UNIVERSIDAD INTERNACIONAL DE LA RIOJA

MAESTRÍA EN ANÁLISIS DE DATOS Y BIG DATA

Actividad grupal: Definición de un Problema Estadístico: Modelización y Propuesta de Soluciones

Autores:

Alcibiades Paredes Martés

Julia Silvana Huayta Gómez

Christian Casallas Benítez

(Todos realizamos el trabajo a la par mediante una reunión)

Nohora Emilce Silva Ramos

(Se retiró de la maestria)

Profesor:

Yeray Mezquita Martin

2024

Definición del problema

La diabetes es una de las enfermedades crónicas más prevalentes en el mundo, afecta a millones de personas cada año y las personas pierden la capacidad de regular eficazmente los niveles de glucosa(azúcar) en la sangre. Si bien no existe una cura para la diabetes existen estrategias para evitarlo o controlarlo como; perder peso, comer saludablemente, hacer ejercicios y recibir tratamientos médicos. Este estudio se enfoca en investigar la relación entre ciertas variables médicas y el diagnóstico de diabetes en mujeres de al menos 21 años de edad y de ascendencia india pima.

Antecedentes

Se calcula que alrededor de 422 millones de individuos en el mundo padecen de Diabetes Mellitus (DM) tipo 2. Este número ha experimentado un incremento notable desde 1980, y según las proyecciones del Diabetes Atlas (2019), se anticipa que alcanzará la cifra de 592 millones para el año 2035.

Modelos Estadísticos

Modelo estadístico 1: Análisis comparativo

Base de datos: Embarazos, Glucosa, Presión arterial, grosor de la piel, insulina, IMC, probabilidad de diabetes según antecedentes, edad, resultado.

El dataser

Esquema de Cálculos Numéricos:

Análisis Descriptivo Comparativo:

Realizar estadísticas descriptivas para cada variable en el conjunto de datos, focalizando en el diagnóstico de diabetes, con el propósito de ofrecer una visión global de las disparidades y similitudes en las características médicas de los individuos.

Visualización Comparativa:

Utilizar gráficos comparativos como gráficos de barras para visualizar la relación entre el diagnostico de diabetes y las variables seleccionadas

Pruebas de Significancia: Realizar pruebas de significancia estadística para comparar las medias o medianas de las variables con diagnostico positivo y negativo de diabetes y determinar si las diferencias observadas son estadísticamente significativas.

Identificación de factores distintivos: Identificar los factores que parecen ser distintivos o significativamente diferentes entre los diagnósticos positivos y negativos de diabetes, lo que podría indicar áreas clave.

Correlaciones Globales: Explorar correlaciones globales entre las variables para entender cómo se relacionan de manera general para un diagnostico positivo y negativo

Variables elegidas en la base de datos

Glucose (Glucosa): La glucosa en sangre es un indicador clave para el diagnóstico de la diabetes. Comparar los niveles de glucosa entre las personas con y sin diabetes puede proporcionar información valiosa.

BMI (Índice de Masa Corporal): El BMI es otro indicador relevante. Las personas con diabetes a menudo tienen problemas de peso, y el BMI puede reflejar esto.

Age (Edad): La diabetes a menudo está asociada con la edad. Comparar las edades de las personas con y sin diabetes puede revelar patrones.

DiabetesPedigreeFunction (porcentaje de antecedentes de diabetes): Esta función expresa la probabilidad de diabetes según antecedentes familiares. Podría ser interesante examinar cómo varía entre los dos grupos.

Insuline (Insulina): Expresa el nivel de insulina en la sangre

Resumen descriptivo

```
> summary(datos$Glucose)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    99.0   117.0   120.9   140.2   199.0
> summary(datos$BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   27.30   32.00   31.99   36.60   67.10
> summary(datos$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   24.00   29.00   33.24   41.00   81.00
> summary(datos$DiabetesPedigreeFunction)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0780  0.2437  0.3725  0.4719  0.6262  2.4200
> summary(datos$Insulin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0    30.5    79.8   127.2   846.0
```

Figura 1. Resumen descriptivo estadístico.

Visualización comparativa

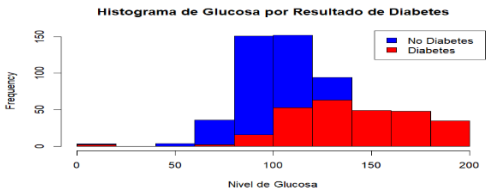

Figura 2. Histograma de Glucosa por Resultado de Diabetes

En el histograma que representa la relación entre los niveles de glucosa y los resultados de diabetes, se observa claramente que a medida que los niveles de glucosa aumentan, la probabilidad de tener diabetes también aumenta. Es notable destacar que todas las personas cuyos niveles de glucosa son iguales o superiores a 150 han sido diagnosticadas con diabetes.

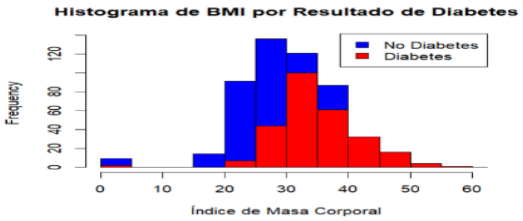

Figura 3. Histograma de BMI por Resultado de diabetes

En el histograma que representa el Índice de Masa Corporal (IMC) en función de los resultados de diabetes, se observa una notable tendencia: a partir de un IMC de 30, se incrementa significativamente el número de casos de personas con diabetes. Esta relación se intensifica aún más para aquellas personas cuyo IMC es igual o superior a 40, donde se evidencia de manera concluyente la presencia de diabetes.

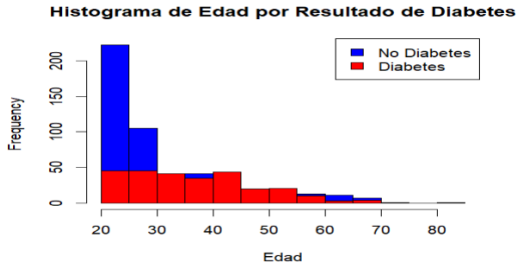

Figura 4. Histograma de Edad por Resultado de Diabetes

En el histograma que representa la edad en relación con los resultados de diabetes, se aprecia que no existe una conexión clara entre la edad de una persona y la probabilidad de desarrollar diabetes, ya que este diagnóstico puede manifestarse en individuos de cualquier grupo etario. Sin embargo, resulta notable que los casos se vuelven más frecuentes a partir de los 20 años, sugiriendo que los diagnósticos de diabetes en personas menores de 20 años son poco comunes.

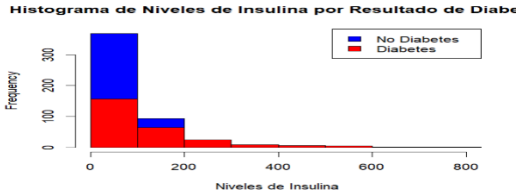

Figura 5. Histograma de Niveles de Insulina por Resultado de Diabetes

En el histograma, se aprecia que alrededor de 150 personas con niveles de insulina en el rango de 0 a 100 tienen un diagnóstico positivo de diabetes, mientras que aproximadamente 200 personas en ese mismo rango no presentan diabetes. Además, se observa un aumento significativo en la cantidad de personas con diabetes en el rango de insulina de 100 a 200, con alrededor de 20 personas sin diabetes en este intervalo. Notablemente, en el rango de 200 en adelante, todas las personas tienen un diagnóstico positivo de diabetes.

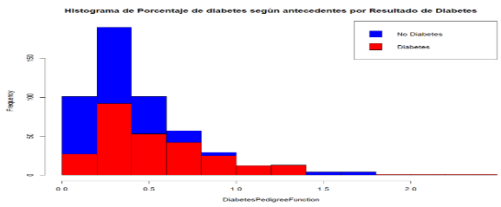

Figura 6. Histograma de porcentaje de diabetes según antecedentes por Resultado de diabetes

En el histograma se puede visualizar que no hay necesariamente una relación directa entre el porcentaje de diabetes según antecedentes y un diagnóstico positivo de diabetes.

Correlación de Pearson

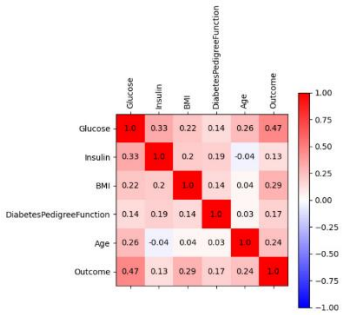

Figura 7. Mapa de calor correlación de Pearson

Interpretación de la correlación de Pearson

Outcome con porcentaje (Resultado de Diabetes): Tiene una correlación positiva con Glucose (0.47), BMI (0.29) y DiabetesPedigreeFunction (0.17). Esto sugiere que a medida que estas variables aumentan, es más probable que el resultado de diabetes también sea positivo.

La correlación con Edad (Age) es moderada (0.24), indicando cierta asociación.

Glucose: Tiene una correlación significativa y positiva con Outcome (0.47), lo que sugiere que niveles más altos de glucosa están asociados con un mayor riesgo de diabetes. También tiene una correlación positiva con BMI (0.22) e Insulin (0.33).

BMI (Índice de Masa Corporal): Tiene una correlación positiva moderada con Glucose (0.22) e Insulin (0.20). También tiene una correlación positiva con Outcome (0.29), sugiriendo que un mayor BMI se asocia con un mayor riesgo de diabetes.

Edad (Age): Tiene una correlación moderada con Glucose (0.26) y una correlación débil con Outcome (0.24).

Insulin: Tiene una correlación positiva con Glucose (0.33) y una correlación moderada con DiabetesPedigreeFunction (0.19).

DiabetesPedigreeFunction: Tiene correlaciones débiles con las otras variables, excepto con Glucose (0.14) e Insulin (0.19).

Modelo Estadístico 2: Análisis de componentes principales (PCA)

Base de datos: Se da uso a las mismas variables mencionadas con anterioridad

Esquema de Cálculos Numéricos:

Normalización de datos:

Normalizar las variables para asegurar que tengan la misma escala y contribución relativa en el análisis.

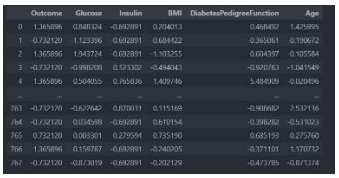

Figura 8. Normalización de los datos

La normalización de la escala de las variables se llevó a cabo con el objetivo de ajustar los datos de manera que tengan una media de 0 y una desviación estándar de 1

Aplicación de PCA:

Aplicar la técnica de PCA para reducir la dimensionalidad de los datos y encontrar los componentes principales que explican la mayor varianza en los datos.

	Outcome	Glucose
0	1.520533	-1.556501
1	-1.661805	-0.197958
2	1.239436	-0.898396
3	-1.617175	0.641906
4	3.379496	2.322944
...	...	...
763	0.059628	-1.356063
764	-0.632437	0.174098
765	-0.832886	0.084607
766	0.662281	-1.691561
767	-1.554492	0.314755

Figura 9. Evidencia de resultados de PCA

Se evidencia que los dos componentes principales del estudio serán la columna de resultado y glucosa, ya que se obtuvo una varianza de 33,68% y 18,85% respectivamente.

Interpretación de componentes principales:

Interpretar los componentes principales identificados, destacando las variables que más contribuyen a cada componente.

Visualización de Resultados de PCA:

Crear gráficos de dispersión o biplots para visualizar la distribución de los países en el espacio de los componentes principales.

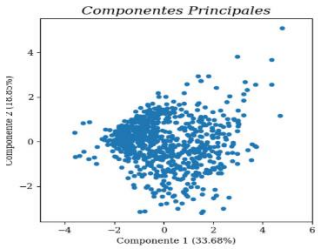

Figura 10. Gráfico de dispersión de componentes principales

En la gráfica de dispersión de los competentes principales se visualiza una correlación positiva y moderada entre los dos componentes.

Correlaciones con Componentes:

La correlación presentada indica un alto agrupamiento en un sector de la gráfica, pero esta no es tan significativa para determinar por si sola el resultado en el diagnóstico de la diabetes.

Discusión

Al comparar los resultados del análisis comparativo y del PCA, se observan patrones clave en las variables relacionadas con el diagnóstico de diabetes en mujeres de ascendencia india pima. El análisis comparativo destaca la influencia de la glucosa, BMI y la edad, mientras que el PCA señala la importancia crucial de la glucosa y el resultado de diabetes como componentes principales. A pesar de reconocer correlaciones, la complejidad de estas destaca la necesidad de un enfoque individualizado para evaluar con precisión los riesgos de diabetes.

Conclusiones

El modelo presentado muestra una correlación entre las variables glucosa, BMI y edad; pero, al contrario, no se observa una correlación con porcentaje de diabetes según precedentes y los niveles de insulina. Esto nos lleva a concluir que el diagnóstico positivo de diabetes va a depender del análisis de cada caso puntual; ya que dependiendo de cada paciente puede que con un solo factor puede dar un resultado de diabetes positivo.

Líneas Futuras

Según lo analizado en el modelo se recomendaría que los países inviertan en los recursos para prevenir y detectar tempranamente los casos de diabetes, esto de la mano de campañas de concientización para que el estilo de vida de sus habitantes se enfoque en comer saludablemente, estar activos y perder peso, que son los principales mitigantes para evitar el daño de la diabetes.

Dataset extraído de: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download

Referencias

Edición, I. D. (2019). *IDF Diabetes Atlas*. Obtenido de https://www.diabetesatlas.org/upload/resources/material/20200302_133352_2406-IDF-ATLAS-SPAN-BOOK.pdf