UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
BIOINGENIERÍA
PROBABILIDAD Y ESTADÍSTICA- 2535331
Jenny Kateryne Aristizábal Nieto
Semestre 2024-1

TERCER PROYECTO

ESTADÍSTICA PREDICTIVA

1. INTRODUCCIÓN

El estudio que se presenta tiene como objetivo desarrollar un modelo predictivo para estimar la esperanza de vida media de los habitantes de una ciudad. Para ello, se cuenta con un conjunto de variables demográficas y socioeconómicas que se consideran relevantes en la determinación de este indicador tan importante para la calidad de vida de una población. Entre las variables disponibles se incluyen: el número total de habitantes, el porcentaje de analfabetismo, los ingresos medios, la tasa de homicidios, el porcentaje de personas con educación universitaria, la cantidad de días de heladas (invierno), el área total de la ciudad y la densidad poblacional.

La esperanza de vida es un indicador que refleja la salud y bienestar de una comunidad, y se encuentra influenciada por una amplia gama de factores. Al analizar estos datos, no solo se busca comprender la relación directa entre cada variable y la esperanza de vida, sino también identificar posibles interacciones y correlaciones entre ellas. Esto permitirá construir un modelo estadístico que capture de manera precisa la dinámica compleja que se atribuye a la esperanza de vida en una ciudad.

2. CONJUNTO DE DATOS

El conjunto de datos proporciona información detallada sobre diversos aspectos demográficos y socioeconómicos de varias ciudades (50 en total), con el objetivo de analizar y predecir la esperanza de vida media de sus habitantes. Las variables que se presentan son:

1. Habitantes (Population): Esta variable numérica indica el número total de habitantes en la ciudad. Representa el tamaño de la población y es un factor fundamental en el análisis demográfico.

2. Analfabetismo (illiteracy): Es una variable que refleja el porcentaje de habitantes que son analfabetos. Este indicador educativo es importante para comprender el nivel de acceso a la educación en la ciudad.

3. Ingresos (Income): Representa los ingresos medios de los habitantes de la ciudad. Esta variable económica proporciona información sobre el nivel socioeconómico de la población.

4. Esperanza de Vida (Life_exp): Es la variable objetivo del estudio. Indica la esperanza de vida media de los habitantes de la ciudad en años. Este es un indicador crítico de la calidad de vida y la salud de la población.

5. Asesinatos (Murder): Representa la tasa de homicidios en la ciudad, expresada como el número de homicidios por cada 100.000 habitantes. Esta variable está relacionada con la seguridad y el nivel de violencia en la comunidad.

6. Universitarios (HS_Grad): Indica el porcentaje de habitantes con educación universitaria. Esta variable proporciona información sobre el nivel de educación de la población.

7. Heladas (Frost): Representa la cantidad de días al año en los que se registran heladas. Este dato climático puede tener implicaciones en la salud y el bienestar de la población.

8. Área (Area): Indica el área total de la ciudad en kilómetros cuadrados. Es un factor relevante para comprender la distribución geográfica de la población.

9. Densidad Poblacional (Density_pob): Representa la densidad de población de la ciudad, expresada como el número de habitantes por kilómetro cuadrado. Esta variable proporciona información sobre la concentración de la población en el territorio.

Este conjunto de datos es de gran relevancia para comprender los factores que pueden influir en la esperanza de vida de los habitantes de la ciudad y servirá como base para la construcción de un modelo predictivo.

3. PROCEDIMIENTO

3.1. Análisis exploratorio de datos (AED)

Utilice la tarea en tiempo real de Matlab “compute by group” para calcular múltiples estadísticas descriptivas de las variables. (media, max, min, std, var, datos perdidos, etc). Se debe realizar un análisis exhaustivo de los conjuntos de datos, incluyendo gráficos descriptivos (histogramas y barras y bigotes), y distribuciones de las variables relevantes. Este paso es crucial para comprender la naturaleza y comportamiento de los datos, identificar posibles outliers y determinar la necesidad de transformaciones.

3.2. Estadística inferencial

1. Realice diferentes modelos de regresión lineal univariada teniendo en cuenta la variable dependiente y cada una de las variables independientes o predictoras. Realice gráficos de predictor vs respuesta real y predicha por el modelo, use las diferentes métricas de

evaluación para definir si hay relación lineal o no entre cada predictor y la variable predecir.

2. Realice un modelo lineal multivariado que incluya todas las variables independientes o predictoras, para predecir la respuesta. Incluya gráficos de dispersión de valores rales vs predichos, y analice las métricas e evaluación.

3. Realice nuevamente un modelo multivariado, pero esta vez con las variables predictoras que considere pertinente, para esto haga uso de lo siguiente:
    a. Verificación de Supuestos: Realizar pruebas de normalidad y de varianza
    b. Test de correlación

4. Pruebe otros modelos de regresión que ofrece la app de Matlab (Arboles de decisión, redes neuronales, etc) y evalúe las métricas en la tabla de comparación que tiene la app.

Tenga en cuenta lo siguiente para cada uno de los modelos a realizar:

Selección de Variables: Es importante explicar la razón por la cual se han seleccionado específicas variables para incluir en el modelo (punto 3). Se debe justificar cómo estas variables están relacionadas con el fenómeno de interés y si su aporte es significativo.

Métricas de Evaluación: Se deben evaluar las métricas apropiadas para la calidad del modelo. Para el análisis de regresión, esto podría incluir el R-cuadrado ajustado, el error estándar de la estimación, la significancia de los coeficientes y las demás vistas en clase.

Validez de los Resultados: Se debe discutir la validez de los resultados obtenidos. Esto implica considerar posibles sesgos, limitaciones del estudio y la generalización de los hallazgos a la población de interés. Además, se debe mencionar cualquier técnica o procedimiento utilizado para mitigar posibles fuentes de sesgo.

Ecuación del modelo: Se debe presentar la ecuación del modelo de regresión para cada punto y explicar la influencia de cada una de las variables sobre la predicción, además debe identificar cual es la variable que más contribuye y si es congruente con lo esperado.

4. PARÁMETROS DE ENTREGA

- Entregar un informe en un livescript de Matlab, se tendrá en cuenta el orden en la presentación de resultados.
- El informe debe contener todos los análisis de cada uno de los resultados presentados, luego incluir un apartado de conclusiones. Recuerde utilizar el lenguaje estadístico apropiado para establecer las conclusiones.