Progresión de la Diabetes

Alexis Hernández Morales

Facultad de Ciencias Físico Matemáticas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, México.

1. Introducción

La diabetes es una enfermedad crónica que afecta directamente el páncreas y causando que el cuerpo sea incapaz de producir insulina. La insulina es la principal responsable de mantener el nivel de glucosa en la sangre. La variedad de factores como la obesidad, sedentarismo, alta presión en la sangre, pueden causar afectaciones en las personas que padecen diabetes. Puede dañar la piel, los ojos y en casos más graves daño renal.

Definir de pronta manera tratamientos precisos y efectivos puede ser retador para profesionales en la medicina. Métodos de aprendizaje automático pueden proveer información valiosa facilitando la asignación del tratamiento y los factores con mayor peso a considerar, reduciendo la carga de trabajo verificando otros factores que tal vez no tengan tanto impacto.

2. Metodología

2.1. Datos

La base de datos (Sklearn.datasets, Load diabetes) consiste de 10 variables siendo, edad, sexo, índice de masa corporal, presión sanguínea promedio, serúm total de colesterol, lipoproteína de baja densidad, lipoproteína de alta densidad, colesterol total, nivel de triglicéridos, y nivel de azúcar en la sangre. Cuenta con la información de 442 pacientes. La variable objetivo es una medida cuantitativa de la progresión del padecimiento.

2.2. K - Medias

El agrupamiento por K - Medias consiste en generar K grupos para n elementos creando una partición en la base de datos en donde cada dato pertenece solamente a un grupo intentando mantener los datos en el agrupamiento lo mas parecidos posible mientras que los agrupamientos manteniéndolos lo mas distantes posibles. Asigna los datos utilizando generalmente como referencia la distancia euclidiana respecto a un centroide.

2.3. Árbol de Decisión

El Árbol de Decisión es un proceso que permiten la construcción de modelos de predicción basándose en su clasificación según ciertas características o propiedades, o en la regresión mediante la relación entre distintas variables para predecir el valor de otra. Esta estructurado por ramas y nodos, los nodos internos representan cada una de las características o propiedades a considerar para tomar una decisión, las ramas representan la decisión en función de una determinada condición y los nodos finales representan el resultado de la decisión.

2.4. Bosque Aleatorio

El Bosque Aleatorio es un método de aprendizaje automático que utiliza Árboles de Decisión para crear predicciones y clasificaciones. Cuando el Bosque Aleatorio esta prediciendo un nuevo ítem basado en ciertos atributos, cada Árbol de Decisión da su propia clasificación como resultado, entonces el resultado general del Bosque Aleatorio será el mayor numero de taxonomía. En el caso de una regresión el resultado sera el valor promedio de todos los Arboles de Decisión.

2.5. Métricas de error

Se utilizó el Error Absoluto Medio Porcentual y el Error de Raíz Cuadrada Media para medir la efectividad del modelo.

2.5.1. Error Absoluto Medio Porcentual

El Error Absoluto Medio Porcentual es un indicador de desempeño que mide el tamaño de error absoluto en términos porcentuales.

$$\text{EAMP} = \frac{100}{N} \cdot \frac{\sum_i \left| Y - \hat{Y} \right|}{Y}$$

2.5.2. Error de Raíz Cuadrada Media

El Error de Raíz Cuadrada Media es la desviación estándar de los valores residuales. Los valores residuales son una medida de la distancia de los puntos de datos de una regresión. El Error de Raíz Cuadrada Media mide cuál es el nivel de dispersión de estos valores residuales.

$$\text{ERCM} = \sqrt{\frac{\sum_i (Y - \hat{Y})^2}{N}}$$

2.6. Selección de Características Exhaustiva

El algoritmo de Selección de Características Exhaustiva evalúa todas las combinaciones que existen entre las variables de la base de datos devolviendo los valores con mejor ajuste al modelo utilizando una métrica de ajuste definida.

3. Resultados

3.1. Selección de Características Exhaustiva

Con el algoritmo de Selección de Características Exhaustiva se definieron las variables a utilizar durante el análisis, con un modelo de regresión lineal y la métrica de Error Medio Absoluto para medir el ajuste al modelo.

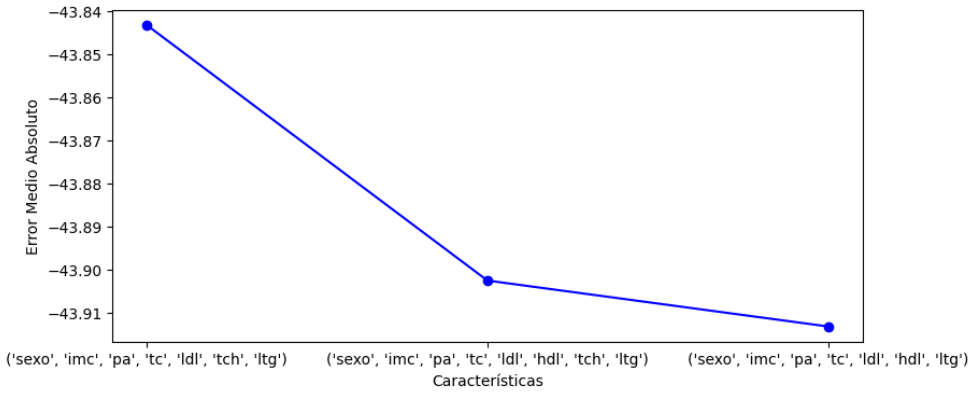

Figura 1: Combinación de características con mayor ajuste.

Se tomaron en cuenta los resultados obtenidos del algoritmo al igual que resultados en investigaciones previas (Shukla AK., 2020) para determinar las variables a utilizar siguientes: índice de masa corporal, presión sanguínea promedio, lipoproteína de baja densidad, colesterol total y nivel de triglicéridos.

3.2. K - Medias

Al utilizar el método de K-medias para la agrupación de los datos y el índice de Davies-Bouldin para definir la cantidad óptima de grupos se puede apreciar una división entre los datos. En la figura 1 vemos como existen tres clasificaciones con respecto al Índice de Masa Corporal y el Objetivo que nos marca el bienestar del paciente.

Demuestra ser muy efectiva la agrupación de los datos utilizando el método de K-medias, creando una visualización clara para la partición de los datos.

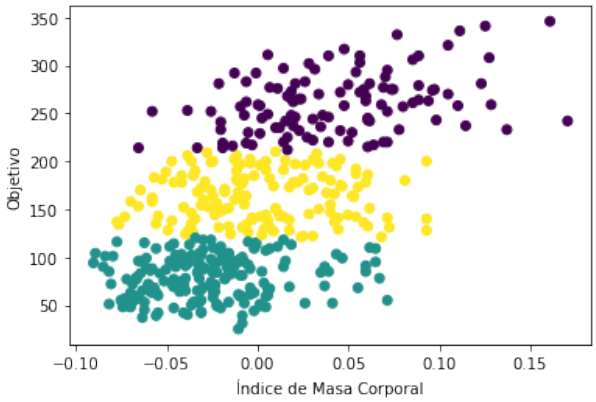

Figura 2: K-medias Agrupación.

3.3. Bosque Aleatorio

Como una forma de predicción de los datos se utilizó el método de Bosque Aleatorio. En el gráfico se muestra el comportamiento de los datos y su diferencia con la predicción realizada por el algoritmo de Bosque Aleatorio.

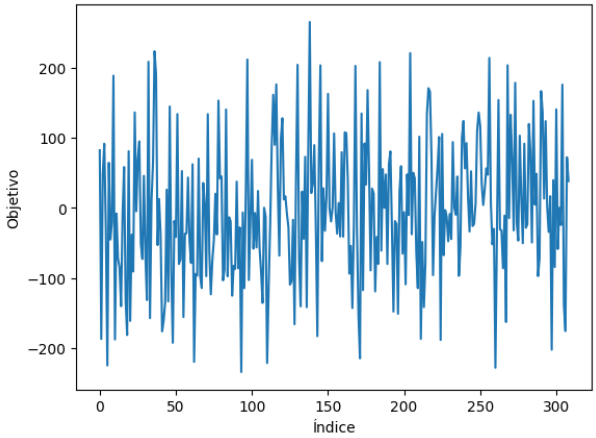

Figura 3: Diferencia de los datos contra la predicción.

Error Absoluto Medio Porcentual entre la predicción y los datos reales: 13.67 %

Error de Raíz Cuadrada Media entre la predicción y los datos reales: 25.37

4. Conclusión

La diabetes es un padecimiento que puede ser difícil de tratar de forma efectiva en cada paciente, crear predicciones muestra ser una forma útil y eficiente de apoyar dichos tratamientos.

Utilizando el algoritmo de Selección Exhaustiva se consiguió definir las variables con mayor significancia para el estudio dando lugar a un manejo mas sencillo y preciso de los datos.

El método de clasificación de K-medias mostró ser una gran herramienta para el agrupamiento de los datos mencionados, creando una división visual sencilla de apreciar y utilizar para mayores investigaciones. Los resultados presentados pueden llegar a ser de utilidad en ámbitos de salud enseñando un indicio de como se ve afectado el bienestar de un paciente dada una condición del mismo y de esta forma predecir lo que se debe atender o poner atención principalmente en los pacientes de diabetes.

El uso del algoritmo de Bosque Aleatorio fue de gran utilidad, creando predicciones con un grado de error bajo, permitiendo apreciar el comportamiento probable de los datos de forma eficiente, dando oportunidad a mayores investigaciones sobre las variables.

Bibliografía

Sklearn.datasets: Load diabetes. (n.d.). Scikit-learn.
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load.diabetes.html

Shukla AK (2020) Patient diabetes forecasting based on machine learning approach. In: Pant M, Sharma TK, Arya R, Sahana BC, Zolfagharinia H (eds) Soft computing: theories and applications. Advances in intelligent systems and computing, vol 1154. Springer, Singapore.

Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9.