FACULTAD DE INGENIERÍAS

Ciencia de Datos y Laboratorio – Simulación de Sistemas

Preinforme de Laboratorio #1

"Análisis de Datos en Cáncer de Próstata: Estadísticas y Visualización"

30 de julio de 2024

(Grupos de máximo 2 personas)

Objetivo del Taller:

Desarrollar habilidades fundamentales en álgebra lineal, probabilidad y estadística, y aplicarlas en la ciencia de datos, incluyendo la exploración y visualización de datos, análisis estadístico y pruebas de hipótesis.

El archivo prostate.csv contiene el conjunto de datos del estudio sobre cáncer de próstata, comúnmente utilizado en análisis estadísticos y aprendizaje automático. Este conjunto de datos incluye información como el nivel de antígeno prostático específico (PSA), volumen de próstata, edad del paciente, entre otros.

Estos datos provienen de un estudio que examinó la correlación entre el nivel de antígeno prostático específico y varias medidas clínicas en hombres que estaban a punto de someterse a una prostatectomía radical. Es un marco de datos con 97 filas y 9 columnas.

El conjunto de datos prostate contiene las siguientes columnas, cada una representando una variable relacionada con el cáncer de próstata:

1. **lcavol**: Logaritmo del volumen de la cápsula de la próstata.
2. **lweight**: Logaritmo del peso de la próstata.
3. **age**: Edad del paciente.
4. **lbph**: Logaritmo del volumen de la hiperplasia prostática benigna.
5. **svi**: Invasión de las vesículas seminales (0 = No, 1 = Sí).
6. **lcp**: Logaritmo del peso de la cápsula prostática.
7. **gleason**: Puntuación de Gleason, una medida de la agresividad del cáncer (un valor más alto indica un cáncer más agresivo).
8. **pgg45**: Porcentaje de las células que son de grado 4 o 5 en la puntuación de Gleason.
9. **lpsa**: Logaritmo del nivel de antígeno prostático específico (PSA), que es un marcador utilizado para detectar el cáncer de próstata.

Con base en la información anterior, desarrollar un script en Python que aborde los siguientes análisis:

1. Cálculo y Visualización de Estadísticas Descriptivas

Objetivo: Calcular estadísticas descriptivas y visualizar la distribución de las variables en el conjunto de datos prostate.

- Utilizar pandas para calcular la media, mediana, moda, desviación estándar, y percentiles para variables como age, lcavol, lpsa, etc.
- Crear histogramas para cada variable utilizando matplotlib o seaborn para visualizar la distribución de los datos.
- Generar box plots para identificar la presencia de valores atípicos en variables como lweight y lbph.

Responde: ¿Qué conclusiones puedes sacar sobre la distribución de las variables, como la edad de los pacientes o los niveles de lpsa? ¿Hay valores atípicos que deban investigarse más a fondo?

2. Análisis de Correlación y Regresión Simple

Objetivo: Determinar la relación entre variables y utilizar modelos de regresión para predicción.

- Calcular la matriz de correlación utilizando pandas para identificar relaciones fuertes entre variables, enfocándose en variables como lcavol, lpsa, y lweight.
- Ajustar un modelo de regresión lineal simple utilizando scikit-learn o statsmodels para predecir lpsa basándose en lcavol.
- Evaluar el modelo de regresión con métricas como el R-cuadrado y los errores cuadráticos medios (MSE).

Responde: ¿Qué tan fuerte es la relación entre lcavol y lpsa? ¿Qué implica el valor del R-cuadrado en el contexto de este estudio?

3. Visualización de Relaciones entre Variables

Objetivo: Visualizar las relaciones entre variables numéricas y observar posibles patrones.

- Crear gráficos de dispersión (scatter plots) utilizando seaborn para visualizar relaciones entre age y lpsa, lweight y lcavol.
- Añadir líneas de regresión a los gráficos de dispersión para mostrar la tendencia de los datos.
- Interpretar las gráficas para entender si hay relaciones lineales o no lineales entre las variables.

Responde: ¿Existen patrones claros o relaciones lineales entre las variables visualizadas? ¿Qué nos indican las pendientes de las líneas de regresión sobre estas relaciones?

4. Pruebas de Hipótesis

Objetivo: Realizar pruebas de hipótesis para comparar medias entre diferentes grupos de datos.

- Formular hipótesis nulas y alternativas, por ejemplo, que la media de lpsa para pacientes con svi positivo no es diferente de la de pacientes con svi negativo.
- Utilizar scipy.stats para realizar pruebas t de dos muestras y calcular intervalos de confianza para las medias de lpsa en ambos grupos.
- Interpretar los resultados de las pruebas de hipótesis para determinar si se rechaza la hipótesis nula.

Responde: ¿La prueba de hipótesis sugiere una diferencia significativa en los niveles de lpsa entre los grupos con svi positivo y negativo? ¿Qué podemos inferir de los intervalos de confianza calculados?

5. Análisis Exploratorio de Datos (EDA)

Objetivo: Realizar un análisis exhaustivo para entender mejor el conjunto de datos.

- Explorar la estructura de los datos, incluyendo la identificación de valores nulos o faltantes y la proporción de datos categóricos frente a datos numéricos.
- Utilizar gráficos de barras para visualizar la distribución de la variable categórica gleason.
- Implementar diagramas de caja y bigotes (box plots) y diagramas de violín (violin plots) para visualizar la distribución y variabilidad de las variables numéricas.

Responde: ¿Qué revela el análisis exploratorio sobre la calidad y características de los datos? ¿Cómo influye la puntuación de Gleason en las demás variables?

Reto.

Usando la base de datos prostate, grafique **lpsa** contra **lcavol**. Ajuste las regresiones de **lpsa** contra **lcavol** y **lcavol** contra **lpsa**. Muestre ambas regresiones. ¿En qué punto se intersectan las dos rectas?

Responde: ¿Qué interpretación se le puede dar al punto de intersección de las dos rectas de regresión? ¿Qué sugiere esto sobre la relación entre lcavol y lpsa?

Realizar el script usando un cuaderno. Iypnb**.** Tener en cuenta el uso de *Markdowns* para separar cada sección del código y para la resolución de preguntas. Comente debidamente el código. Se evaluará la estética y el orden del script.

LUIS ESTEBAN GOMEZ CADAVID

Director de Programa – Líder de Analítica de Datos

Facultad de Ingenierías - Universidad de San Buenaventura Medellín

Email: analitica@usbmed.edu.co

Medellín - Colombia