UNIVERSIDAD DE ANTIOQUIA
Facultad de Ingeniería

Programa de Bioingeniería

Probabilidad y estadística

Tercer proyecto: Aprendizaje supervisado

Introducción al aprendizaje supervisado: Clasificación y Regresión

En el marco de este proyecto, abordaremos un enfoque integral para analizar conjuntos de datos heterogéneos, empleando técnicas de clasificación y regresión. Para la tarea de clasificación, utilizaremos un conjunto de datos para un análisis detallado mediante regresión logística, con el objetivo de discernir categorías claramente definidas. Asimismo, para el análisis de regresión, contaremos con un segundo conjunto de datos que requerirá la aplicación de modelos de regresión lineal múltiple, permitiendo así explorar relaciones más complejas entre las variables y prever resultados continuos. Este enfoque multidisciplinario nos proporcionará una comprensión más profunda y matizada de los fenómenos estudiados, brindando así un panorama completo de los factores que influyen en los resultados de interés.

MODELO DE CLASIFICACIÓN

Introducción

El cáncer de mama es una afección de gran relevancia en el ámbito de la salud, y su detección temprana es crucial para mejorar las tasas de supervivencia y calidad de vida de las pacientes. En este contexto, la aplicación de técnicas estadísticas desempeña un papel fundamental al permitirnos analizar y comprender los factores que influyen en la naturaleza del crecimiento anormal de células mamarias. El objetivo de este estudio es emplear un conjunto de variables clave, incluyendo características como el radio, textura, perímetro, área, suavidad y otros, para predecir si el crecimiento es benigno o maligno.

Para llevar a cabo esta tarea, se han recolectado datos de múltiples células mamarias, y se han calculado treinta características distintas para cada una de ellas. Estas características incluyen medidas como el radio medio, la textura, el perímetro, el área y otras propiedades relevantes. Adicionalmente, se han considerado medidas de la media y el error estándar para cada una de estas características. La columna "diagnóstico" indica si el crecimiento es maligno (M) o benigno (B).

El objetivo de este trabajo es desarrollar un modelo predictivo que, basado en estas características, pueda determinar con precisión si un crecimiento es benigno o maligno. Para lograr esto, se deben utilizar técnicas avanzadas de estadística inferencial y machine learning, aprovechando la riqueza de información contenida en estas variables para tomar decisiones clínicas informadas.

Conjunto de datos

En el marco del estudio sobre el cáncer de mama, disponemos de un conjunto de datos definidos de la siguiente manera:

1. Diagnóstico (Clase): Esta variable categórica indica la clasificación de los casos en dos grupos distintos: "Maligno" y "Benigno". Esta clasificación es fundamental en nuestro análisis, ya que determina la naturaleza del crecimiento celular.

2. Radio Medio (mean_radius): Representa las distancias promedio desde el centro hasta los puntos en el perímetro de la célula mamaria. Se trata de una variable numérica continua.

3. Textura Media (mean_texture): Esta variable numérica continua refleja la desviación estándar de los valores de escala de grises en la célula mamaria.

4. Perímetro Medio (mean_perimeter): Indica el perímetro promedio de la célula mamaria y es una variable numérica continua.

5. Área Media (mean_area): Representa el área media de la célula mamaria y es una variable numérica continua.

6. Suavidad Media (mean_smoothness): Esta variable numérica continua describe la variación local en las longitudes de radio.

7. Compactación Media (mean_compactness): Calculada como (perímetro^2 / área - 1.0), esta variable numérica continua indica la compacidad promedio de la célula mamaria.

8. Concavidad Media (mean_concavity): Representa la gravedad de las porciones cóncavas del contorno de la célula mamaria y es una variable numérica continua.

9. Puntos Cóncavos Medios (mean_concave_points): Indica el número promedio de porciones cóncavas en el contorno de la célula mamaria. Es una variable numérica continua.

10. Simetría Media (mean_symmetry): Esta variable numérica continua refleja la simetría promedio de la célula mamaria.

11. Dimensión Fractal Media (mean_fractal_dimension): Es una variable numérica continua.

12. Radio del Error Estándar (se_radius): Esta variable numérica continua indica el error estándar de las distancias desde el centro hasta los puntos en el perímetro de la célula mamaria.

13. Textura del Error Estándar (se_texture): Refleja el error estándar de la desviación estándar de los valores de escala de grises.

14. Perímetro del Error Estándar (se_perimeter): Indica el error estándar del perímetro de la célula mamaria y es una variable numérica continua.

15. Área del Error Estándar (se_area): Representa el error estándar del área de la célula mamaria y es una variable numérica continua.

16. Suavidad del Error Estándar (se_smoothness): Esta variable numérica continua describe el error estándar de la variación local en las longitudes de radio.

17. Compactación del Error Estándar (se_compactness): Calculada como el error estándar de la compactación de la célula mamaria, es una variable numérica continua.

18. Concavidad del Error Estándar (se_concavity): Representa el error estándar de la gravedad de las porciones cóncavas del contorno de la célula mamaria y es una variable numérica continua.

19. Puntos Cóncavos del Error Estándar (se_concave_points): Indica el error estándar del número de porciones cóncavas en el contorno de la célula mamaria. Es una variable numérica continua.

20. Simetría del Error Estándar (se_symmetry): Esta variable numérica continua refleja el error estándar de la simetría de la célula mamaria.

21. Dimensión Fractal del Error Estándar (se_fractal_dimension): Representa el error estándar de la dimensión fractal y es una variable numérica continua.

22. Peor Radio (worst_radius): Indica el peor valor del radio de la célula mamaria y es una variable numérica continua.

23. Peor Textura (worst_texture): Refleja el peor valor de la desviación estándar de los valores de escala de grises.

24. Peor Perímetro (worst_perimeter): Representa el peor valor del perímetro de la célula mamaria y es una variable numérica continua.

25. Peor Área (worst_area): Indica el peor valor del área de la célula mamaria y es una variable numérica continua.

26. Peor Suavidad (worst_smoothness): Esta variable numérica continua describe el peor valor de la variación local en las longitudes de radio.

27. Peor Compactación (worst_compactness): Calculada como el peor valor de la compactación de la célula mamaria, es una variable numérica continua.

28. Peor Concavidad (worst_concavity): Representa el peor valor de la gravedad de las porciones cóncavas del contorno de la célula mamaria y es una variable numérica continua.

29. Peor Puntos Cóncavos (worst_concave_points): Indica el peor valor del número de porciones cóncavas en el contorno de la célula mamaria. Es una variable numérica continua.

30. Peor Simetría (worst_symmetry): Esta variable numérica continua refleja el peor valor de la simetría de la célula mamaria.

31. Peor Dimensión Fractal (worst_fractal_dimension): Representa el peor valor de la "aproximación a la costa" menos uno y es una variable numérica continua.

Estas variables nos proporcionan una visión detallada de las características de las células mamarias, lo que es esencial para el análisis y la predicción de la naturaleza de su crecimiento. Cada una de ellas aporta información valiosa que puede ser utilizada en la construcción de modelos predictivos precisos y relevantes para la detección temprana y el tratamiento efectivo del cáncer de mama.

MODELO DE REGRESIÓN

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 crucial 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, 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 buscaremos 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 subyace a la esperanza de vida en esta ciudad.

Conjunto de datos

El conjunto de datos proporciona información detallada sobre diversos aspectos demográficos y socioeconómicos de una ciudad, con el objetivo de analizar y predecir la esperanza de vida media de sus habitantes.

1. Habitantes: 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: 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: 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: 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: 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: Indica el porcentaje de habitantes con educación universitaria. Esta variable proporciona información sobre el nivel de educación de la población.

7. Heladas: 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: 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: 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.

Procedimiento

En este estudio, se requiere un riguroso proceso de análisis que abarque desde la exploración inicial hasta la validación de los resultados obtenidos. A continuación, se detalla el enfoque metodológico a seguir:

1. **Análisis Exploratorio de Datos (AED):** Se debe realizar un análisis exhaustivo de los conjuntos de datos, incluyendo gráficos descriptivos, estadísticas resumen 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.

2. **Verificación de Supuestos**: Se deben verificar los supuestos subyacentes a los modelos seleccionados.

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

4. **Métricas de Evaluación:** Se deben definir métricas apropiadas para evaluar la calidad del modelo. Para el análisis de regresión, esto podría incluir el R-cuadrado ajustado, el error estándar de la estimación y la significancia de los coeficientes. En el caso de la regresión logística, la sensibilidad, exactitud, precisión y f1-score.

5. **Descripción del Modelo Implementado:** Se debe proporcionar una descripción detallada del modelo implementado, incluyendo la especificación de las variables independientes, la forma funcional del modelo y cualquier consideración especial en la selección de los términos.

6. 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.

7. Ecuación del modelo: Se debe presentar la ecuación del modelo de regresión 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.

Parámetros de entrega

Entregar un informe en un Notebook, se tiene en cuenta el orden en la presentación de resultados, análisis exploratorio, gráficos seleccionados, análisis de tendencias, evaluación de supuestos.

Adicionalmente se debe referenciar que enunciado se está desarrollando y el objetivo del desarrollo.

El informe debe contener conclusiones de cada uno de los resultados presentados, sin el apartado de conclusiones no se revisa y la nota final es 0.

Finalmente, la nota estará definida por el trabajo escrito y la sustentación (presencial).

La nota del trabajo se define en la sustentación. Trabajo sin sustentación tiene como nota final 0.

Trabajo escrito 70%, sustentación 30%.