Informe Técnico – CONCLUSIONES

Análisis Exploratorio y Modelado Predictivo del TEA en Niños

1. Conclusiones del Análisis Exploratorio de Datos (EDA)

Durante el análisis exploratorio se identificaron patrones claros en la relación entre las variables del cuestionario AQ-10 y la variable objetivo Clase_TEA. Entre los hallazgos más relevantes:

- Las variables con mayor proporción de respuestas afirmativas en niños con TEA fueron A4_Score, A9_Score, A10_Score y A8_Score, lo que indica su fuerte relación con la condición.
- La edad promedio del conjunto fue de 6 años, y se encontró un pequeño número de valores nulos, que fueron imputados con la mediana.
- Se eliminaron variables como Resultado_Diagnóstico, Grupo_Edad o Etnia por contener redundancias, problemas de representación local o riesgo de sesgo.
- Se observaron correlaciones relevantes entre ciertas preguntas del cuestionario:
    - A3_Score y A1_Score presentan una correlación de **0.71**, lo que sugiere que ambas capturan una sensibilidad sensorial y de atención conjunta.
    - A10_Score y A9_Score tienen una correlación de **0.55**, lo que podría reflejar una dimensión social compartida (dificultades en interacciones).

      Estas correlaciones indican que algunas variables pueden estar representando dimensiones similares del comportamiento. En modelos como **Regresión Logística**, esto podría causar colinealidad, pero **en modelos como Árboles o Random Forest no representa un problema** significativo.

Estas conclusiones guiaron la selección de variables y la limpieza del dataset, resultando en una estructura clara de 15 columnas numéricas sin valores nulos.

---

2. Conclusiones del Modelado Predictivo

Se entrenaron y compararon tres modelos supervisados:

Regresión Logística

- **Accuracy:** 98%
- **Recall (TEA):** 100%
- **F1-score (TEA):** 0.98

- **Fortaleza:** Alta sensibilidad, sin falsos negativos.
- **Limitación:** Puede sobreajustarse al conjunto de test si los datos están desequilibrados.

Árbol de Decisión (profundidad limitada)

- **Accuracy:** 78%
- **F1-score:** 0.79
- **Ventaja:** Alta interpretabilidad.
- **Limitación:** Menor rendimiento general. Cometió tanto falsos positivos como falsos negativos.

Random Forest (Optimizado)

- **Hiperparámetros:**
  - n_estimators=200
  - max_depth=10
  - min_samples_split=5
- **Validación cruzada (5-fold):** Accuracy promedio ≈ 94.8% ± 1.5% (precisión)
- **F1-score (TEA):** 0.90, reflejando un alto balance entre precision y recall.
- **Recall (TEA):** 0.93, lo que indica una excelente capacidad para evitar falsos negativos.
- **AUC-ROC:** 0.99, lo que muestra una gran habilidad para distinguir correctamente entre las clases.
- **Curva Precisión-Recall:** Precisión media ≈ 0.99

Este último modelo demostró ser el más equilibrado y generalizable para la tarea.

Justificación del Modelo Seleccionado

Aunque la Regresión Logística mostró un rendimiento sobresaliente sobre el conjunto de test (98% de accuracy y sin falsos negativos), su desempeño se evaluó en una única partición de los datos. En cambio, el modelo de Random Forest optimizado alcanzó una **precisión promedio de 94.8% en validación cruzada**, lo que indica una mayor capacidad de generalización ante nuevos datos.

Además, mientras la Regresión Logística podría verse afectada por correlaciones entre variables (colinealidad), el Random Forest tolera sin problemas estas relaciones internas y **captura patrones más complejos** gracias a su arquitectura basada en múltiples árboles.

También se destaca que este modelo logró un **AUC-ROC de 0.99** y una **precisión promedio en la curva PR de 0.99**, confirmando su excelente habilidad para separar correctamente las clases.

Por estas razones, el **Random Forest optimizado fue seleccionado como el mejor modelo para el problema**, tanto por robustez técnica como por confiabilidad práctica.

3. Conclusión General del Informe

El análisis exploratorio permitió limpiar y comprender el dataset, identificando variables clave asociadas a conductas relacionadas con el TEA, así como correlaciones significativas entre algunas preguntas del test.

El proceso de modelado mostró que, si bien la Regresión Logística logró métricas muy altas sobre el conjunto de test, el **Random Forest optimizado fue el modelo con mejor capacidad de generalización**, respaldado por validación cruzada y métricas avanzadas como AUC-ROC y curva PR.

Recomendación: Aplicar este modelo en contextos donde se requiera priorizar derivaciones clínicas, como en Tierra del Fuego, teniendo en cuenta que no reemplaza un diagnóstico profesional, pero sí puede apoyar la toma de decisiones tempranas en salud y educación.