TP 3 Aprendizaje Supervisado y/o NO Supervisado

Modelos de ML para Aprendizaje Supervisado

Modelos Predictivos (para Tasa de Mortalidad):

- *Regresión Lineal Múltiple:* *para predecir la tasa de mortalidad en función de diferentes variables como la disponibilidad de especialidades médicas, características poblaciones, etc. Este modelo permitirá entender cómo las variables independientes afectan la tasa de mortalidad.*

Modelos de Clasificación:

- *Árboles de Decisión y Random Forest:* *Estos modelos pueden ser útiles para clasificar provincias en categorías basadas en la disponibilidad de especialidades médicas y la tasa de mortalidad. También permiten entender qué características son más importantes para la clasificación.*

- *Máquinas de Soporte Vectorial (SVM):* *Pueden usar SVM para clasificar provincias según la necesidad de especialidades médicas y la tasa de mortalidad. Este modelo es eficaz en casos con muchas características y un conjunto de datos pequeño.*

Modelos de ML para Aprendizaje Supervisado

Modelos de Clustering:

- *K-means Clustering:* *Para segmentar provincias en grupos basados en similitudes en la tasa de mortalidad y la disponibilidad de especialidades. Esto puede ayudar a identificar regiones con necesidades similares.*

- *Algoritmos Jerárquicos:* *Para una segmentación más flexible, donde los grupos se forman de manera jerárquica, permitiendo una comprensión más profunda de cómo se relacionan las provincias entre sí.*

TP 3 Aprendizaje Supervisado y/o NO Supervisado

Aprendizaje Supervisado

1. Preparación de los datos para aplicar modelos de regresión/clasificación:

a. Generar a partir del dataset, los conjuntos de train y test.

b. Analizar las proporciones consideradas para cada conjunto con respecto al total del dataset.

c. En el caso de optar por un algoritmo de clasificación, para cada conjunto (train y test), verificar la distribución de la variable objetivo (se debe buscar que las clases estén balanceadas, es decir, con distribuciones sean similares entre sí).

2. Predicción con un modelo baseline:

a. Entrenar un modelo "baseline", esto es, un modelo que sea técnicamente lo más simple posible, para con ello tener un punto de partida con el cual comparar modelos más complejos. Fijar la semilla aleatoria para hacer repetible el experimento.
(Hint: se puede usar, por ejemplo, la clase `DummyRegressor` de scikit-learn)

Documentación:

- https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyRegressor.html

b. Evaluar sobre el conjunto de entrenamiento y test reportando:

i. Mean Absolute Error (MAE)

ii. Mean Squared Error (MSE)

iii. Root Mean Squared Error (RMSE)

iv. R-squared (R^2)

c. Reflexionar sobre cuál métrica es conveniente optimizar en este problema en función de la elección del algoritmo de regresión/clasificación. Ej. ¿Sería el RMSE una buena métrica? ¿Qué métrica elegirían y por qué?

3. Predicción con modelos supervisados:

a. Entrenar modelos de regresión/clasificación para predecir la variable objetivo. Fijar la semilla aleatoria para hacer repetible el experimento

b. Evaluar sobre el conjunto de entrenamiento y test reportando:

i. MAE

ii. MSE

iii. RMSE

iv. R^2

c. Elaborar conclusiones en base a la métrica a optimizar y comparar con el modelo baseline.

4. Ajuste por hiperparámetros:

a. Para el "mejor modelo" obtenido en el punto anterior, seleccionar valores para los hiperparámetros principales de dichos modelos (ajustar por lo menos con 3 parámetros). Utilizar grid-search y k-fold cross-validation.

b. Mencionar el mejor modelo obtenido de la Optimización de Hiperparámetros y con cuáles parámetros se obtuvo ese resultado.

c. Con el mejor modelo obtenido realizar las predicciones sobre test y val.

d. Reportar las métricas del mejor modelo, incluyendo las evaluaciones de MAE, MSE, RMSE y R^2.

TP 3 Aprendizaje Supervisado y/o NO Supervisado

Aprendizaje NO Supervisado

1. Preparación de los datos para aplicar modelos de clustering/agrupamiento:

a. Limpiar y preprocesar los datos: verificar la presencia de valores nulos, escalado de características y manejo de outliers.
b. Seleccionar las variables o características relevantes que se utilizarán para el agrupamiento.
c. Dividir el dataset en un conjunto de entrenamiento y test (si es necesario, aunque generalmente en aprendizaje no supervisado se utiliza todo el conjunto de datos para la modelización inicial).

2. Aplicación de un modelo baseline:

a. Entrenar un modelo de agrupamiento sencillo como K-Means con un número predeterminado de clusters (por ejemplo, 2 o 3). Fijar la semilla aleatoria para hacer repetible el experimento.
b. Evaluar el modelo utilizando métricas como Coeficiente de Silhouette
c. Reflexionar sobre las agrupaciones obtenidas: ¿Son las diferencias entre clusters significativas? ¿Qué tan bien separadas están las agrupaciones según las métricas seleccionadas?

3. Aplicación de modelos no supervisados:

a. Entrenar otros modelos de clustering como DBSCAN o Hierarchical Clustering
b. Evaluar el desempeño de estos modelos utilizando la misma métrica que en el punto 2
c. Comparar los resultados obtenidos con los diferentes modelos y elaborar conclusiones sobre cuál modelo agrupa mejor los datos en función de la/s métrica/s seleccionada/s.

4. Ajuste de hiperparámetros y evaluación:

a. Mencionar el mejor modelo obtenido tras la optimización de hiperparámetros y especificar con qué parámetros se obtuvo ese resultado.
b. Visualizar los clusters en función de las variables más importantes, utilizando técnicas de reducción de dimensionalidad como PCA para facilitar la interpretación.
c. Reportar las métricas finales del modelo optimizado, enfocándose en la calidad de la agrupación y la estabilidad de los clusters.

TP 3 Aprendizaje Supervisado y/o NO Supervisado

Características que debe cumplir el entregable:

- Se debe ir desarrollando cada punto en la misma notebook donde se escriba el código. Dicho notebook debe contar con un índice, con sus diferentes apartados y el código debe ser fácil de leer, estar probado y comentado (esto último, en función de la necesidad).
- Se debe enviar el link directo del archivo .ipynb ó alternativamente subir el entregable a un repositorio GitHub mediante la integración con Google Colab. Recordar que al compartir el notebook, queden habilitados los permisos de edición, para poder dejar comentarios/correcciones.
- Tener en cuenta que si bien, pueden practicar con diversos modelos, se debe dejar en el entregable sólo aquello que sea relevante.
- Luego de cada modelo es importante poder obtener una conclusión y/o breve interpretación de los resultados.

Fecha de Entrega: 16/09