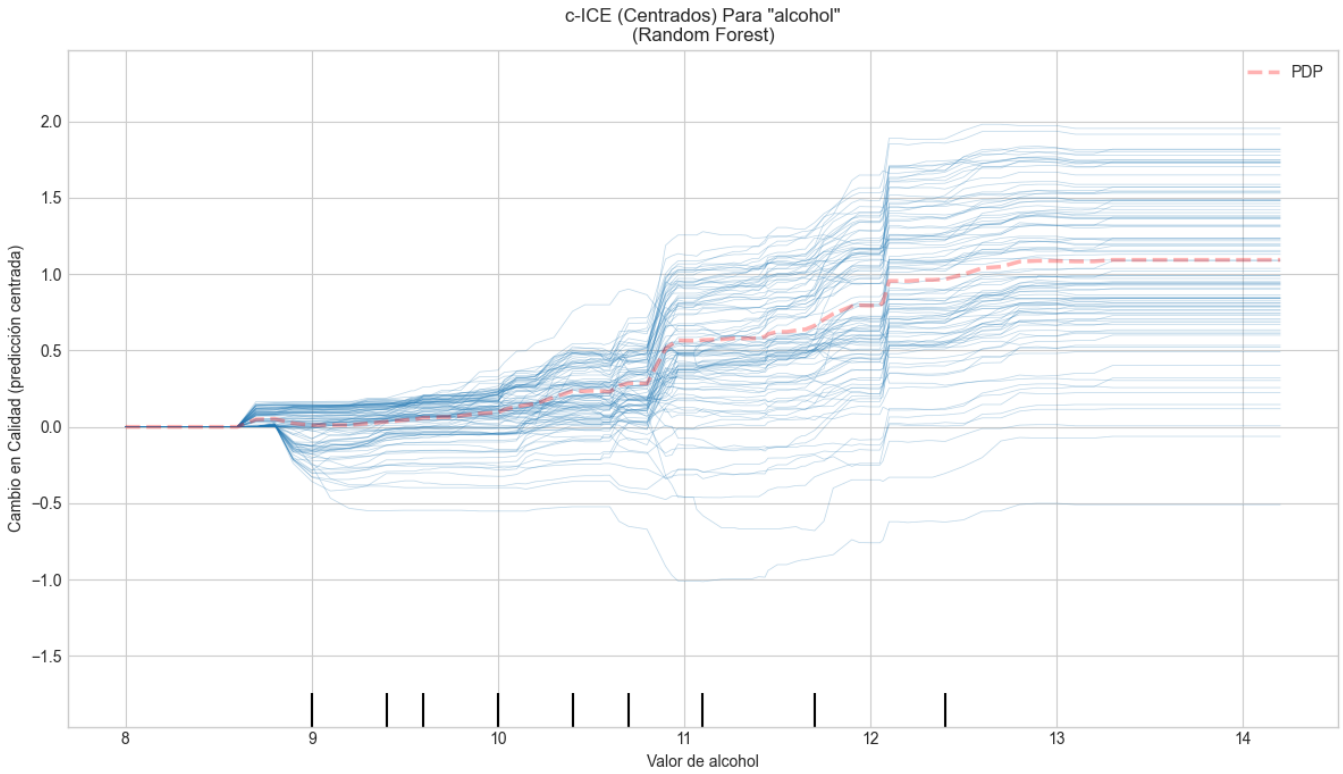

Ilustración 8:c-ICE&PDP-RF-WineQuality-Alcohol

2.2.6. Gráficos ICE Derivados (d-ICE)

Los gráficos ICE Derivados (d-ICE) son otra variante de los ICE que buscan resaltar la heterogeneidad en las respuestas del modelo, esta vez enfocándose en la tasa de cambio de la predicción.

Los gráficos d-ICE visualizan las derivadas parciales individuales de la función de predicción del modelo con respecto a la característica de interés, para cada instancia. La derivada indica si la predicción está cambiando, y en qué dirección (positiva o negativa) y magnitud lo hace. Si no existen interacciones significativas, las derivadas individuales deberían ser similares para todas las instancias. Las diferencias en estas derivadas entre instancias son, por lo tanto, indicativas de interacciones.

El propósito es facilitar la identificación de rangos de valores de la característica donde las predicciones del modelo cambian de manera diferente para distintas instancias, señalando así la presencia de efectos heterogéneos.

A pesar de su potencial para revelar interacciones, los gráficos d-ICE pueden ser computacionalmente muy costosos y, a menudo, se consideran poco prácticos para el uso rutinario, especialmente en comparación con los ICE y c-ICE. Debido a su complejidad computacional y menor uso práctico en comparación con ICE y c-ICE, los d-ICE se mencionan aquí por completitud, pero no serán el foco principal de las implementaciones prácticas en este trabajo.

3. Metodología y Herramientas

La metodología empleada en este estudio se centra en la aplicación práctica de las técnicas de interpretabilidad PDP e ICE sobre modelos de regresión entrenados para predecir la calidad del vino blanco. Se sigue un enfoque comparativo, analizando los resultados obtenidos tanto en el software Orange Data Mining como mediante programación en Python.

3.1. Conjunto de Datos Principal Seleccionado: "White Wine Quality"

Para la demostración práctica de PDP e ICE, se ha seleccionado el dataset "*White Wine Quality*", con información referente a la calidad del vino blanco, obtenido del repositorio UCI Machine Learning (https://archive.ics.uci.edu/dataset/186/wine+quality). Nos hemos decantado por este dataset por la naturaleza del problema, el hecho de que todas las variables predictoras son numéricas, lo que simplifica el preprocesamiento y permite una aplicación directa de los gráficos PDP e ICE, y por el potencial de interacciones que podríamos encontrar en las cualidades físico-químicas de los componentes, que es ideal para explorar las diferencias entre PDP e ICE.

3.1.1. Descripción del Dataset

El dataset "White Wine Quality" contiene 4898 instancias, cada una representando una muestra de vino blanco portugués "Vinho Verde". Dispone de 12 atributos en total: 11 características físico-químicas y una variable objetivo que indica la calidad.

3.1.2. Variable Objetivo y Características Predictoras

- **Variable Objetivo (Y)**: *quality*. Es una puntuación sensorial asignada por expertos, que varía típicamente entre 3 (muy malo) y 9 (excelente). En este estudio, se trata como una variable continua para los modelos de regresión.

- **Características Predictoras (X)**: Las 11 variables predictoras son:

    - fixed acidity (g/dm³ de ácido tartárico)
    - volatile acidity (g/dm³ de ácido volátil)
    - citric acid (g/dm³)
    - residual sugar (g/dm³)
    - chlorides (g/dm³ de cloruro de sodio)
    - free sulfur dioxide (mg/dm³)
    - total sulfur dioxide (mg/dm³)
    - density (g/cm³)
    - pH
    - sulphates (g/dm³ de sulfato de potasio)
    - alcohol (% vol.)

3.1.3. Preprocesamiento Básico Realizado

Dado que todas las características predictoras son numéricas, el preprocesamiento requerido es mínimo (en python, ya que en Orange no es necesario realizar ningún procesamiento manual):

1. **Carga de Datos**: El archivo *White-wine-quality.csv*, disponible en el enlace anterior, o directamente en el repositorio preparado, es cargado gestionando las cabeceras y metadatos iniciales para obtener un DataFrame limpio.

2. **Separación de Variables**: Se distingue entre el conjunto de características predictoras (X) y la variable objetivo (y).

3. **División del Conjunto de Datos**: El dataset se divide en un conjunto de entrenamiento (utilizado para ajustar los modelos) y un conjunto de prueba (utilizado para una evaluación preliminar del rendimiento de los modelos). Se emplea una proporción de 75% para entrenamiento y 25% para prueba, utilizando un random_state fijo para garantizar la reproducibilidad de los experimentos.

3.2. Modelos de Machine Learning Utilizados

Para ilustrar la aplicación de PDP e ICE, seleccionamos los siguientes modelos de regresión:

3.2.1. Árbol de Decisión (Decision Tree Regressor)

Como ejemplo de modelo interpretable de forma inherente, seleccionamos un Árbol de Decisión, el cual se configuró con una profundidad máxima limitada para facilitar la visualización de su estructura y la naturaleza escalonada de sus gráficos PDP e ICE, tal como se describe en la sección 2.1.3.

3.2.2. Random Forest (Random Forest Regressor)

Para tener un ejemplo en el que se evidenciara de mayor manera la utilidad de las herramientas de estudio, seleccionamos un Random Forest como ejemplo de un modelo de ensamble más complejo y, a menudo, más preciso. Este tipo de modelo, al promediar las predicciones de múltiples árboles, tiende a producir PDP e ICE más suaves y es capaz de capturar relaciones no lineales e interacciones de forma más robusta. El Random Forest también se utilizó para derivar una medida de la importancia de las características, guiando la selección de variables para un análisis más profundo con PDP/ICE.

3.3. Herramientas de Software

3.3.1. Orange Data Mining

Se utilizó **Orange Data Mining** (versión 3.38.1) para una fase inicial de exploración visual e interactiva. La interfaz gráfica de Orange, basada en la conexión de widgets, permitió construir rápidamente flujos de trabajo para el entrenamiento de modelos y la generación de gráficos PDP e ICE. Esta herramienta fue particularmente útil para desarrollar una intuición inicial sobre los datos y el comportamiento de los modelos, así como para explorar visualmente las interacciones mediante las funcionalidades incorporadas en el widget ICE.

3.3.2. Python y Librerías Clave

El análisis principal, la implementación detallada de los modelos y la generación de los gráficos para este informe se realizaron utilizando **Python** (versión 3.11.5), principalmente dentro de un entorno de **Google Colab**. Las librerías fundamentales empleadas son:

- **Pandas**: Para la carga, manipulación y gestión eficiente de los datos tabulares.
- **NumPy**: Para operaciones numéricas y el manejo de arrays.
- **Scikit-learn**: Para la implementación de los algoritmos de Machine Learning (*DecisionTreeRegressor*, *RandomForestRegressor*), la división de datos (*train_test_split*), el cálculo de métricas de error (*mean_squared_error*), y, crucialmente, para la generación de los Gráficos de Dependencia Parcial y de Expectativa Condicional Individual a través del módulo *sklearn.inspection.PartialDependenceDisplay*.
- **Matplotlib** y **Seaborn**: Para la creación y personalización de todas las visualizaciones estáticas, incluyendo los gráficos de importancia de características y los propios gráficos PDP e ICE.

4. Experimentación en Orange

La exploración con Orange Data Mining sirvió como una etapa inicial para visualizar de forma interactiva los conceptos de PDP e ICE aplicados al dataset "*White Wine Quality*". Esta herramienta facilitó una comprensión preliminar de los datos y el comportamiento de los modelos gracias a su interfaz visual.

4.1. Configuración del Flujo de Trabajo

Se diseñó un flujo de trabajo en Orange comenzando con un widget File para cargar el dataset "White Wine Quality". Los datos se conectaron a widgets de modelos como Tree (Árbol de Decisión, configurado con max_depth=5), Random Forest y SVM. Las salidas de datos y de los modelos entrenados se enlazaron a widgets ICE dedicados para cada modelo, permitiendo una comparación directa. Se utilizaron widgets adicionales como Tree Viewer para inspeccionar la estructura del árbol de decisión y Feature Importance (conectado al modelo Random Forest) para obtener una clasificación de la relevancia de las variables predictoras. La capacidad de cambiar entre diferentes datasets de prueba (como "Nota-final-estudiantes" o "Food-Info") se gestionó mediante la conexión a un widget Data Table intermedio, permitiendo una exploración comparativa ágil.

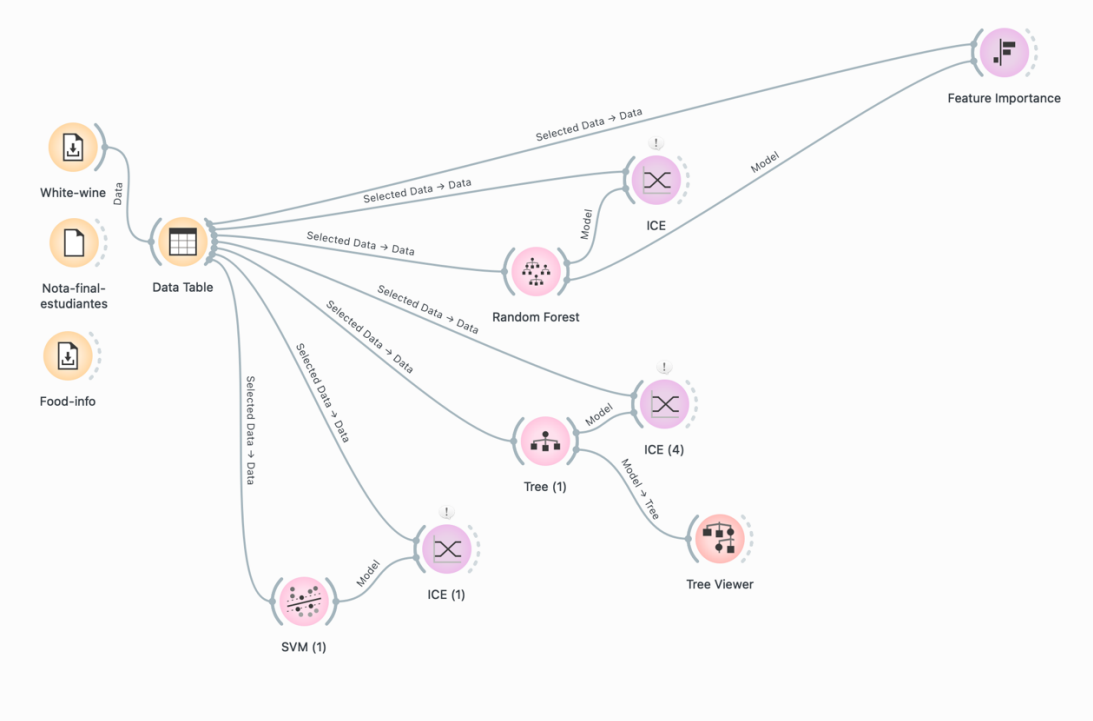

Ilustración 9: Configuración del Entorno en Orange

Es importante comentar que, en Orange, no tenemos widgets diferenciados para realizar PDPs y gráficos ICE, sino que todo se realiza a través del mismo widget “ICE”, y es mediante configuraciones en este que podemos obtener los diferentes tipos de gráficos, además de que realiza un submuestreo automático cuando se le intentan introducir más de 300 instancias.

4.2. Análisis con Árbol de Decisión

Los Gráficos de Dependencia Parcial (PDP) para alcohol mostraron la forma escalonada esperada (aunque limitada a 300 instancias), donde la predicción promedio de quality aumentaba en umbrales discretos. Las líneas ICE individuales, aunque también escalonadas, muestran una importante variedad en la predicción que el PDP no es capaz de reflejar.

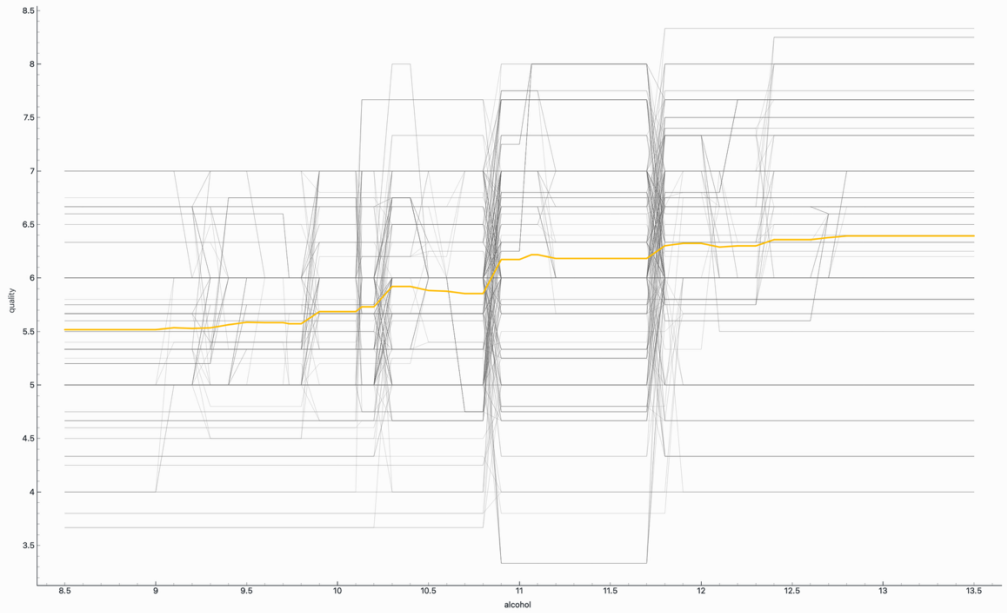

Ilustración 10: PDP e ICE en Orange DT

Generar un c-ICE en Orange es tan sencillo como seleccionar la opción “Centered” dentro del widget ICE:

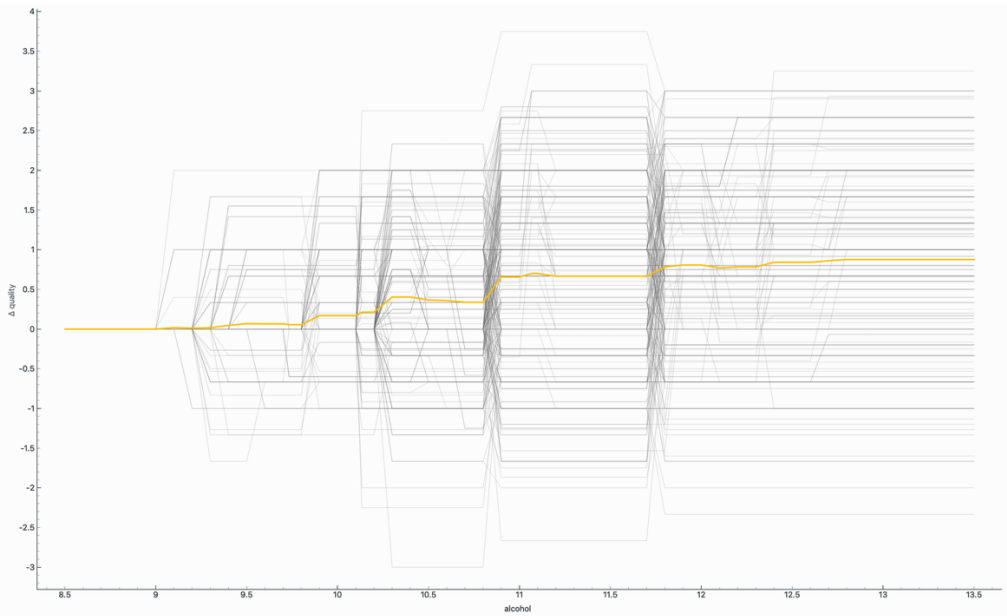

Ilustración 11: PDP y c-ICE en Orange DT

4.3. Análisis con Random Forest

Siguiendo el mismo procedimiento que para el árbol de decisiones, clickando sobre el widget ICE (esta vez conectado al Random Forest) podemos observar los diagramas generados por Orange:

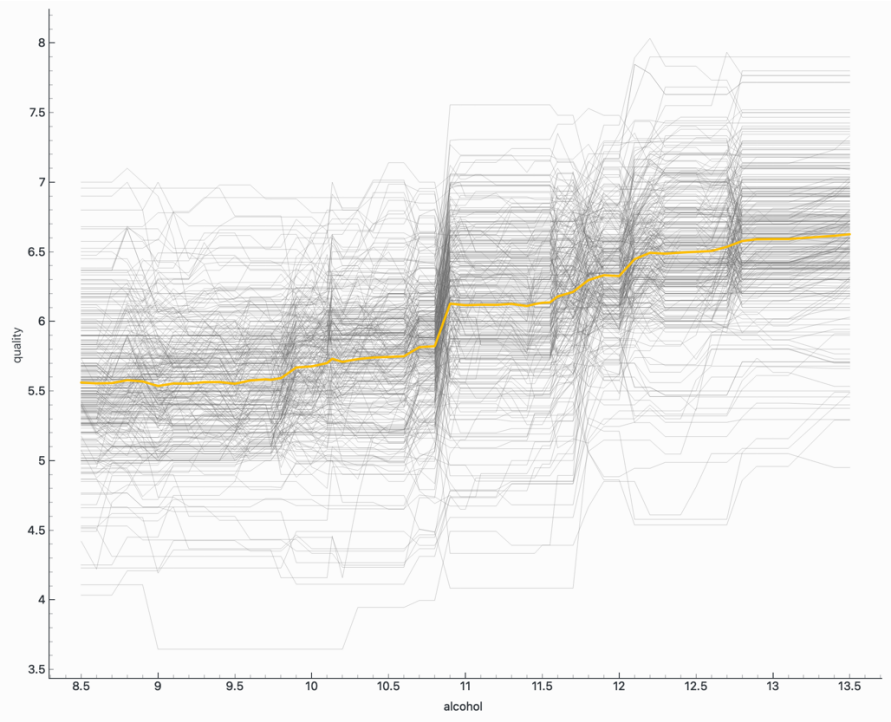

Ilustración 12: PDP e ICE en Orange RF

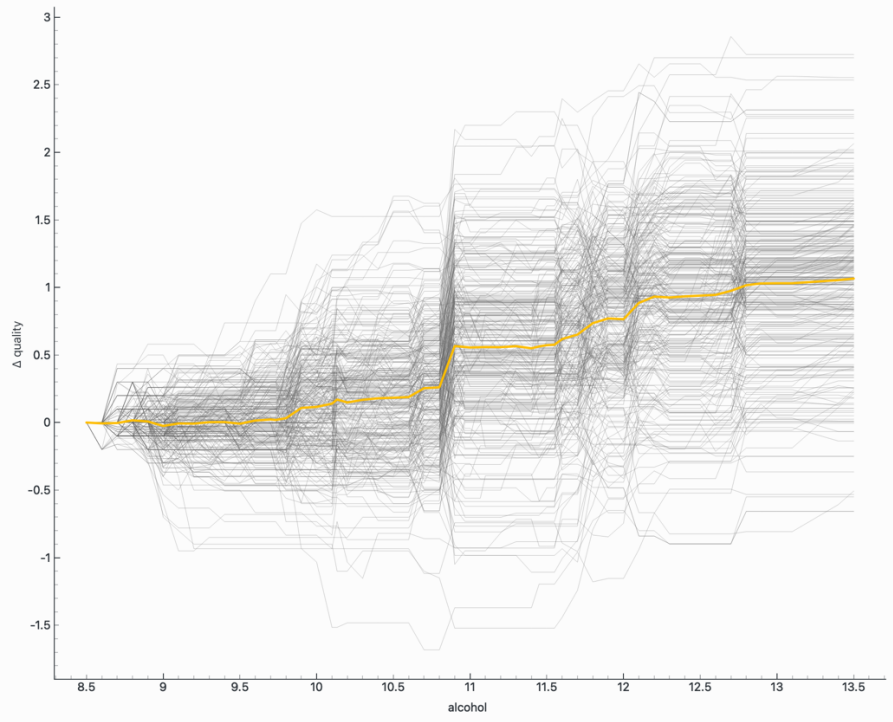

Ilustración 13: PDP y c-ICE en Orange RF

Los gráficos c-ICE para alcohol con el Random Forest son particularmente informativos, mostrando un claro "abanico" de líneas. Esto realta que la sensibilidad de la quality predicha ante cambios en el alcohol es heterogénea entre las instancias, lo que se puede traducir en un fuerte indicio de interacciones con otras características que el modelo estaba capturando.

4.4. Análisis con SVM

Por completitud, se exploró brevemente un modelo SVM. En este caso, el análisis de importancia de características sugirió que la variable density podría tener una influencia más destacada en comparación con los modelos basados en árboles, sin embargo, consideramos que la exploración de este modelo excede el ámbito de la práctica, por lo que tan solo adjuntaremos la gráfica c-ICE con PDP generada por esta para la principal característica observada en este proyecto: el alcohol.

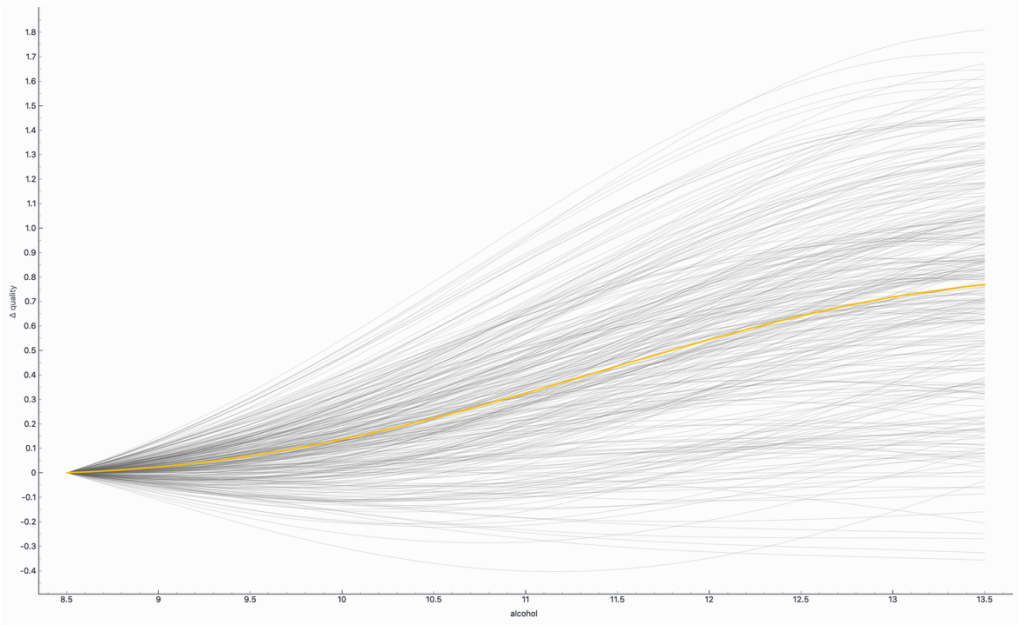

Ilustración 14: PDP y c-ICE en Orange SVM

Podemos apreciar una suavidad mucho mayor en las líneas generadas por este modelo en comparación con las generadas por el Random Forest.

5. Experimentación en Python

Complementando la exploración visual e interactiva realizada en Orange, se llevó a cabo un análisis más profundo y programático utilizando Python. Este enfoque permitió un mayor control sobre los parámetros de los modelos y las visualizaciones, así como la creación de un flujo de trabajo reproducible y documentado.

5.1. Entorno de Trabajo y Configuración de Scripts

El desarrollo y la ejecución de los análisis en Python se realizaron principalmente en **Google Colaboratory (Colab)**, un entorno basado en Jupyter Notebook que facilita la colaboración y el acceso a recursos computacionales. Se elaboró un notebook que contiene todo el proceso, desde la carga de datos hasta la generación de los gráficos de interpretabilidad.

Para asegurar la reproducibilidad y facilitar la ejecución en diferentes entornos, se consideraron dos enfoques para la obtención del dataset "White Wine Quality":

1. **Carga desde Repositorio de GitHub (para Colab)**: En el notebook de Colab, el dataset se carga directamente desde una URL pública del repositorio del proyecto en GitHub (https://raw.githubusercontent.com/jaxlvl/Exploring-ICE-PDP/main/DataBases/White-wine-quality.csv). Esto asegura que el notebook sea autocontenido y ejecutable por cualquier persona con acceso a internet.
2. **Script .py con Carga Local (para ejecución en repositorio)**: Adicionalmente, se preparó un script de Python (PDP_ICE_Analisis_Local.py) equivalente al contenido del notebook. La principal diferencia radica en que este script está configurado para cargar el dataset (White-wine-quality.csv) desde la estructura de directorios local del repositorio (ej., desde una carpeta DataBases/), lo que es ideal para ejecuciones locales o como parte de la estructura del proyecto en GitHub.

Las librerías clave de Python utilizadas para este análisis, ya mencionadas en la sección 3.3.2. han sido Pandas para la manipulación de datos, NumPy para operaciones numéricas, Scikit-learn para el modelado y la generación de PDP/ICE (PartialDependenceDisplay), y Matplotlib/Seaborn para las visualizaciones.

5.2. Entrenamiento de Modelos

Sin entrar demasiado, ya que la finalidad de este informe no es explicar el entrenamiento de modelos, se entrenaron 2 modelos de regresión:

- **Árbol de Decisión Regresor (DecisionTreeRegressor)**: Configurado con max_depth=5 y random_state=42 para facilitar la comparación con los resultados de Orange y mantener una estructura interpretable.

- **Random Forest Regresor (RandomForestRegressor)**: Configurado con n_estimators=100, max_depth=10, min_samples_leaf=5 y random_state=42 para obtener un modelo más robusto y capaz de capturar relaciones complejas.

5.3. Generación de gráficos PDP, ICE y c-ICE con Scikit-learn

Para guiar la selección de variables en los análisis PDP e ICE, se calcularon las importancias de las características utilizando el RandomForestRegressor entrenado. Las características más importantes identificadas fueron (en orden decreciente): alcohol, volatile acidity, density, free sulfur dioxide y residual sugar.

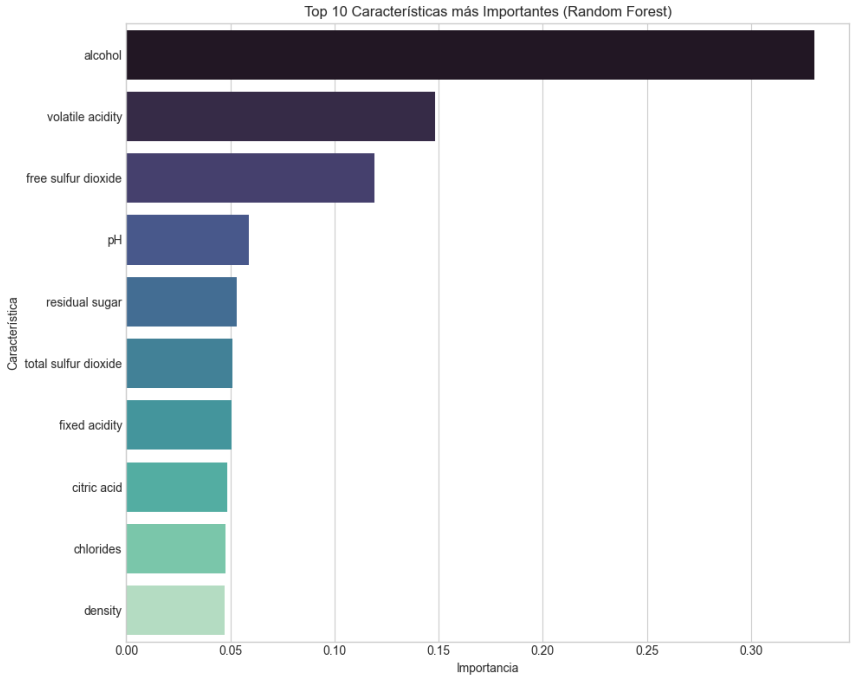

Ilustración 15: Importancia de las características del dataset

Utilizando la función PartialDependenceDisplay de Scikit-learn, se generaron los diferentes tipos de gráficos de interpretabilidad para las características seleccionadas (principalmente las más importantes y aquellas con potencial de interacción).

5.3.1. Árbol de decisión

- **PDP**: Los gráficos de dependencia parcial para características individuales (ej. alcohol) mostraron la distintiva forma escalonada, donde cada "escalón" corresponde a un punto de corte en el árbol de decisión. Esto ilustra cómo el modelo agrupa rangos de valores de una característica para asignarles una predicción promedio similar.

- **ICE y c-ICE**: Las líneas ICE individuales también fueron escalonadas. Los gráficos c-ICE ayudaron a visualizar cómo, a pesar de la naturaleza escalonada, las predicciones para diferentes instancias podían variar en magnitud dentro de las mismas "regiones" definidas por el árbol.

Podemos encontrar el resultado de estos gráficos en las ilustraciones 2, 5 y 7.

5.3.2. Random Forest

- **PDP**: Los PDPs generados a partir del Random Forest fueron notablemente más suaves, reflejando el efecto de promediar las predicciones de múltiples árboles. Esto permitió observar tendencias más graduales y relaciones no lineales más matizadas (ej. la relación entre alcohol y quality).

- **ICE y c-ICE**: Estos gráficos fueron cruciales para este modelo. Las líneas ICE individuales mostraron una dispersión considerable alrededor del PDP, indicando que el efecto de una característica no era uniforme para todos los vinos. Los gráficos c-ICE, al normalizar el punto de partida, resaltaron claramente la heterogeneidad en las *pendientes* y *formas* de las curvas individuales. Para la característica alcohol, por ejemplo, se pudo observar un "abanico" de líneas, sugiriendo que la sensibilidad de la calidad predicha al contenido de alcohol variaba significativamente entre las distintas muestras de vino, probablemente debido a interacciones con otros componentes.

Podemos encontrar el resultado de estos gráficos en las ilustraciones 1, 6 y 8.

5.3.3. Exploración de Interacciones con PDP 2D

Para visualizar directamente las interacciones entre pares de características, se generaron Gráficos de Dependencia Parcial 2D. Se seleccionaron pares como (alcohol, volatile acidity) y, de forma destacada, (alcohol, density), basándose en la importancia de las características y en hipótesis sobre la química del vino. Estos gráficos de contorno o mapas de calor mostraron cómo la predicción de la calidad (quality) variaba en función de la combinación de los valores de las dos características, revelando regiones de alta o baja calidad predicha y patrones de interacción. Por ejemplo, el PDP 2D para residual sugar y pH ilustró cómo el impacto del azúcar residual en la calidad percibida podía depender críticamente del nivel de acidez del vino.

La ilustración 4 muestra el PDP 2D que relaciona las características alcohol y volatile acidity con respecto a la característica objetivo quality, mientras que la ilustración 16 muestra la relación alcohol-densidad.

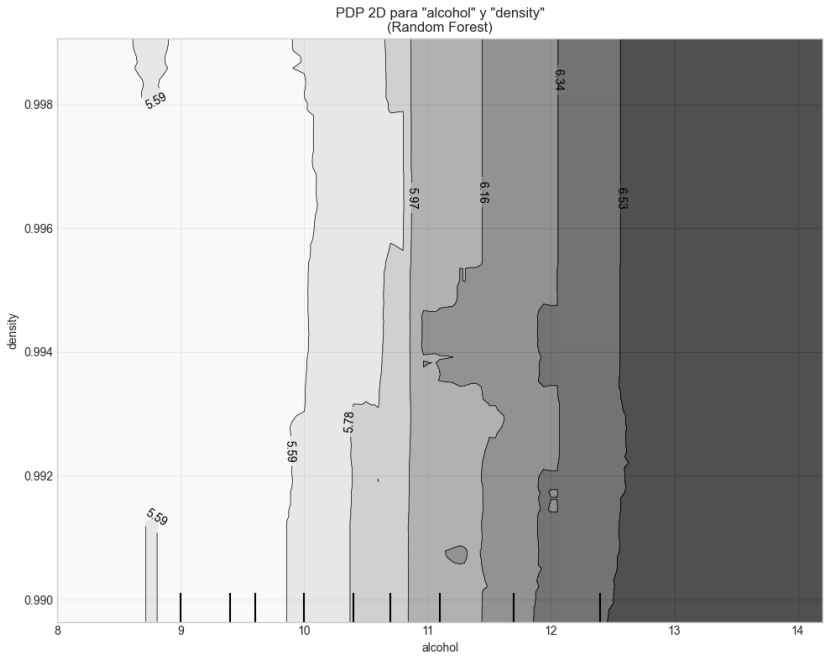

Ilustración 16:PDP 2D alcohol-density RF

6. Comparativa

6.1 PDP vs. ICE vs. c-ICE

A lo largo de los experimentos con el dataset "White Wine Quality", se ha podido observar consistentemente que, mientras los PDP ofrecen una valiosa visión del efecto promedio de una característica, los gráficos ICE son cruciales para descubrir la heterogeneidad en las respuestas del modelo, especialmente los c-ICE, facilitando una comprensión más rápida.

Para características como alcohol, el PDP mostraba una tendencia general que aumentaba la predicción de calidad, sin embargo, las líneas ICE individuales (y los c-ICE) revelaron patrones ocultos, revelando que un grupo de instancias no seguía el comportamiento promedio. Esto subraya cómo el promedio del PDP puede enmascarar comportamientos individuales importantes y la presencia de interacciones complejas.

La siguiente tabla agrupa y sintetiza las principales diferencias entre las herramientas estudiadas (omitimos los d-ICE, ya que a penas los hemos tratado durante el desarrollo de la práctica):

Aspecto	PDP (Gráfico de Dependencia Parcial)	ICE (Gráfico de Expectativa Condicional Individual)	c-ICE (Gráfico ICE Centrado)
Definición	Efecto marginal promedio de 1-2 características.	Cómo cambia la predicción de cada instancia al variar una característica.	Similar al ICE, pero cada curva se centra en y=0 para comparar formas.
Granularidad	Global (promedio sobre todas las instancias).	Local (una línea por instancia).	Local, con punto de partida normalizado.
Propósito Principal	Entender tendencia general y efecto promedio.	Descubrir efectos heterogéneos e interacciones que PDP podría ocultar.	Mejorar visualización de heterogeneidad eliminando diferencias en niveles base.
Visualización	Línea única (1D) o mapa de contorno/calor (2D).	Múltiples líneas (1D).	Múltiples líneas (1D) comenzando en y=0.
Ventaja Clave	Simplicidad, visión general.	Revela interacciones y comportamiento a nivel de instancia.	Facilita comparación de formas de curvas ICE y detección de patrones divergentes.
Limitación Clave	Oculta interacciones; supone independencia de características.	Hacinamiento visual; solo para 1D; una única característica por gráfico.	Similar al ICE; supone independencia de características.

6.2 Orange vs. Python

Ambas herramientas han resultado de gran interés y utilidad, ofreciendo una funcionalidad complementaria y no tan solapada como se podría pensar en un principio. Individualmente:

- **Orange:** Ha destacado por su facilidad de uso para una exploración rápida y visual. La construcción de flujos de trabajo es intuitiva, y la generación de PDP/ICE es directa. La capacidad de colorear líneas ICE por otra variable en el widget de Orange fue muy útil para investigar interacciones visualmente de forma preliminar. Es una herramienta con un gran poder didáctico e ideal para aprender conceptos de ML sin la barrera que supone la implementación de código y el uso de librerías especializadas. El flujo de trabajo y la interactividad de Orange ha sido una de las características más interesantes del trabajo con esta herramienta.

- **Python (con Scikit-learn):** Ofreció mayor flexibilidad, control programático sobre los parámetros de los gráficos (como subsample, cmap para PDP 2D), y la capacidad de integrar el análisis de interpretabilidad en un flujo de trabajo de ML más amplio y reproducible. La generación de gráficos para la memoria fue más personalizable, aunque la curva de aprendizaje es considerablemente más inclinada.

7. Conclusiones

Este estudio se ha centrado en el análisis y la aplicación de los Gráficos de Dependencia Parcial (PDP) y de Expectativa Condicional Individual (ICE) como métodos para mejorar la interpretabilidad de modelos de aprendizaje automático. Para ello, se han utilizado el dataset "White Wine Quality" y modelos como Árboles de Decisión y Random Forest, implementando los análisis tanto en Orange Data Mining como en Python.

A través de la experimentación, se ha podido constatar la utilidad de los PDP para obtener una visión general del efecto promedio de las características sobre la predicción de la calidad del vino. Sin embargo, el verdadero valor para una comprensión más profunda proviene de los gráficos ICE y c-ICE. Estos permiten observar la heterogeneidad en las predicciones individuales, revelando que el efecto de una misma característica, como el alcohol o el residual sugar, no es uniforme para todas las muestras de vino. En particular, el análisis de interacciones (por ejemplo, entre residual sugar y pH mediante PDP 2D, o al considerar cómo las líneas ICE divergen) demuestra que el impacto de una variable puede estar condicionado por otras, un aspecto que los PDP unidimensionales tienden a simplificar.

La forma de los gráficos también refleja la naturaleza de los modelos: los Árboles de Decisión producen las esperadas funciones escalonadas, mientras que los Random Forest generan tendencias más suaves, capaces de capturar relaciones más matizadas. En cuanto a las herramientas, Orange ha facilitado una exploración visual e interactiva inicial muy ágil, mientras que Python ha proporcionado la flexibilidad necesaria para un análisis más detallado y la personalización de las visualizaciones. Ambas herramientas se muestran complementarias.

Desde una perspectiva personal, el desarrollo de esta práctica ha subrayado que la generación de estos gráficos es solo el primer paso; el reto y el aprendizaje principal residen en su correcta interpretación y en la habilidad para extraer conocimiento útil sobre el comportamiento del modelo. Asimismo, destaca la importancia de ser consciente de las limitaciones de estas técnicas, como la suposición de independencia de características.

En definitiva, los PDP e ICE son herramientas valiosas en el ámbito de la Inteligencia Artificial Explicable. Facilitan una inspección más detallada de los modelos predictivos, lo cual es fundamental no solo para mejorar los propios modelos, sino también para fomentar una mayor confianza y comprensión en su aplicación en contextos prácticos.

8. Bibliografía

1. **Amat Rodrigo, J.** (Consultado el 10 de mayo de 2025). *Interpretación de modelos predictivos: Gráficos PDP e ICE (Python)*. Cienciadedatos.net. Recuperado de https://cienciadedatos.net/documentos/py16-interpretacion-modelos-graficos-pdp-ice

2. **Géron, A.** (2022). *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow* (3ª ed.). O'Reilly Media / Anaya Multimedia.

3. **HACARUS INC.** (2021, 7 de diciembre). *Visualizing AI: PDP Reliability using d-ICE*. HACARUS INC. AI Lab. Recuperado el 14 de mayo de 2025, de https://hacarus.com/ai-lab/20211207-d-ice/

4. **Maheshwarappa, A.** (2020, 11 de octubre). *Explainable AI with ICE (Individual Conditional Expectation) Plots*. Medium. Recuperado el 1 de mayo de 2025, de https://abhishek-maheshwarappa.medium.com/explainable-ai-with-ice-individual-conditional-expectation-plots-c71e8fc1f1c2

5. **Maheshwarappa, A.** (2020, 27 de septiembre). *Explainable AI with PDP (Partial Dependence Plot)*. Medium. Recuperado el 5 de mayo de 2025, de https://abhishek-maheshwarappa.medium.com/explainable-ai-with-pdp-partial-dependence-plot-fecf09b0e947

6. **Molnar, C.** (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Recuperado el 3 de mayo de 2025, de https://christophm.github.io/interpretable-ml-book/ (Capítulo específico sobre ICE: https://christophm.github.io/interpretable-ml-book/ice.html)

7. **Scikit-learn Developers.** (s.f.). *1.11. Partial Dependence and Individual Conditional Expectation Plots*. Scikit-learn User Guide. Recuperado el 3 de mayo de 2025, de https://scikit-learn.org/stable/modules/partial_dependence.html

8. **[Nombre del Canal de YouTube o Autor del Vídeo]**. (2024, 29 de abril). *[Partial Dependence (PDPs) and Individual Conditional Expectation (ICE) Plots | Intuition and Math]*. YouTube. Recuperado el 16 de abril de https://youtu.be/dEhbS37Kglc?si=eluK8-pQOPbI8FaP