



Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura en Inteligencia Artificial

Aprendizaje Automático 1

Trabajo Práctico: Predicción de Lluvia en Australia.

Objetivo

Familiarizarse con la librería scikit-learn y las herramientas que brinda para el pre-procesamiento de datos, la implementación de modelos y la evaluación de métricas, y con TensorFlow para el entrenamiento de redes neuronales.

Dataset

El dataset se llama weatherAUS.csv y contiene información climática de Australia de los últimos diez años, incluyendo si para el día siguiente llovió o no y la cantidad de lluvia en las columnas **'RainTomorrow'** y **'RainfallTomorrow'**. El objetivo es la predicción de estas dos variables en función del resto de las características que se consideren adecuadas.

Tiene una columna 'Location' que indica la ciudad y el objetivo es predecir la condición de lluvia en las ciudades de Sydney, SydneyAirport, Canberra, Melbourne y MelbourneAirport (costa sureste). Pueden considerarse como una única ubicación. **Descartar** el resto de los datos.

Consignas

1. Armar grupos de hasta dos personas para la realización del trabajo práctico. Dar aviso al cuerpo docente del equipo. En caso de no tener compañero, informar al cuerpo docente.
2. Realizar un análisis descriptivo, que ayude a la comprensión del problema, de cada una de las variables involucradas en el problema detallando características, comportamiento y rango de variación.
Debe incluir:
 - Análisis y decisión sobre datos faltantes
 - Visualización de datos (por ejemplo histogramas, scatterplots entre variables, diagramas de caja)
 - ¿Está *balanceado* el dataset?
 - Codificación de variables categóricas (si se van a utilizar para predicción).
 - Matriz de correlación
 - Selección de características para la predicción.
 - Estandarización de datos.
3. Implementar la solución del problema de regresión con regresión lineal múltiple.
 - Probar con el método LinearRegression.
 - Probar con métodos de gradiente descendiente.
 - Probar con métodos de regularización (Lasso, Ridge, Elasticnet).
 - Obtener las métricas adecuadas (entre R2 Score, MSE, RMSE, MAE, MAPE).
4. Implementar la solución del problema de clasificación con regresión logística.
 - Obtener las métricas adecuadas (entre Accuracy, precision, recall, F1 Score, ROC-AUC, entre otras).

5. Implementar un modelo base para clasificación y uno para regresión.
6. Implementar las soluciones con una red neuronal.
 - Obtener las métricas adecuadas.
7. Optimizar la selección de hiperparámetros.
 - Probar validación cruzada.
 - Utilizar grid search, random search u optuna. Justificar su uso.
8. Implementar explicabilidad del modelo.
 - Utilizar SHAP o similar.
9. MLOps (a definir).
10. Escribir una conclusión del trabajo

Entregas parciales

Las entregas parciales se realizan mediante GitHub (suben el código y nos devuelven el link del repositorio). Se deben hacer commits con el asunto "Entrega hasta ítem x" (se pueden hacer commits parciales).

Hasta el 15/09: ítem 1. Aviso por correo.

Hasta el 13/10: ítems 2 y 3. Entrega de notebook.

Hasta el 27/10: ítem 4. Entrega de notebook.

Hasta el 20/11: ítem 5 y 6. Entrega de notebook.

Hasta el 7/12: ítems 7, 8, 9 y 10. Entrega de notebook, app.py y MLOps.

Los ítems se pueden ir perfeccionando a medida que se va avanzando, aunque no se tiene la misma consideración para la nota si fue editado luego de la fecha de entrega. **Pero sí importa.**