Laboratorio: Árboles y random forest para regresión y clasificación

**Objetivos**

Mediante este laboratorio se pretende que aplique los conocimientos adquiridos en los temas de Árboles y Random forest trabajando con un conjunto de datos médico.

**Descripción**

El conjunto de datos con el cual vamos a trabajar se encuentra en el siguiente link:

<https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.names>

<https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv>

En esta primera actividad se trata de familiarizarse con los pasos generales a realizar para generar un modelo de aprendizaje automático. Podemos resumir estos pasos en:

* Análisis descriptivo de los datos
* Tratamiento de valores faltantes o nulos
* Aplicación del algoritmo de aprendizaje automático
* Evaluación

Para ello, se deben seguir los pasos que indicamos a continuación.

**Metodología**

1. **Análisis descriptivo de los datos:** Se adjunta un archivo .txt con la descripción de cada una de las columnas.
   1. Comentar de manera general qué se puede observar en la figura 1.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 1. Resumen del conjunto de datos

* 1. En la figura 2 se pueden observar las estadísticas de las columnas numéricas. ¿Si se tienen 768 observaciones, a qué conclusiones podríamos llegar con estos datos? ¿Podríamos eliminar alguna variable?Interfaz de usuario gráfica

     Descripción generada automáticamente con confianza media

Figura 2. Estadísticas de columnas numéricas.

* 1. En la figura 3 se muestran los histogramas de cada una de las columnas. ¿Qué se puede decir de la distribución de las variables?

Gráfico

Descripción generada automáticamente

Figura 3. Histogramas de cada columna del conjunto de datos.

* 1. En la figura 3 tenemos el mapa de calor de la matriz de correlaciones, por favor revise cuáles son las variables que mayor correlación tienen y si se puede eliminar alguna columna con base en este mapa de calor. Justifique su respuesta:

Imagen que contiene Gráfico

Descripción generada automáticamente

Figura 4. Matriz de correlación

1. **Tratamiento de valores faltantes**:
2. en la figura 5 se puede observar que en este conjunto de datos no existen valores faltantes.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 5. Columnas con valores faltantes o NA.

En la figura 6 podemos ver que existen columnas con ceros. Puede comentar ¿Qué puede estar ocurriendo con este conjunto de datos?

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 6. Cantidad de ceros por columna.

¿Puede completar la descripción de la Figura 7? ¿Qué acciones se realizan?

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 7. ¿?

1. **Entrenamiento de algoritmos**
2. Al aplicar árboles de decisión y Random Forest con el dataframe en los que se **eliminaron todas las filas con valores faltantes**. Las métricas obtenidas a partir de la matriz de confusión se ven en las figuras 8 y 9:

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 8. Métricas obtenidas de árboles de decisión. Validación cruzada 5 folds.

Tabla

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 9. Métricas obtenidas al aplicar Random Forest. Validación cruzada 5 folds.

Comente los resultados.

1. Al aplicar árboles de decisión y Random Forest con el dataframe en los que se **eliminaron columnas con un % de valores faltantes**. Las métricas obtenidas a partir de la matriz de confusión se ven en las figuras 10 y 11:

Tabla

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 10. Métricas obtenidas al aplicar árboles de decisión. Validación cruzada 5 folds.

Tabla

Descripción generada automáticamente Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 11. Métricas obtenidas al aplicar Random Forest. Validación cruzada 5 folds.

Comente los resultados.

1. Al aplicar árboles de decisión y Random Forest con el dataframe en los que **se imputan valores con la media**. Las métricas obtenidas a partir de la matriz de confusión se ven en las figuras 12 y 13:

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 12. Métricas obtenidas al aplicar árboles de decisión. Validación cruzada 5 folds.

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 13. Métricas obtenidas al aplicar Random Forest. Validación cruzada 5 folds.

Comente los resultados.

1. Al aplicar árboles de decisión y Random Forest con el dataframe en los que **se imputan valores con una función de interpolación**. Las métricas obtenidas a partir de la matriz de confusión se ven en las figuras 14 y 15:

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 14. Métricas obtenidas al aplicar árboles de decisión. Validación cruzada 5 folds.

Tabla

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 15. Métricas obtenidas al aplicar Random Forest. Validación cruzada 5 folds.

Comente los resultados.

1. **Comentarios adicionales**

* Según los resultados obtenidos, ¿Cuál sería la forma más correcta de trabajar con los valores faltantes teniendo en cuenta que son datos médicos?
* Comente si tiene alguna sugerencia para mejorar los resultados.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Realiza el análisis descriptivo de los datos de manera adecuada | 2 | 20% |
| Criterio 2 | El tratamiento de los valores faltantes es correcto | 2 | 20% |
| Criterio 3 | Analiza muy bien los resultados de aplicar los clasificadores al trabajar con la eliminación de columnas o filas con valores faltantes | 2 | 20% |
| Criterio 4 | Analiza muy bien los resultados de aplicar los clasificadores al trabajar con imputación de valores faltantes con la media y la interpolación. | 2 | 20% |
| Criterio 5 | Los comentarios sobre los resultados son pertinentes y adecuados | 2 | 20% |
|  |  | **10** | **100 %** |

**Extensión** máxima de la actividad: 20 páginas.