|  |
| --- |
| Universidad Politécnica de Madrid |
| Actividad 2 SHAP & LIME |
| Inteligencia Artificial Explicable   |  | | --- | |  | | Sergio Arroni del Riego – s.arroni@alumnos.upm.es  Autor: Sergio Arroni del Riego  Descripción: Actividad 2, SHAP & LIME  Fecha: 20/11/2023 |   Grupo 1 |



Master Universitario en Inteligencia Artificial - MUIA

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**SHAP & LIME – Grupo 1**

**Autores:**

**Sergio Arroni del Riego**

Graduado en Ingeniería Informática de Software

|  |  |  |
| --- | --- | --- |
| **Versión** | **Fecha** | **Comentarios de versión** |
| V1 | 20/11/2023 | Finalizadas las respuestas |
| V2 | 30/11/2023 | Finalizada la revisión |

Contenido

[Pregunta a) 3](#_Toc151537265)

[Pregunta b) 3](#_Toc151537266)

[Pregunta d) 3](#_Toc151537267)

[Pregunta e) 3](#_Toc151537268)

[Pregunta f) 3](#_Toc151537269)

### Pregunta a)

1. Enfoque de eliminación de características
2. Comportamiento del modelo
3. Técnica de resumen

### Pregunta b)

Las opciones de diseño específicas incluyen:

* El diseño de arquitecturas de red "interpretables".
* La incorporación de una red de atención.
* El uso de una red de módulos neuronales para dividir el problema en subproblemas.

Además, se propone el uso de métricas de evaluación automáticas como la eliminación y la inserción para medir la calidad de las explicaciones.

### Pregunta d)

Es la incorporación de una red de atención para aprender dónde mirar en una imagen antes de producir cada palabra de la descripción. Esta opción de diseño permite que el modelo aprenda a enfocarse en áreas específicas de la imagen, lo que puede proporcionar una explicación más detallada y localizada de la importancia de ciertos rasgos.

### Pregunta e)

Las dos ventajas de las explicaciones locales son:

1. Permiten una comprensión detallada de cómo el modelo toma decisiones en un nivel granular, lo que puede ser crucial para identificar y corregir posibles sesgos o errores en el proceso de toma de decisiones.
2. Permiten a los usuarios comprender exactamente qué características o regiones de una entrada contribuyen a una decisión específica, proporcionando una mayor transparencia y confianza en el modelo.

### Pregunta f)

Esta afirmación se basa en la interpretación de los valores de Shapley en SHAP como una aplicación específica de la técnica de ajustes ponderados empleada en LIME. Ambos métodos comparten la idea de asignar importancias locales a las características, utilizando ajustes ponderados para explicar las predicciones de modelos en entornos locales. En LIME, esto se logra mediante la creación de modelos interpretables locales, mientras que en SHAP se emplean valores de Shapley para asignar contribuciones de características de manera justa. Los valores de Shapley son equivalentes a LIME con