Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura en Inteligencia Artificial

Aprendizaje Automático 1

Trabajo Práctico para condición de Libre: Predicción de lluvia en Australia.

Objetivos

Familiarizarse con la biblioteca scikit-learn y las herramientas que brinda para el pre-procesamiento de datos, la implementación de modelos y la evaluación de métricas, con TensorFlow para el entrenamiento de redes neuronales y con docker para la puesta en producción del modelo seleccionado como el más adecuado, entre otras.

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 **Albury, Sydney, SydneyAirport, Canberra, Melbourne y MelbourneAirport**. Pueden considerarse como una única ubicación. **Descartar el resto de los datos.**

Consignas

1. Para todos los ítems, incorporar una cantidad de texto adecuado en forma de comentarios, ya sea para la comprensión del código (usualmente una línea de comentario por cada celda) como para explicar las decisiones tomadas a lo largo del trabajo (por ejemplo, la justificación de la imputación de valores faltantes, la elección de las métricas adecuadas, entre otros).

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? ¿Por qué creen que hacemos esta pregunta?
   - Codificación de variables categóricas (si se van a utilizar para predicción).
   - Matriz de correlación de variables.
   - Estandarización de datos.
   - Validación cruzada train - test. Realizar una división del conjunto de datos en conjuntos de entrenamiento y prueba (y si se quiere, se puede incluir validación, que luego será útil) **en el MOMENTO donde usted lo crea adecuado.**

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, Elastic Net**).

- Obtener las métricas adecuadas (entre R2 Score, MSE, RMSE, MAE, MAPE, elegir) tanto para entrenamiento como para prueba. ¿Por qué para ambos conjuntos?

4. Implementar la solución del problema de clasificación con regresión logística.
    - Obtener las métricas adecuadas (entre Accuracy, precision, recall, F1 Score, otras).
    - Trazar curvas ROC. Comente cuáles serían los umbrales adecuados a utilizar.

5. Implementar las soluciones con una red neuronal.
    - Obtener las métricas adecuadas.

6. Optimizar la selección de hiperparámetros. Elegir un sólo método para hacerlo. Justificar la selección de ese método y de los hiperparámetros a recorrer.

7. Comparación de modelos.
    - Incluyan en su análisis una comparación de modelos: de todos los modelos de regresión, ¿cuál es el mejor? Escoger u**na métrica adecuada p**ara poder compararlos. Lo mismo con los de clasificación.

8. Implementar **explicabilidad** del modelo.
    - Utilizar SHAP o similar.

9. MLOps

Para realizar el deployment del trabajo deben utilizar Docker. Dentro del repo del TP, ademas del notebook con el desarrollo, deben agregar una carpeta "docker" que contenga:

- script de inferencia (debe llamarse **inferencia.py**)
- requeriments.txt (unicamente incluir librerias necesarias para la inferencia)
- binarios con pipeline o modelo, imputers y scalers (.pkl, .joblib, pueden escoger el formato de serializacion que consideren adecuado)
- Dockerfile
- readme.md con instrucciones para poder construir la imagen de docker y ejecutar el container (docker build y docker run)
- todo lo que consideren necesario para realizar la inferencia, **no incluir archivos/carpetas innecesarios, ni la imagen de docker.**

10. Escribir una conclusión del trabajo

Entrega

Enviar un link con el repositorio donde esté subido el notebook del trabajo y docker al correo electrónico joelspak45@gmail.com, por lo menos con 24 hs de anticipación al horario de la defensa pactado.

En la defensa, con el trabajo ya revisado por el cuerpo docente, se harán preguntas sobre lo realizado, haciendo hincapie en errores pero también para que detallen cuestiones particulares sobre lo realizado, con enfoque teórico incluido. Tener en cuenta que la nota depende tanto del trabajo como de la defensa, si no se demuestra en la defensa conocimiento suficiente, por más que el trabajo esté perfecto, la condición será de desaprobado.