Estimación de niveles de obesidad en México, Perú y Colombia mediante redes neuronales: un enfoque basado en hábitos alimenticios y condiciones físicas

Alejandro Álvarez Patiño

Universidad Nacional de Colombia sede Manizales

2023

# RESUMEN

Este estudio presenta un conjunto de datos diseñado para estimar los niveles de obesidad en individuos de Colombia, México y Perú. Se desarrolló una red neuronal basada en hábitos alimenticios y condiciones físicas, utilizando un conjunto de datos integral con 17 atributos y 2111 registros. El 77% de los datos se generó sintéticamente mediante la herramienta Weka y el filtro SMOTE, mientras que el 23% se recopiló directamente de usuarios a través de una plataforma web. La red neuronal, compuesta por capas de entrada, ocultas y salida, fue entrenada durante 20 épocas, mostrando una mejora sustancial en pérdida y precisión. En la época final, la precisión de validación alcanzó el 93.85%, respaldando la capacidad del modelo para generalizar a nuevos datos. Posteriormente, evaluado con datos de prueba, el modelo logró una precisión adicional del 94.56%. Estos resultados demuestran la eficacia de la red neuronal en la estimación de niveles de obesidad, brindando una herramienta para abordar la salud pública en la región.

ÍNDICE

[RESUMEN 1](#_Toc151728849)

[DATOS 3](#_Toc151728850)

[I. Descripción 3](#_Toc151728851)

[II. Transformación de los datos 3](#_Toc151728852)

[RED NEURONAL 4](#_Toc151728853)

[I. Selección de Arquitectura y Hiperparámetros: 4](#_Toc151728854)

[CONCLUSIÓN 5](#_Toc151728855)

[REFERENCIAS 6](#_Toc151728856)

# DATOS

## Descripción

El conjunto de datos se centra en la estimación de los niveles de obesidad en individuos con edades comprendidas entre los 14 y 61 años, capturando una diversidad de hábitos alimenticios y condiciones físicas. Se recopilaron respuestas anónimas a través de una encuesta en línea que abordaba preguntas relacionadas con el género, edad, altura, peso, antecedentes familiares de sobrepeso, frecuencia de consumo de alimentos calóricos, consumo de vegetales, número de comidas principales, consumo de alimentos entre comidas, consumo diario de agua, consumo de alcohol, monitoreo de la ingesta calórica, frecuencia de actividad física, tiempo de uso de dispositivos tecnológicos y medio de transporte utilizado.

La estructura de datos consta de atributos que abordan tanto los hábitos alimenticios como la condición física, proporcionando un conjunto completo para análisis y aplicaciones de minería de datos, como clasificación, predicción, segmentación y asociación. La variable de clase "NObesity" clasifica a los individuos en siete categorías distintas de obesidad. Además, se realizó un proceso de equilibrio de clases mediante la generación sintética del 77% de los datos utilizando la técnica SMOTE de la herramienta Weka, asegurando un conjunto de datos más representativo para técnicas de aprendizaje automático. [1]

## Transformación de los datos

El proceso de transformación de datos se llevó a cabo en varias etapas para asegurar la preparación adecuada del conjunto de datos para su posterior uso en la construcción de modelos, particularmente en el contexto de la creación de una red neuronal. A continuación, se detallan cada una de las etapas del proceso:

* Carga de Datos:
  + Se utilizó la biblioteca pandas para cargar el conjunto de datos original.
* Categorización de Datos Categóricos:
  + Se identificaron las columnas categóricas relevantes, como el género, historial familiar de sobrepeso, etc.
  + Se aplicó la técnica de one-hot encoding para convertir estas columnas categóricas en variables binarias.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.

# RED NEURONAL

## Selección de Arquitectura y Hiperparámetros:

La elección de la arquitectura de la red neuronal y los hiperparámetros es una tarea crucial que influye directamente en el rendimiento del modelo. A continuación, se explican las decisiones tomadas para cada aspecto de la arquitectura y los hiperparámetros en el código proporcionado:

* Capas de Entrada:
  + La capa de entrada tiene 27 unidades, que corresponde al número de características en el conjunto de datos procesado.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.
* Normalización de Datos Numéricos:
  + Se seleccionaron las columnas numéricas, como la edad, altura, peso, etc.
  + Se normalizaron estas variables numéricas para llevarlas a una escala común entre 0 y 1.

Estas elecciones se realizaron mediante un proceso de ajuste y prueba, teniendo en cuenta la naturaleza del problema de clasificación multiclase y buscando un equilibrio entre la capacidad de representación del modelo y la prevención del sobreajuste.

# CONCLUSIÓN

Los resultados obtenidos durante el entrenamiento de la red neuronal proporcionan una visión clara del rendimiento del modelo a lo largo de las épocas. A continuación, se resumen los resultados clave y se presentan algunas conclusiones:

Imagen que contiene Calendario

Descripción generada automáticamenteGráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

La pérdida y la precisión en el conjunto de entrenamiento y validación muestran una mejora significativa a lo largo de las épocas, indicando que el modelo está aprendiendo de manera efectiva los patrones en los datos. La precisión de validación alcanzada alrededor del 93.85% en la época 20 sugiere que el modelo generaliza bien a datos no vistos. La disminución constante de la pérdida en ambos conjuntos indica que el modelo está convergiendo y mejorando su capacidad predictiva. Las gráficas de Loss y Accuracy confirman la efectividad del modelo al aprender y generalizar a partir de los datos proporcionados. 

Después de evaluar el modelo con datos de prueba, se obtuvo una precisión adicional del 94.56%, confirmando la habilidad del modelo para realizar predicciones precisas en datos no vistos. La red neuronal ha aprendido de manera efectiva y generaliza bien a nuevos datos, proporcionando una herramienta para la estimación de niveles de obesidad basada en hábitos alimenticios y condiciones físicas.

# REFERENCIAS

1. https://www.sciencedirect.com/science/article/pii/S2352340919306985?ref=pdf\_download&fr=RR-2&rr=82b33df55fba257d.