Práctica 2: Limpieza y validación de los datos

Integrantes:

- Diana Sánchez.
- Carlos Simbaña.

1. Detalles de la actividad

El presente documento muestra un caso práctico orientado a identificar, limpiar, integrar, preparar, validar y analizar los datos dentro de un proyecto analítico en el cual se utiliza herramientas de código abierto para la consecución de los objetivos.

1.1. Descripción

Para la realización de la práctica, se han elegido los datos de la competencia “Titanic ML” la cual originalmente busca la realización de un modelo capaz de predecir los pasajeros que sobrevivieron al naufragio del 15 de abril de 1912. En esta actividad se realizará la limpieza de los datos para corregir valores erróneos que pudieren encontrarse; la validación de los mismos y el análisis descriptivo para proveer información de valor en relación a las probabilidades de supervivencia de los pasajeros del Titanic.

1.2. Objetivos

Los objetivos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y la capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3. Competencias

En esta práctica se desarrollan las siguientes competencias:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.

- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Resolución

2.1. Descripción del dataset

El dataset contiene dos archivos referentes al naufragio del Titanic, train.csv (891 registros) test.csv (418 registros) y correspondientes a la competencia “Titanic ML ” disponible en el enlace https://www.kaggle.com/c/titanic.

Los archivos de datos se encuentran constituidos de la siguiente manera:

- **PassengerId:** Es el identificador del pasajero.
- **Survived:** Indica si el pasajero sobrevivió o no. Este campo solo se encuentra en el fichero train.csv debido a que este se utilizará como archivo de entrenamiento para predecir la supervivencia de cada uno de los pasajeros del archivo test.csv. Cero (0) es igual a “No sobrevive”. Uno (1) es igual a “Sobrevive”.
- **Pclass:** Es el tipo de ticket que adquirió el pasajero. 1 es Upper, 2 es Middle y 3 es Lower.
- **Name:** Es el nombre del pasajero.
- **Sex:** Es el sexo del pasajero.
- **Age:** Es la edad del pasajero expresada en años.
- **SibSp:** Es el número de hermanos y/o esposas(os) a bordo, por pasajero.
- **Parch:** Es el número de padres y/o hijos a bordo, por pasajero.
- **Ticket:** Es el número de ticket del pasajero.
- **Fare:** Es la tarifa pagada por el pasajero.
- **Cabin:** Es el número de camarote del pasajero.
- **Embarked:** Es el Puerto de embarque del pasajero. C = Cherbourg, Q = Queenstown, S = Southampton.

2.2. Importancia y objetivos de los análisis

La importancia de este conjunto de datos radica en encontrar la relación entre las diferentes variables las cuales permiten que aumenten o disminuyan la probabilidad de los pasajeros de sobrevivir al naufragio. Para esto es importante realizar la creación y cambios que fueren necesarios en las variables utilizando métodos como la normalización, discretización y transformaciones necesarias para posteriormente llevar a cabo el análisis de los datos. Debido a que en total los datos de los dos archivos suman 1309 registros y el objetivo del análisis es el contar con la probabilidad individual de cada pasajero para sobrevivir, no se utilizará el método de reducción de la cantidad.

La realización de este análisis permitirá que se realicen otros modelos predictivos aplicados a otros naufragios con la finalidad que puedan establecerse nuevas normativas y planes prevención en temas de navegación fluvial en el mundo. No obstante, también sirve de motivación para que aplicando técnicas similares, se realicen modelos enfocados en la navegación terrestre, aérea y ferroviaria.

2.3. Limpieza de los datos

Previa a la ejecución de la limpieza de los datos, se requiere realizar un análisis visual que permita determinar la ruta a seguir para obtener la mejor calidad de datos posibles. Para esto se han realizado las siguientes observaciones:

<table>
<tr><th>Campo</th><th>Tipo de variable</th><th>Observaciones</th></tr>
<tr><td>PassengerId</td><td>Descriptiva</td><td>Es un número secuencial de cada pasajero. En el archivo train.csv inicia en 1 y finaliza en 891. En el archivo test.csv inicia en 892 y finaliza en 1309.</td></tr>
<tr><td>Survived</td><td>Categórica</td><td>No existen valores en blanco.</td></tr>
<tr><td>Pclass</td><td>Categórica</td><td>No existen valores en blanco.</td></tr>
<tr><td>Name</td><td>Cualitativa</td><td>Contiene las abreviaturas Mr. (hombres adultos), Mrs. (mujeres adultas casadas), Miss. (mujeres solteras), Master. (hombres jóvenes); datos que pueden ser de utilidad para determinar la edad en caso de no encontrarse en los archivos iniciales.</td></tr>
<tr><td>Sex</td><td>Categórica</td><td>No existen valores en blanco.</td></tr>
<tr><td>Age</td><td>Cuantitativa</td><td>Existen las siguientes cantidades:
<table>
<tr><th>Detalle</th><th>train.csv</th><th>test.csv</th></tr>
<tr><td>Valores menores a uno</td><td>7</td><td>5</td></tr>
<tr><td>Valores con decimales xx,5</td><td>18</td><td>15</td></tr>
<tr><td>Campos en blanco</td><td>177</td><td>86</td></tr>
</table>
</td></tr>
<tr><td>SibSp</td><td>Cuantitativa</td><td>No existen valores en blanco.</td></tr>
<tr><td>Parch</td><td>Cuantitativa</td><td>No existen valores en blanco.</td></tr>
</table>

Ticket	Categórica	No existen valores en blanco.
Fare	Cuantitativa	No existen valores en blanco.
Cabin	Categórica	Existen valores en blanco: test.csv (327), train.csv (687)
Embarked	Categórica	Existen valores en blanco: train.csv (2)

2.3.1. Selección de los datos de interés

Los atributos presentes en el análisis, contienen información de cada pasajero del Titanic en los cuales de pueden encontrar características físicas, emocionales y restricciones que pudieron afectar a cada pasajero al momento del naufragio y por ende afectaron en su supervivencia. Por ejemplo en el caso de tener niños, seguramente los padres, trataron de salvarlos primero incluso poniendo en riesgo su propia supervivencia.

De las columnas provistas en los conjuntos de datos, se puede obviar: XX, XXX y XXXX en vista que son variables que no tienen relación con la probabilidad de supervivencia.

2.3.2. Ceros y elementos vacíos

2.3.3. Valores extremos

2.3.4. Exportación de los datos preprocesados

2.4. Análisis de los datos

2.4.1. Selección de los grupos de datos a analizar
2.4.2. Comprobación de la normalidad y homogeneidad de la varianza
2.5. Pruebas estadísticas
2.5.1. ¿Qué variables cuantitativas influyen más en la probabilidad de supervivencia al naufragio?
2.5.2. ¿Es mayor la probabilidad de supervivencia para las mujeres y niños?
2.5.3. ¿Cuáles son las variables dependientes en el análisis?
2.5.4. Modelo de regresión lineal
2.6. Conclusiones

3. Recursos