Universidad Católica Boliviana "San Pablo"
Unidad Académica Cochabamba
Departamento de Posgrado
Ing. Mauricio Alejandro Quezada Bustillo

DML-004 – GUÍA DE LABORATORIO 6

Despliegue y comparación de 4 endpoints (Best Model + 3 modelos) en Azure ML con 100 predicciones binarias

Objetivo

- Reutilizar el Best Model obtenido en la Guía 5 (AutoML).
- Seleccionar otros tres modelos del ranking de AutoML.
- Desplegar 4 endpoints (uno por modelo) como Managed Online Endpoints en Azure ML.
- Enviar 100 registros para inferencia binaria a cada endpoint y comparar Accuracy/Precision/Recall/F1 entre modelos.

1) Elegir los 4 modelos

Desde el ranking de AutoML en base a la guía 5:

- Modelo A: Best Model.
- Modelo B: segundo del ranking.
- Modelo C: intermedio.
- Modelo D: lejano al tope.

2) Despliegue como Managed Online Endpoints

1. En ml.azure.com → Automated ML → tu experimento → Run completado → Models.
2. En cada modelo (A, B, C, D), usa Deploy → Real-time endpoint.
3. Crea un endpoint distinto por modelo (ep-best, ep-top2, ep-mid, ep-low).
4. Selecciona instance type pequeño (ej. Standard_F2s_v2).
5. Espera estado Healthy.

3) Formato del request y prueba rápida

Formato JSON típico:

```
{
  "input_data": {
    "columns": ["col1","col2","..."],
    "data": [[v11, v12, "..."]]
  }
}
```

Universidad Católica Boliviana "San Pablo"
Unidad Académica Cochabamba
Departamento de Posgrado
Ing. Mauricio Alejandro Quezada Bustillo

4) Envío de 50-100 registros con Postman o Codigo

URL: <SCORING_URI_DE_TU_ENDPOINT>
Headers:
Content-Type: application/json
Authorization: Bearer <API_KEY_DEL_ENDPOINT>
azureml-model-deployment: v1 (opcional)

Body (JSON) Ejemplo:

```
{
  "input_data": {
    "columns":
["age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal"],
    "data": [
      [63,1,3,145,233,1,0,150,0,2.3,0,0,1],
      [37,1,2,130,250,0,1,187,0,3.5,0,0,2],

      ……. Hasta 50 a 100 datos
    ]
  }
}
```

5) Comparación y análisis

1. Tabla de métricas (Accuracy, Precision, Recall, F1).
2. Matriz de confusión.
3. Conclusión: verificar si Best Model mantiene superioridad.

6) Entregables

- Capturas de los 4 endpoints.
- Capturas con las predicciones.
- **Documento breve (PDF/Word)** con:
    - Ranking original y elección de los 4 modelos (cita run de AutoML).
    - Configuración de despliegue (endpoint y deployment por modelo).
    - Comparativa de métricas + conclusión.

7) Troubleshooting

- 400/422: revisa esquema de columnas.
- 401: usa Primary/Secondary key del endpoint.
- Timeout: dividir batch en 50.
- Región: mantener misma región del workspace..