Analítica Avanzada de Datos.

Proyecto Final (30%).

Descripción del proyecto:

El objetivo de este proyecto es "predecir los proveedores potencialmente fraudulentos" basándonos en las reclamaciones presentadas por ellos. Además, descubrir variables importantes que ayuden a detectar el comportamiento de los proveedores potencialmente fraudulentos. Por otra parte, analizar patrones fraudulentos en las reclamaciones de los proveedores para comprender su comportamiento futuro.

DataSet. "*Health Insurance Fraud Detection*" el cual contiene: características numéricas y categóricas, valores perdidos y un desbalance de clases.

https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?resource=download&select=Test_Outpatientdata-1542969243754.csv

Descripción DataSet:

Para este proyecto, tenemos en cuenta las reclamaciones de pacientes hospitalizados, las reclamaciones de pacientes ambulatorios y los datos de los beneficiarios de cada proveedor. Veamos sus detalles:

A) **Datos de pacientes hospitalizados:** Estos datos proporcionan información sobre las reclamaciones presentadas por los pacientes ingresados en los hospitales. También proporciona detalles adicionales como las fechas de admisión y alta y el código de diagnóstico de admisión.
B) **Datos de pacientes ambulatorios:** Estos datos proporcionan información detallada sobre las reclamaciones presentadas por pacientes que visitan hospitales y no están ingresados en ellos.
C) **Datos de los beneficiarios:** Estos datos contienen información sobre los beneficiarios, como su estado de salud, la región a la que pertenecen, etc.

Pasos del proyecto:

Exploración y preparación de datos:

- Realizar un análisis exploratorio de los datos para comprender la distribución de las características y las clases de fraude.
- Evaluar y tratar los valores perdidos en el conjunto de datos utilizando técnicas de imputación o eliminación de filas/columnas.
- Evaluar el desbalance de clases y considerar técnicas de muestreo (sobremuestreo, submuestreo, generación de datos sintéticos) si es necesario.

Ingeniería de características:

- Realizar una selección de características relevantes para la detección de fraude en seguros de salud.
- Transformar características categóricas en variables numéricas utilizando técnicas como: codificación one-hot o codificación ordinal.

Construcción y evaluación de modelos:

- Experimentar con diferentes algoritmos de clasificación (uno por cada integrante):
    - K-NN (al menos dos K's)
    - Random Forest
    - Regresión Logística
    - Support Vector Machines (SVM)
    - Redes Neuronales

Validación y presentación de resultados:

- Realizar una evaluación de los modelos utilizando 10 – Fold Cross-Validation Estratificado
- Presentar los resultados del proyecto, y visualizaciones adecuadas (precisión, recall, F1-score, matriz de confusión, gráficos de rendimiento, curvas ROC, etc.).
- Explicar las conclusiones (individuales) obtenidas y discutir posibles mejoras o futuras direcciones.

```
Confusion Matrix Val:
 [[ 60  45]
 [ 51 926]]
Accuracy Val:  0.911275415896488
Sensitivity Val:  0.5714285714285714
Specificity Val:  0.9477993858751279
Kappa Value : 0.50631647988137
AUC         : 0.7596139786518497
F1-Score Val  :  0.5555555555555556
```

Reporte Ejecutivo:

1. Portada: La portada debe incluir el título del informe, el nombre del proyecto, la fecha y el nombres del equipo.
2. Resumen ejecutivo: Un resumen breve pero completo de los principales hallazgos y conclusiones del proyecto de ciencia de datos. Debe proporcionar una visión general de los resultados sin entrar en demasiados detalles.
3. Índice

4. Introducción: Una introducción que describa el contexto y los objetivos del proyecto. Explicar el problema que se abordó, los datos utilizados y los objetivos específicos que se persiguieron.
5. Metodología: Una descripción de los métodos y técnicas utilizados en el proyecto. Esto puede incluir una explicación de los algoritmos y técnicas de análisis de datos empleadas, así como detalles sobre la limpieza y transformación de los datos.
6. Análisis exploratorio de datos: Aquí incluir gráficos, tablas y visualizaciones que resuman las características de los datos y muestren patrones, correlaciones o anomalías relevantes con sus respectivas interpretaciones.
7. Desarrollo del modelo: Una descripción del proceso de desarrollo de los modelos. Esto puede incluir la selección de características, la división de los datos en conjuntos de entrenamiento y prueba, así como los detalles del proceso de entrenamiento y validación del modelo.
8. Resultados: La presentación de los resultados obtenidos a través de los modelos. Esto debe incluir métricas de evaluación del rendimiento de los modelos. Además, incluir visualizaciones o gráficos que ayuden a interpretar los resultados.
9. Discusión: Una discusión detallada sobre los resultados y su interpretación.
10. Conclusiones (individuales) y trabajo a futuro: Conclusiones generales del proyecto y las recomendaciones para futuras acciones o investigaciones.
11. Referencias: APA

¿Qué van a entregar?

1. Reporte escrito
2. Explicación del código desarrollado

Fecha límite de entrega: 23 de junio