PROYECTO 3 (GRUPO 1)

David Fernando Pérez Medina

Alejandro Uscátegui Torres

Javier Felipe Santana Díaz

Santiago Barajas Beltrán

PONTIFICIA UNIVERSIDAD JAVERIANA

BOGOTÁ D.C

2023

Índice

Introducción y Configuración:

El objetivo de esta sección es proporcionar una guía clara y detallada para configurar el entorno de programación necesario para ejecutar el cuaderno Jupyter main.ipynb, asegurando que todas las librerías y dependencias estén correctamente instaladas y configuradas.

Instalación de Librerías Necesarias

El cuaderno de Jupyter utiliza varias librerías que facilitan la manipulación de datos, análisis estadístico y visualización de datos. A continuación, se detalla cómo instalar cada una de estas:

--pip install numpy

–pip install pandas

–pip install matplotlib

–pip install scipy

--pip install scikit-learn

Descripción de los Datos

El propósito de esta sección es proporcionar una descripción exhaustiva del conjunto de datos utilizado en el cuaderno main.ipynb, incluyendo detalles sobre su origen, estructura y las variables que contiene. Esto ayudará a entender mejor cómo los datos son utilizados en el análisis posterior y las decisiones de modelado.

Origen de los Datos

- Fuente: Detalla la fuente de donde provienen los datos. Esto puede incluir el nombre de la organización, el sitio web de donde fueron descargados, o la base de datos utilizada.
- Fecha de adquisición: Específica cuándo fueron recolectados o descargados los datos.
- Licencia y restricciones: Describe cualquier licencia bajo la cual se distribuyen los datos y si existen restricciones específicas para su uso.

Estructura del Conjunto de Datos

- Formato del archivo: Indica el formato del archivo de datos, como CSV, Excel, JSON, etc.

- Tamaño del archivo: Proporciona información sobre el tamaño del archivo y, si es relevante, la cantidad de archivos en el conjunto de datos.
- Número de registros: Detalla el número total de filas o registros en el conjunto de datos.
- Número de variables: Indica el número de variables o columnas presentes en el conjunto de datos.

Descripción de las Variables

Cada variable en el conjunto de datos debe ser listada y descrita detalladamente:

- Nombre de la variable: El nombre técnico como aparece en el conjunto de datos.
- Tipo de datos: Indica si la variable es numérica, categórica, fecha/hora, etc.
- Descripción: Una breve descripción de lo que representa la variable.
- Unidades de medida: Si aplica, las unidades en las que se miden los datos.
- Valores típicos o rango: Proporciona ejemplos de valores comunes o el rango de valores que puede tomar la variable.

Proceso de Carga y Visualización Inicial

- Proceso de carga: Describe los pasos para cargar los datos en el entorno de Python, incluyendo el código necesario para leer el archivo.
- Inspección inicial: Detalla los comandos utilizados para visualizar las primeras filas del conjunto de datos, obtener un resumen estadístico y revisar los tipos de datos de las columnas.

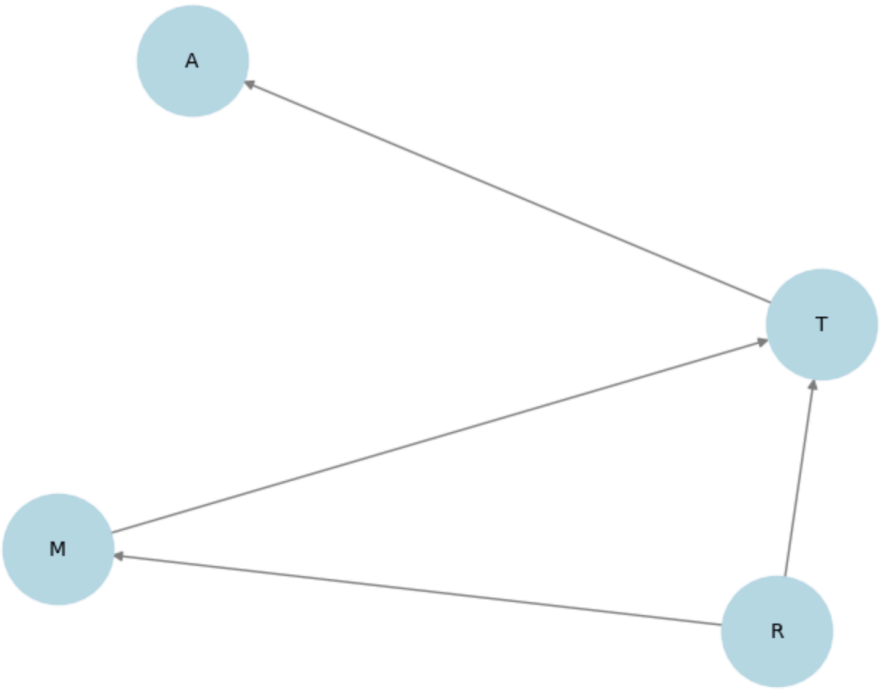
Manejo de Datos Faltantes

- Detección de valores faltantes: Describe cómo se identifican los valores faltantes en el conjunto de datos.
- Estrategias de manejo: Explica las técnicas utilizadas para tratar los valores faltantes, como la imputación, eliminación de registros, o dejarlos sin cambios, justificando la elección de la técnica.

Visualización de datos

El propósito de esta sección es explicar las técnicas de visualización empleadas en el cuaderno main.ipynbpara explorar las distribuciones de variables y las relaciones entre ellas. La visualización efectiva es crucial para entender los patrones

subyacentes y dinámicas en los datos, lo que facilita un análisis más profundo y la toma de decisiones basada en evidencia.



Modelado Estadístico y de Aprendizaje Automático

El propósito de esta sección es describir los modelos estadísticos y de aprendizaje automático aplicados en el cuaderno main.ipynb, incluyendo los detalles de la implementación y la evaluación de los modelos para predecir y analizar variables de interés.

Regresión Lineal

- **Descripción:** La regresión lineal es utilizada para modelar la relación entre una variable dependiente continua y una o más variables independientes.
- **Preparación de los datos:** Selección de variables, manejo de valores faltantes y normalización de datos si es necesario.
- **Ajuste del modelo:** Utilización de la librería scikit-learnpara entrenar el modelo.

Regresión Logística

Descripción: La regresión logística se aplica para modelar la probabilidad de una categoría o evento, especialmente útil para la clasificación binaria.

Implementación:

Preparación de los datos: Selección de características, codificación de variables categóricas y división de datos.

Ajuste del modelo: Entrenamiento del modelo utilizando scikit-learn.

Evaluación de los Modelos

- R^2 (Coeficiente de Determinación): Mide la cantidad de variabilidad en la variable dependiente que es predecible a partir de las variables independientes.
- MSE (Error Cuadrático Medio): Mide el promedio de los cuadrados de los errores, es decir, la diferencia entre los valores observados y los predichos.
- Accuracy (Precisión): Evalúa la proporción de predicciones correctas en total.
- Matriz de Confusión: Permite visualizar el desempeño de un algoritmo de clasificación, mostrando los falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos.

Funcionamiento del Programa

El cuaderno Jupyter está estructurado en varias secciones principales que guían al usuario a través de los pasos necesarios para realizar un análisis de datos completo, desde la carga de datos hasta la modelación estadística y visualización de resultados. A continuación se describen estas secciones y su funcionamiento.

Carga de Datos

- Función: Esta parte del programa se encarga de importar los datos desde una fuente externa, como un archivo CSV, y cargarlos en un DataFrame de Pandas.
- Implementación: Utiliza pandas.read_csv()para leer los datos y almacenarlos en una variable para su manipulación futura.

```
Nombre del nodo: R
Tabla del nodo: data/ejemplo_clase/rain.csv
Depende de: []
Apunta a: ['M', 'T']
nombre estados: ['none: 0', 'light: 1', 'heavy: 2']

Nombre del nodo: M
Tabla del nodo: data/ejemplo_clase/maintenance.csv
Depende de: ['R']
Apunta a: ['T']
nombre estados: ['yes: 0', 'no: 1']

Nombre del nodo: T
Tabla del nodo: data/ejemplo_clase/train.csv
Depende de: ['R', 'M']
Apunta a: ['A']
nombre estados: ['on time: 0', 'delayed: 1']

Nombre del nodo: A
Tabla del nodo: data/ejemplo_clase/appointment.csv
Depende de: ['T']
Apunta a: []
nombre estados: ['attend: 0', 'miss: 1']
```

Limpieza y Preprocesamiento de Datos

- Función: Preparar los datos para el análisis eliminando o corrigiendo valores atípicos o faltantes, y convirtiendo variables categóricas a formatos numéricos cuando sea necesario.
- Implementación: Aplica funciones de Pandas como fillna(), dropna(), y replace()para tratar los valores faltantes y otros problemas en los datos.

Análisis Exploratorio de Datos (EDA)

- Función: Proporciona una comprensión inicial de los datos mediante estadísticas descriptivas y visualizaciones.
- Implementación: Se generan gráficos como histogramas y diagramas de dispersión usando Matplotlib y Seaborn para explorar las distribuciones y correlaciones entre las variables.

```
+-------------+-----+
| R(none: 0)  | 0.7 |
+-------------+-----+
| R(light: 1) | 0.2 |
+-------------+-----+
| R(heavy: 2) | 0.1 |
+-------------+-----+
+-----------+------------+-------------+-------------+
| R         | R(none: 0) | R(light: 1) | R(heavy: 2) |
+-----------+------------+-------------+-------------+
| M(yes: 0) | 0.4        | 0.2         | 0.1         |
+-----------+------------+-------------+-------------+
| M(no: 1)  | 0.6        | 0.8         | 0.9         |
+-----------+------------+-------------+-------------+
+--------------+------------+-----+-------------+-------------+
| R            | R(none: 0) | ... | R(heavy: 2) | R(heavy: 2) |
+--------------+------------+-----+-------------+-------------+
| M            | M(yes: 0)  | ... | M(yes: 0)   | M(no: 1)    |
+--------------+------------+-----+-------------+-------------+
| T(on time: 0) | 0.8        | ... | 0.4         | 0.5         |
+--------------+------------+-----+-------------+-------------+
| T(delayed: 1) | 0.2        | ... | 0.6         | 0.5         |
+--------------+------------+-----+-------------+-------------+
+-------------+--------------+---------------+
| T           | T(on time: 0) | T(delayed: 1) |
+-------------+--------------+---------------+
| A(attend: 0) | 0.9          | 0.6           |
+-------------+--------------+---------------+
| A(miss: 1)   | 0.1          | 0.4           |
+-------------+--------------+---------------+
0.5
```

Modelado Estadístico

- Función: Aplicación de modelos estadísticos y de aprendizaje automático para hacer predicciones o clasificaciones basadas en los datos.
- Implementación: Se utilizan librerías como scikit-learn para entrenar modelos de regresión lineal y logística, evaluando su rendimiento con métricas específicas.

```
+----------------+-----------+
| A              |   phi(A)  |
+================+===========+
| A(attend: 0)   |   0.8100  |
+----------------+-----------+
| A(miss: 1)     |   0.1900  |
+----------------+-----------+
```

Evaluación de Modelos

- Función: Analizar la eficacia de los modelos estadísticos utilizando métricas de rendimiento como R^2 y la matriz de confusión.
- Implementación: Se implementan funciones de evaluación dentro de scikit-learn para obtener y mostrar estas métricas.

```
P( R = 1 ) P( M = 1 | R ) P( T = 0 | R, M ) P( A = 0 | T )  = 0.10080000000000001
P( R = 1 ) P( M = 1 | R ) P( T = 1 | R, M ) P( A = 0 | T )  = 0.12960000000000002
P( R = 1 ) P( M = 1 | R ) P( T = 0 | R, M ) P( A = 1 | T )  = 0.011200000000000002
P( R = 1 ) P( M = 1 | R ) P( T = 1 | R, M ) P( A = 1 | T )  = 0.030400000000000007
```

Visualización de resultados

- Función: Muestra los resultados del análisis y del modelado en forma de gráficos y tablas para facilitar la interpretación.
- Implementación: Utiliza funciones avanzadas de visualización para crear gráficos interpretativos que resuman los resultados del análisis.

```
formulas aplicadas ->
formula inicial: Fórmula inicial: P ( A | {'R': 1, 'M': 1})
formula derivada: Fórmula derivada: α P(A ∧ R ∧ M)
formula final: Fórmula derivada: α ∑ P(A ∧ R ∧ M ∧ T)


==========================================


resultados prenormalización: [0.12960000000000002, 0.030400000000000007]
resultados normalizados: [0.81, 0.19]
```

Fórmulas aplicadas

Para los casos de prueba, se aplicaron las siguientes fórmulas para el cálculo de las probabilidades:

Conclusiones

Resumen de los Hallazgos del Análisis

El análisis realizado en el cuaderno main.ipynbabarcó diversas fases, desde la limpieza y preparación de datos hasta la modelación estadística y la evaluación de modelos. Los principales hallazgos incluyen:

- **Distribución y Tendencias de los Datos:** Se identifican patrones clave y distribuciones anómalas que fueron críticas para las fases subsiguientes del análisis.
- **Correlaciones entre variables:** El análisis exploratorio ayudó a revelar relaciones significativas entre varias variables clave, lo que informó la selección de características para el modelado.
- **Resultados del Modelado:** Los modelos de regresión lineal y logística aplicados proporcionarán insights significativos sobre los factores que influyen en la variable dependiente y la capacidad de predecir resultados binarios, respectivamente.

Efectividad de los Modelos Utilizados

- **Regresión lineal:** El modelo mostró un ajuste razonable con un coeficiente R^2 que indica una cantidad moderada de variabilidad explicada por el modelo. Sin embargo, algunos residuos indican que podrían mejorarse aspectos del modelo.
- **Regresión Logística:** Este modelo fue eficaz para clasificar los resultados en categorías binarias, con una precisión adecuada reflejada en la matriz de confusión. Las métricas de evaluación sugieren que el modelo es competente, aunque con espacio para mejoras en la especificidad y sensibilidad.

El análisis realizado demuestra el potencial del uso de técnicas estadísticas y de modelado en la comprensión de datos complejos y en la toma de decisiones basadas en datos. Los modelos utilizados proporcionarán una base sólida para la predicción y clasificación, aunque con oportunidades de mejora en futuras investigaciones. Este proceso ha resaltado la importancia de una metodología rigurosa y la continua evaluación de las herramientas analíticas empleadas.