Proyecto II

1. Healthcare dataset (healthcare_dataset.csv)

Contexto:

Este conjunto de datos sintéticos de salud se creó como un recurso valioso para los entusiastas de la ciencia de datos, el aprendizaje automático y el análisis de datos. Está diseñado para imitar datos de salud del mundo real, lo que permite a los usuarios practicar, desarrollar y demostrar sus habilidades de manipulación y análisis de datos en el contexto del sector sanitario.

Inspiración:

La inspiración detrás de este conjunto de datos radica en la necesidad de datos de salud prácticos y diversos para fines educativos y de investigación. Los datos de salud suelen ser sensibles y estar sujetos a regulaciones de privacidad, lo que dificulta su acceso para el aprendizaje y la experimentación. Para abordar esta deficiencia, se aprovechó la biblioteca Faker de Python para generar un conjunto de datos que refleja la estructura y los atributos que se encuentran comúnmente en los historiales médicos. Al proporcionar estos datos sintéticos, se espera fomentar la innovación, el aprendizaje y el intercambio de conocimientos en el ámbito del análisis de la salud.

Información del conjunto de datos:

Cada columna proporciona información específica sobre el paciente, su ingreso y los servicios de salud prestados, lo que hace que este conjunto de datos sea adecuado para diversas tareas de análisis y modelado de datos en el ámbito sanitario. A continuación, se presenta una breve explicación de cada columna del conjunto de datos:

- **Name**: Esta columna representa el nombre del paciente asociado con el historial médico.
- **Age**: La edad del paciente al momento del ingreso, expresada en años.
- **Gender**: Indica el género del paciente, ya sea "Masculino" o "Femenino".
- **Blood Type**: El grupo sanguíneo del paciente, que puede ser uno de los más comunes (p. ej., "A+", "O-", etc.).
- **Medical Condition**: Esta columna especifica la afección médica o diagnóstico principal asociado con el paciente, como "Diabetes", "Hipertensión", "Asma", etc.

- **Date of Admission**: La fecha en que el paciente ingresó al centro de salud.
- **Doctor**: El nombre del médico responsable de la atención del paciente durante su ingreso.
- **Hospital**: Identifica el centro de salud u hospital donde ingresó el paciente. Proveedor de Seguros: Esta columna indica el proveedor de seguros del paciente, que puede ser uno de varios, como "Aetna", "Blue Cross", "Cigna", "UnitedHealthcare" y "Medicare".
- **Billing Amount**: El monto facturado por los servicios de atención médica del paciente durante su ingreso. Se expresa como un número de punto flotante.
- **Room Number**: El número de habitación donde se alojó al paciente durante su ingreso.
- **Admission Type**: Especifica el tipo de ingreso, que puede ser "Emergencia", "Electiva" o "Urgente", según las circunstancias del ingreso.
- **Discharge Date**: La fecha en que el paciente fue dado de alta del centro de atención médica, según la fecha de ingreso y un número aleatorio de días dentro de un rango realista.
- **Medication**: Identifica un medicamento recetado o administrado al paciente durante su ingreso. Algunos ejemplos incluyen "Aspirina", "Ibuprofeno", "Penicilina", "Paracetamol" y "Lipitor".
- **Test Results**: Describe los resultados de una prueba médica realizada durante el ingreso del paciente. Los valores posibles incluyen "Normal", "Anormal" o "No concluyente", lo que indica el resultado de la prueba.

2. Diabetes prediction dataset (diabetes_prediction_dataset.csv)

El conjunto de datos de predicción de diabetes recopila datos médicos y demográficos de pacientes, junto con su estado de diabetes (sea positivo o negativo). Los datos incluyen características como edad, sexo, índice de masa corporal (BMI), hipertensión, cardiopatías, antecedentes de tabaquismo, nivel de HbA1c y glucemia. Este conjunto de datos permite crear modelos de aprendizaje automático para predecir la diabetes en pacientes basándose en su historial médico e información demográfica. Esto resulta útil para los profesionales sanitarios a la hora de identificar pacientes con riesgo de desarrollar diabetes y desarrollar planes de tratamiento personalizados. Además, los investigadores pueden utilizar el conjunto de datos para explorar la relación entre diversos factores médicos y demográficos y la probabilidad de desarrollar diabetes.

3. Loan approval dataset (loan_data.csv)

Este conjunto de datos es una versión sintética inspirada en el conjunto de datos original de Riesgo Crediticio de Kaggle y enriquecida con variables adicionales basadas en datos de

Riesgo Financiero para la Aprobación de Préstamos. Se utilizó SMOTENC para simular nuevos puntos de datos y ampliar las instancias. El conjunto de datos está estructurado para características categóricas y continuas.

Column	Description	Type
`person_age`	Age of the person	Float
`person_gender`	Gender of the person	Categorical
`person_education`	Highest education level	Categorical
`person_income`	Annual income	Float
`person_emp_exp`	Years of employment experience	Integer
`person_home_ownership`	Home ownership status (e.g., rent, own, mortgage)	Categorical
`loan_amnt`	Loan amount requested	Float

`loan_intent`	Purpose of the loan	Categorical
`loan_int_rate`	Loan interest rate	Float
`loan_percent_income`	Loan amount as a percentage of annual income	Float
`cb_person_cred_hist_length`	Length of credit history in years	Float
`credit_score`	Credit score of the person	Integer
`previous_loan_defaults_on_file`	Indicator of previous loan defaults	Categorical
`loan_status` (target variable)	Loan approval status: 1 = approved; 0 = rejected	Integer