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

En la figura 1 se observan todas las variables que forman el dataset junto a sus tipos, todas las variables son numéricas, la mayoría de carácter entero, las únicas variables en formato decimal flotante son el IMC (índice de masa muscular), y la variable *funcionPediDiabe* que representa la probabilidad que una persona tenga diabetes debido a antecedentes familiares.

* 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.

Los datos anteriores son las medidas de tendencia para analizar todas las variables numéricas del dataset.

Si bien, esta información es muy preliminar para eliminar alguna columna (variable) de un dataset, sí nos permite hacer énfasis en las variables que debemos estudiar más detenidamente.

Por ejemplo, observe que las variables *IMC*, *concentracionGlucosa*, *presionArterialSistolica*, tienen como valor mínimo 0. En la realidad, dichos datos no pueden formar parte de una medición real, podemos considerarlos outliers y debemos analizar sí es conveniente tenerlos en el dataset o no.

Además, la columna *insulinaSerica* tiene un valor máximo totalmente alejado de los valores presentados en los cuartiles anteriores a él, debemos analizar más detenidamente este valor para determinar cómo proceder con el outlier.

* 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.

Se observan tres tipos de distribuciones:

En primer lugar, vemos que las variables: *IMC*, *presionArterialSistólica* y *concentracionGlucosa* siguen una distribución normal o que podemos aproximarla como *normal*.

En cambio, las variables *nEmbarazos*, *pliegueCutaneo*, *insulinaSerica*, *funcionPediDiabe* y edad tienen una distribución con una cola en el lado derecho de la distribución, estas se conocen como *'right-skewed'* y generalmente muestran un comportamiento en el que *moda < mediana < media*.

Por último, observe que la variable diabetes, nuestra variable a predecir, tiene una distribución categórica entre dos valores: 0 y 1.

* 1. En la figura 4 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

Al utilizar el comando *df.corr()* para visualizar el valor de las correlaciones numéricas tenemos el siguiente resultado:

Interfaz de usuario gráfica

Descripción generada automáticamente

Ninguna de las variables presenta una correlación alta (>0.8), recordemos que una alta correlación entre variables podría indicar que existe información repetida por lo que podríamos realizar una reducción de dimensionalidad. Debido a la baja correlación entre variables, la eliminación no se recomienda para este dataset.

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.

El código anterior suma todos los valores que son cero por cada columna.

Algunos cero tienen sentido mientras que para otros debe aplicarse un tratamiento para eliminar valores nulos.

Aunque no existe ninguna entrada de datos con valores nulos, hay variables en los que un valor cero no tiene ningún sentido.

En ese sentido, deben analizarse los ceros de las columnas en: ***pliegueCutaneo***, ***insulinaSerica***, ***IMC***, ***presionArterialSistolica*** y ***concentracionGlucosa***.

Estos valores cero son outliers de los valores esperados en cada columna, como nos mostró el df.describe() del enunciado a)

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

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 7. Reemplazamiento de los valores faltantes de cero a *NaN* (not a number) en las columnas donde el valor cero es un *outlier*.

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.

La eliminación de todas las filas con valores faltantes o nulos, es muy significativa, de 768 registros quedan solo 79 \* 5 = 395 registros. Esta eliminación influye a la hora de calcular el *accuracy* de los modelos de ML. El F1-Score es una métrica que combina los resultados de *precision* y *recall* y es muy útil al tener un conjunto de datos no balanceado como el presentado acá donde hay más valores de personas sin diabetes (0) que con diabetes ( 1) como se mostraba en el histograma de la variable en la parte 1.

En cuanto a las gráficas del accuracy, se observa que la validación cruzada se desempeña de peor manera que el accuracy de los propios datos de entrenamiento, esto nos permite afirmar que no es el mejor modelo para predecir resultados en un conjunto de datos no conocidos por el modelo.

Por último, en cuanto a los modelos de ML, se observa mejores resultados en las métricas al utilizar Random Forests que árboles de decisión, lo cual es lo esperado ya que el modelo de Random Forests combina muchos árboles de decisión sobre diversos subconjuntos del dataset y luego usa la moda para predecir la variable de salida, *diabetes*. Por tanto, el modelo de Random Forests siempre se desempeñará mejor.

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.

La eliminación de columnas con un alto porcentaje de valores faltantes no mejora los resultados anteriores, esto es debido a que la baja correlación entre las variables afecta los resultados del modelo al eliminarlas. Aunque también es importante considerar que no se tiene información del porcentaje umbral que se utilizó para considerar la eliminación de las columnas. Las métricas mostradas en los resultados anteriores nos permiten afirmar que la eliminación de columnas no es lo más adecuado en este dataset por lo que deben considerarse otras opciones para el tratamiento de valores nulos. Una segunda consideración es modificar el umbral o porcentaje para la eliminación de las columnas.

Las gráficas del desempeño del accuracy también permiten identificar que no hay una mejoría significativa y que el modelo se comporta de forma muy distinta según el subconjunto de datos del dataset tomados en el modelo.

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.

En este caso los valores nulos o faltantes se han sustituido por un promedio de los valores de la columna correspondiente. Si bien es cierto, esto es una técnica de tratamiento de valores nulos, los resultados de las métricas muestran que no hay una mejora significativa comparada a las otras técnicas anteriores.

La técnica de Random Forests sí presenta una mejoría del F1-Score para la predicción del valor 0. Sin embargo, el valor a la hora de predecir 1 hay un *recall* bajo, aproximadamente el 50%, esto es la proporción de predicciones correctas entre todas las predicciones marcadas con positivo (1). En el caso de un dataset de enfermedades, donde queremos la mayor relación de verdaderos positivos, necesitamos un *recall* alto.

Las gráficas muestran que en general, el modelo se desempeña en una peor manera ante todo el dataset, esto puede comprobarse al observar la tendencia de más baja accuracy en la validación cruzada.

En conclusión, este modelo mejora con Random Forests el F1-Score de la predicción de ceros, pero dada la importancia de tener una alta proporción de verdaderos positivos no puede utilizarse en este dataset médico.

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.

Cuando se utiliza una función de interpolación para obtener los valores faltantes se observa que es cuando se tienen mejores resultados, esto era esperable debido a que la interpolación estima el valor faltante comparando con los valores de su alrededor.

Esto produce que los resultados de las métricas mejoren en conjunto, en especial el modelo de Random Forests al ser un mejoramiento de los árboles de decisión.

Si consideramos especial atención a la métrica del *recall* en específico veremos que son los árboles de decisión los que han obtenido mejor desempeño.

Observe también que la grafica muestra que el *accuracy* de los resultados de la validación cruzada son más altos que el resultado del accuracy en un dataset de entrenamiento. Esto nos permite tener una idea de como se comportará el modelo con el dataset en su totalidad arrojando buenos 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?

Teniendo en cuenta que son datos médicos, se observa que hay un mejor desempeño del modelo y de los resultados en las métricas al sustituir los valores nulos con una función de interpolación. Al ser un dataset médico se debe tener especial consideración en la métrica de *recall* ya que se busca tener la mayor tasa de verdaderos positivos.

* Comente si tiene alguna sugerencia para mejorar los resultados.

Como recomendación final, se puede mejorar el modelo de Random Forest presentado al final, con técnicas de interpolación, por medio de la configuración de sus hiperparámetros. Se recomienda utilizar una búsqueda de hiperparámetros utilizando *GridSearchCV* , ajustando la profundidad del árbol, el tamaño de los subconjuntos del dataset, considerar la poda, y también considerar otras funciones de interpolación.

**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.