¿Cómo realizar la limpieza y análisis de datos?

Práctica 2

William Gabriel Granda Betancourt

Oscar Augusto Diaz Triana

Universitat Oberta de Catalunya

Índice

Descripción Práctica 2

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Competencias

En esta PEC se desarrollan las siguientes competencias del Máster de Data Science:

- 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.

Objetivos

Los objetivos concretos de la práctica son:

- Aprender a aplicar los conocimientos adquiridos y su 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

Desarrollo de la Práctica

1. Descripción del dataset

El conjunto de datos sobre pacientes y las posibilidades de sufrir un infarto es de gran importancia debido a su relevancia para la salud cardiovascular. Al analizar este conjunto de datos, el objetivo es identificar las características o factores que influyen en la propensión de una persona a sufrir un ataque al corazón.(¿Cuáles son las características que influyen en la propensión de una persona a sufrir un ataque cardíaco?)

Mediante la implementación de un modelo de clasificación, es posible predecir con antelación si una persona es propensa a sufrir un ataque cardíaco. Esto permitiría a los especialistas médicos intervenir de manera oportuna y tomar medidas preventivas para reducir el riesgo. Al identificar las características que están fuertemente asociadas con los ataques cardíacos, se podrían desarrollar estrategias y programas de prevención más efectivos para mejorar la salud cardiovascular de la población estudiada.

2. Integración y selección

La integración y selección de datos permite utilizar información relevante y de calidad para un análisis preciso y significativo, mejorando la toma de decisiones y obteniendo resultados confiables.

Para realizar la integración y selección de datos en R con el dataset "Heart Attack Analysis & Prediction", se comienza descargando el dataset desde el enlace proporcionado. Luego, se cargan las bibliotecas necesarias, en este caso, el paquete "tidyverse", que permitirá trabajar de manera eficiente con los datos.

Después de cargar el dataset, se explora los datos para familiarizarse con su contenido. Se visualizan las primeras filas, se obtiene información sobre las variables y se calculan estadísticas descriptivas para tener una idea general de los datos.

Luego, se seleccionan las variables relevantes para el análisis de los ataques cardíacos. Basándose en el conocimiento previo y el análisis exploratorio, se identifican las características que podrían estar relacionadas con los ataques cardíacos. Algunas variables de interés podrían ser la edad, el sexo, los niveles de colesterol y la presión arterial.

Después de seleccionar las variables, se verifica si hay valores faltantes en el dataset seleccionado y se decide cómo manejarlos. Se utiliza la función "complete.cases" para eliminar las filas que contengan valores faltantes. Esto asegurará que los datos estén completos y listos para el análisis.

Es importante documentar cada paso realizado durante este proceso y prestar atención a cualquier necesidad adicional de preprocesa

3. Limpieza de los datos.

La limpieza de datos es crucial para eliminar errores, valores faltantes y ruido, lo que mejora la calidad de los datos y evita sesgos en el análisis, proporcionando resultados más precisos y confiables.

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Es recomendable realizar una verificación exhaustiva para asegurarse de que no haya ningún valor cero o elemento vacío oculto en los datos. Puedes utilizar funciones como is.na() en R.

Al verificar los datos no se encuentran ceros ni elementos vacíos, esto indica que el dataset está completo y no requiere ninguna acción adicional en ese aspecto. Es positivo que los datos no contengan valores faltantes, ya que esto facilita el análisis y evita posibles sesgos.

Igualmente se verifica si hay valores duplicados, con la función duplicated(), lo cual es importante gestionarlos para evitar cualquier distorsión o sesgo en el análisis. Al realizar la se encontraron 2 valores duplicados. Como los valores duplicados no aportan información adicional y no son necesarios para el análisis, se elimina para obtener un dataset único y sin duplicados, quedando un dataset de 302 registros.

3.2. Identifica y gestiona los valores extremos.

Al identificar valores extremos en el dataset, es importante gestionarlos adecuadamente para evitar que afecten el análisis. En este caso, se han identificado valores atípicos en la variable "caa" con los valores 3 y 4. Sin embargo, la definición de esta variable establece que su dominio es de 0 a 3, por lo que el único valor atípico es 4. Para tratar este valor atípico, se propone reemplazarlo por el valor máximo permitido en el dominio, que es 3.

En cuanto a las variables "chol", "oldpeak" y "trtbps", se ha determinado que existen valores atípicos. En lugar de utilizar la media para tratar estos valores atípicos, se utilizará la mediana de cada una de las variables. Esto se debe a que la media puede verse sesgada por la presencia de valores atípicos, mientras que la mediana es una medida más robusta ante la presencia de valores extremos.

4. Análisis de los datos.

El análisis de datos en el dataset que estamos trabajando es importante porque nos permite obtener información crucial sobre los factores que pueden influir en los ataques cardíacos. Al explorar y examinar los datos, podemos identificar patrones, relaciones y características relevantes que ayudarán a comprender mejor los riesgos y tomar medidas preventivas adecuadas para la salud cardiovascular de los pacientes, lo que potencialmente puede salvar vidas y mejorar la calidad de vida de las personas.

4.1. Selección de los grupos de datos

En este paso del análisis, se seleccionan los grupos de datos que se desean analizar y comparar, en este caso, en función de la variable "output". Se aplicará un análisis de variables numéricas, específicamente se examinará el comportamiento de las variables "age", "trtbps", "thalachh" y "oldpeak" en relación con la variable "output". Se utilizará la técnica de diagramas de cajas para visualizar las diferencias entre los grupos. Se observa que hay diferencias significativas en estas variables con respecto a "output". Para confirmar estas hipótesis, se realizarán pruebas estadísticas, incluyendo la transformación de Box-Cox en las variables no normalmente distribuidas.

Además, se realizará un análisis de las variables categóricas "sex", "cp", "thall", "restecg", "exng", "slp" y "thall" para determinar si existen diferencias significativas entre los grupos definidos por "output". Se presentarán diagramas de frecuencias para visualizar el comportamiento de estas variables en relación con "output" y se realizarán pruebas estadísticas, como el test de chi-cuadrado y el análisis de frecuencias relativas, para identificar posibles asociaciones o diferencias entre los grupos.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Para comprobar la normalidad y homogeneidad de la varianza en el análisis de datos, se aplican 2 pruebas estadísticas específicas.

Se utilizará la prueba de normalidad, como la prueba de Shapiro-Wilk o la prueba de Kolmogorov-Smirnov, para verificar si los datos siguen una distribución normal. Si los datos no siguen una distribución normal, se pueden aplicar técnicas de transformación, como la transformación de Box-Cox, para lograr una distribución más cercana a la normalidad.

Para verificar la homogeneidad de la varianza entre grupos, se utiliza la prueba de Levene o la prueba de Bartlett. Estas pruebas evalúan si las varianzas de los grupos son iguales. En caso de encontrar diferencias significativas en las varianzas, se pueden aplicar técnicas de análisis que tengan en cuenta la heterogeneidad de la varianza, como el análisis de varianza robusto o los modelos de efectos mixtos.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

Pruebas Estadísticas

En variables numéricas

Se ha realizado un estudio gráfico que se basa en la representación de diagramas de caja (boxplots) para visualizar la distribución de las variables numéricas en relación a la variable "output". Este tipo de gráfico es útil para comparar las medianas, cuartiles y la variabilidad de los datos entre diferentes grupos. En cuanto a las pruebas estadísticas aplicadas, se utilizaron principalmente dos pruebas.La prueba de Shapiro-Wilk, se utilizó para evaluar la normalidad de los datos. Se aplicó a las variables "ageT", "trtbpsT" (transformadas mediante la transformación de Box-Cox), "thalachhT" (transformada) y "oldpeakT" (transformada). La prueba de Shapiro-Wilk determina si una muestra sigue una distribución normal. Si el valor de p es menor que el nivel de significancia (α), se rechaza la hipótesis nula de normalidad. La prueba de Wilcoxon (prueba de rangos con signo): Se aplicó esta prueba no paramétrica a las variables "age", "chol", "trtbps" y "oldpeak". La prueba de Wilcoxon se utiliza para comparar las medianas de dos grupos independientes cuando los datos no siguen una distribución normal. Proporciona una medida de la diferencia estadística entre los grupos.

En base a los resultados obtenidos, se puede concluir lo siguiente:

1. Variable "age":
Hay diferencias estadísticamente significativas en la edad de los pacientes con respecto a la variable "output". Los pacientes más propensos a sufrir un ataque cardíaco tienen una edad promedio diferente a los menos propensos.

2. Variable "chol":
Se observan diferencias estadísticamente significativas en los niveles de colesterol entre las personas menos propensas y más propensas a sufrir un ataque cardíaco.

3. Variable "trtbps":
No se encontraron diferencias estadísticamente significativas en la presión arterial entre los pacientes más propensos y menos propensos a sufrir ataques cardíacos.

4. Variable "thalachh":
Se observan diferencias estadísticamente significativas en la frecuencia cardíaca máxima alcanzada entre las personas menos propensas y más propensas a sufrir un ataque cardíaco. El promedio estimado del grupo de personas menos propensas es menor al del grupo de personas más propensas.

5. Variable "oldpeak":
Se observan diferencias estadísticamente significativas en la depresión del segmento ST de los pacientes más probables y menos probables a sufrir ataques cardíacos.

Las variables "age", "chol", "thalachh" y "oldpeak" presentan diferencias significativas con respecto a la variable "output", lo que indica que estas variables tienen influencia en la presencia de enfermedad cardíaca. Por otro lado, la variable "trtbps" no muestra una diferencia significativa entre los grupos.

En variables categóricas

En función de los datos y los objetivos del estudio, se han aplicado pruebas estadísticas como el test $\chi 2$ para determinar si existen diferencias significativas entre las variables

categóricas y la variable "output" en el estudio. A continuación, se resumen los resultados obtenidos para cada variable categórica en relación con la variable "output":

1. Variable "sex":
Se ha encontrado una asociación significativa entre la variable "sex" (género) y la variable "output" (presencia de enfermedad cardíaca). El 75% de las mujeres son más propensas a sufrir un ataque cardíaco, mientras que en los hombres solo el 49.83% lo son. Se ha aplicado el test $\chi 2$, y el resultado indica que existe una asociación significativa entre las variables sex y output.

2. Variable "cp":
Se ha encontrado una asociación significativa entre la variable "cp" (tipo de dolor torácico) y la variable "output". La tasa de personas sanas es mayor en pacientes asintomáticos (72.72%) en comparación con los pacientes con angina típica (18%), angina atípica (20.69%), y dolor no anginoso (30.43%). La prueba $\chi 2$ ha confirmado la asociación significativa entre las variables cp y output.

3. Variable "thall":
Se ha encontrado una asociación significativa entre la variable "thall" (resultados de la prueba de esfuerzo con talio) y la variable "output". La tasa de salud es menor cuando los pacientes poseen un defecto fijo en la prueba de esfuerzo con talio (22.02%), mientras que es alta cuando los resultados son normales o presentan defectos reversibles (66.67% y 76.07%, respectivamente). La prueba $\chi 2$ confirma la asociación significativa entre las variables thall y output.

4. Variable "restecg":
Se ha encontrado una asociación significativa entre la variable "restecg" (resultados electrocardiográficos en reposo) y la variable "output". Cuando el paciente posee anomalías en la onda ST-T, la tasa de salud es menor en comparación con la tasa de enfermedad. La prueba $\chi 2$ ha confirmado la asociación significativa entre las variables restecg y output.

5. Variable "exng":
Se ha encontrado una asociación significativa entre la variable "exng" (angina producida por ejercicio) y la variable "output". Si el paciente presenta angina, la tasa de salud es mayor en comparación con la tasa de sufrir un ataque cardíaco. La prueba $\chi 2$ ha confirmado la asociación significativa entre las variables exng y output.

6. Variable "slp":
Se ha encontrado una asociación significativa entre la variable "slp" (pendiente de un segmento de electrocardiograma) y la variable "output". Cuando la pendiente es ascendente, la tasa de salud es menor en comparación con la tasa de enfermedad. En cambio, cuando la pendiente es descendente o normal, las tasas de salud son mayores. La prueba $\chi 2$ ha confirmado la asociación significativa entre las variables slp y output.

7. Variable "fbs":
No se ha encontrado una asociación significativa entre la variable "fbs" (nivel de azúcar en sangre en ayunas) y la variable "output". En ambas categorías, la tasa de salud es menor que la tasa de enfermedad. La prueba $\chi 2$ no ha confirmado una asociación significativa entre las variables fbs y output.

Los resultados han mostrado asociaciones significativas entre las variables sex, cp, thall, restecg, exng y slp con la variable "output", lo que indica que estas variables tienen influencia en la presencia de enfermedad cardíaca. Sin embargo, no se ha encontrado una asociación significativa entre la variable fbs y output.

Regresión logística

Se ha realizado un análisis estadístico utilizando la regresión logística para predecir la probabilidad de que una persona sea propensa a sufrir un ataque cardíaco. El valor del estadístico de Kolmogorov-Smirnov obtenido es 0.6956019, y este valor se alcanza en un punto de corte de 0.6077056.

En base a esto, se ha decidido clasificar como pacientes propensos a sufrir un ataque cardíaco aquellos cuya probabilidad estimada sea mayor a 0.6077056. Este punto de corte permite separar eficientemente los casos con mayor riesgo de los casos con menor riesgo.

El modelo utilizado ha demostrado un desempeño satisfactorio, con una exactitud del 83.05%, una precisión del 86.67% y un F1-score del 83.87%. Estos indicadores muestran la capacidad del modelo para predecir correctamente los casos de pacientes propensos a sufrir un ataque cardíaco.

Además, se ha calculado el área bajo la curva (AUC) y se ha obtenido un valor de 0.92. Este valor indica que el modelo tiene una buena capacidad para distinguir entre casos positivos y negativos, lo que refuerza su utilidad en la predicción.

Por último, se ha evaluado la presencia de multicolinealidad utilizando el Factor de Inflación de la Varianza Generalizado (GVIF) para los parámetros estimados. Se ha encontrado que ningún valor del GVIF supera 4, lo que sugiere que no existen problemas significativos de multicolinealidad entre las variables utilizadas en el modelo.,

5. Representación de los resultados a partir de tablas y gráficas.

A lo largo de la práctica, se utilizaron varias representaciones gráficas y tablas para mostrar los resultados. A continuación, se muestran algunos ejemplos de las representaciones utilizadas:

1. Representación gráfica de la distribución de variables numéricas: Se utilizaron gráficos de densidad y diagramas de caja para representar la distribución de las variables numéricas (age, trtbps, chol, thalachh, oldpeak) en relación a la variable "output". Estos gráficos permiten visualizar la forma de la distribución, los valores atípicos y las diferencias entre los grupos.

2. Representación gráfica de variables categóricas: Se utilizaron gráficos de barras para representar la frecuencia de las variables categóricas (sex, cp, fbs, restecg, exng, slp, thall, output). Estos gráficos permiten comparar las proporciones de diferentes categorías y observar posibles patrones o tendencias.

3. Tabla de correlaciones: Se calculó la matriz de correlaciones (método de Spearman) entre las variables numéricas (age, trtbps, chol, thalachh, oldpeak) y se mostró en una tabla. Esta tabla permite identificar las relaciones de correlación entre las variables.

4. Gráfico de cajas para variables numéricas y output: Se utilizó un gráfico de cajas para representar la distribución de las variables numéricas (age, trtbps, chol, thalachh, oldpeak) en relación a la variable "output". Este gráfico muestra las diferencias en las distribuciones entre los grupos de "output".

5. Gráfico ROC (Receiver Operating Characteristic): Se utilizó un gráfico ROC para evaluar el rendimiento del modelo predictivo. Este gráfico muestra la sensibilidad frente a la especificidad del modelo a diferentes puntos de corte y calcula el área bajo la curva (AUC), que indica la capacidad de discriminación del modelo.

6. Resolución del problema.

A partir de los resultados obtenidos en el análisis estadístico y la evaluación del modelo de regresión logística, podemos llegar a las siguientes conclusiones:

- El valor del estadístico de Kolmogorov-Smirnov obtenido, con un valor de 0.6956019, indica que el modelo tiene una buena capacidad para distinguir entre los pacientes propensos y no propensos a sufrir un ataque cardíaco.

- El punto de corte identificado en 0.6077056 nos permite clasificar eficientemente a los pacientes propensos a sufrir un ataque cardíaco. Aquellos con una probabilidad estimada superior a este punto de corte se consideran propensos.

- El modelo de regresión logística ha mostrado un desempeño satisfactorio, con una exactitud del 83.05%, una precisión del 86.67% y un F1-score del 83.87%. Estos indicadores demuestran la capacidad del modelo para predecir correctamente los casos positivos (pacientes propensos).

- El valor del área bajo la curva (AUC) obtenido, con un valor de 0.92, indica que el modelo tiene una buena capacidad de discriminación entre los casos positivos y negativos.

- El análisis del Factor de Inflación de la Varianza Generalizado (GVIF) ha revelado que no existen problemas significativos de multicolinealidad entre las variables utilizadas en el modelo. Esto refuerza la confiabilidad de las estimaciones de los coeficientes de regresión.

7. Código.

Durante el desarrollo del ejercicio, se empleó el lenguaje de programación R, el cual se encuentra alojado en un repositorio de GitHub. Se utilizaron diversas herramientas y técnicas para abordar los siguientes elementos:

Descripción del dataset: Se realizó una descripción detallada de los datos utilizados en el análisis, incluyendo la naturaleza de las variables, la estructura de los datos y la distribución de los valores.

Limpieza de datos: Se realizaron tareas de limpieza y preprocesamiento de los datos para garantizar su calidad y adecuación para el análisis. Esto incluyó el manejo de valores faltantes, la corrección de errores y la transformación de variables si fuera necesario.

Variables numéricas: Se llevaron a cabo técnicas como el imputado de valores faltantes, la detección y tratamiento de outliers, y la normalización o estandarización de variables si fuera requerido.

Variables categóricas: Se realizó el manejo de variables categóricas, incluyendo la codificación de variables categóricas en variables numéricas utilizando técnicas como la codificación one-hot o la codificación ordinal.

Correlaciones: Se exploraron las correlaciones entre las variables para identificar posibles relaciones o dependencias entre ellas.

Comparación entre grupos: Se realizaron comparaciones entre grupos de datos utilizando técnicas estadísticas adecuadas. Esto incluyó la comparación de variables numéricas entre diferentes grupos utilizando pruebas de hipótesis o análisis de varianza, y la comparación de variables categóricas utilizando tablas de contingencia y pruebas de chi-cuadrado.

Variables numéricas: Se compararon las distribuciones de variables numéricas entre diferentes grupos para evaluar posibles diferencias estadísticamente significativas.

Variables categóricas: Se compararon las frecuencias de las categorías de variables categóricas entre diferentes grupos para determinar si existían diferencias significativas.

Regresión logística: Se aplicó el modelo de regresión logística para predecir la probabilidad de que una persona sea propensa a sufrir un ataque cardíaco. Se realizaron técnicas de selección de variables para identificar las variables más relevantes en el modelo.

Conclusiones: A partir de los resultados obtenidos, se elaboraron conclusiones sobre el análisis realizado. Estas conclusiones pueden incluir hallazgos importantes, relaciones identificadas entre variables, el rendimiento del modelo de regresión logística y su capacidad para predecir la probabilidad de ataques cardíacos, entre otros aspectos relevantes.

8. Vídeo.

https://drive.google.com/drive/folders/1nm67WNiMnT7WIBGh5kqJjXXaBpQsYaqV?usp=sharing

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de

- los datos. Editorial UOC.
- Megan Squire (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and
- techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with
- Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science &
- Business Media.
- Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.
- Herramienta para realización de gráficas: https://www.data-to-viz.com/

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los apartados es la siguiente:

Apartado	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Puntos	0.5	0.5	0.2	0.25	1.5	0.5	2	0.5

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

REFERENCIAS

Calvo, M., Subirats, L., & Pérez, D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

Squire, M. (2015). Clean Data. Packt Publishing Ltd.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. Morgan Kaufmann.

Osborne, J. W. (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews, 10(1), 15-27.

Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.

McKinney, W. (2012). Python for Data Analysis. O'Reilly Media, Inc.

GitHub. (n.d.). Tutorial de Github. Recuperado de https://guides.github.com/activities/hello-world

Data to Viz. (n.d.). Herramienta para realización de gráficas. Recuperado de https://www.data-to-viz.com/

Contribuciones	Firma Integrantes
Investigación previa	WGGB, OADT
Redacción de las respuestas	WGGB, OADT
Desarrollo del código	WGGB, OADT
Participación en el vídeo	WGGB, OADT