Universidad del Valle de Guatemala
Minería de datos
Sección: 20

Luis Quezada 18028
Jennifer Sandoval 18962
Esteban del Valle 18221

Avances del proyecto

Tema de proyecto:

Análisis de datos para la predicción de diabetes en una persona en base a sus antecedentes médicos y condiciones presentes. El dataset contiene 769 casos, con información sobre las condiciones físicas de cada paciente. La idea es hacer un análisis exploratorio, luego revisar algunos agrupamientos (no supervisado). Finalmente, se realizará una prueba con los algoritmos de machine learning seleccionados para construir un clasificador contundente que permita diagnosticar a una persona con diabetes, sin necesidad de que sea revisado por un médico.

Resultados de EDA

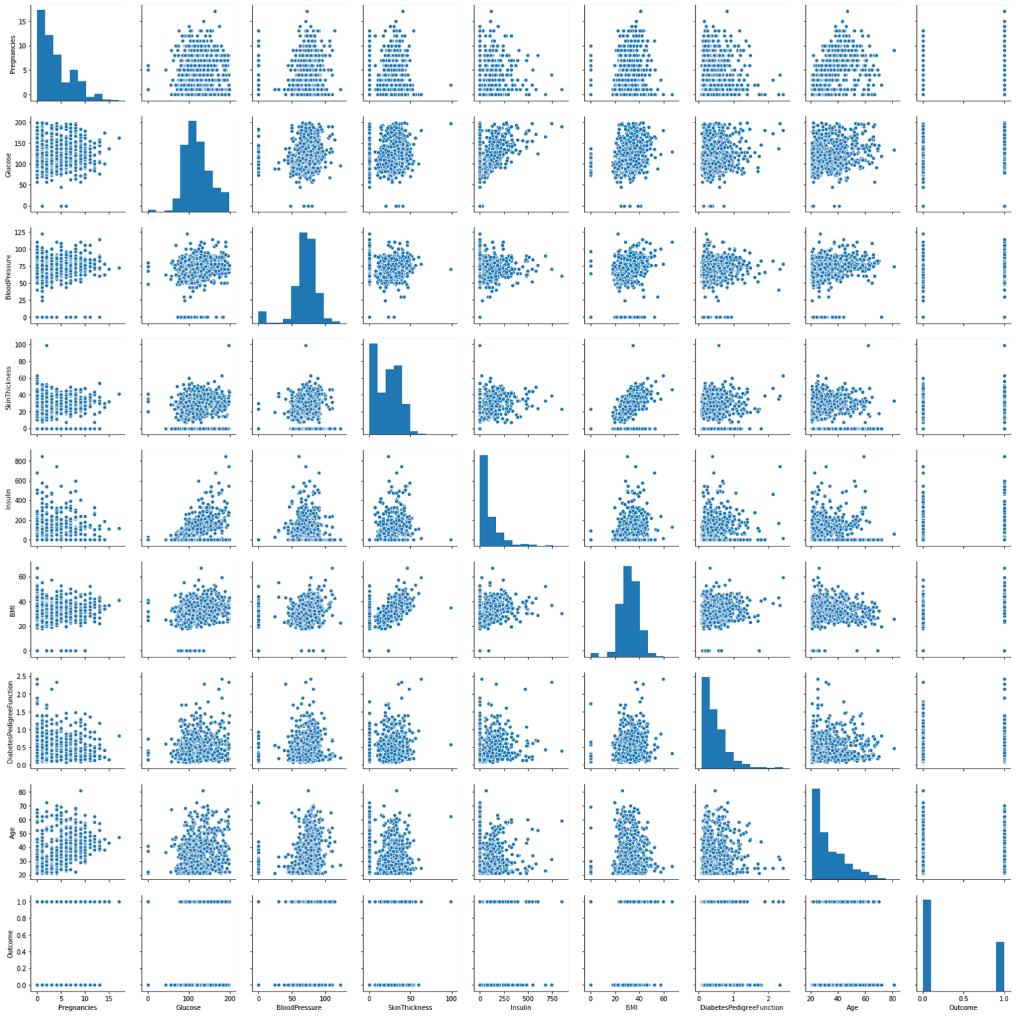

Figura 1. Pairplot de todas las variables involucradas en el set de datos utilizado

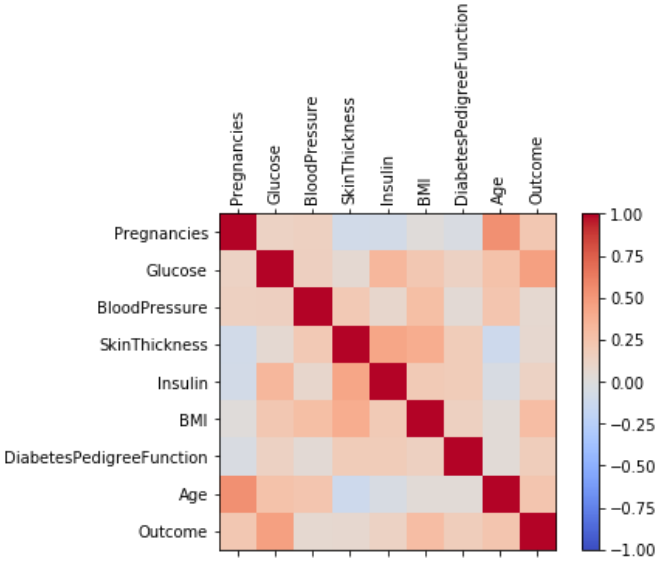

Figura 2. Diagrama de correlación de variables para el análisis de multicolinealidad

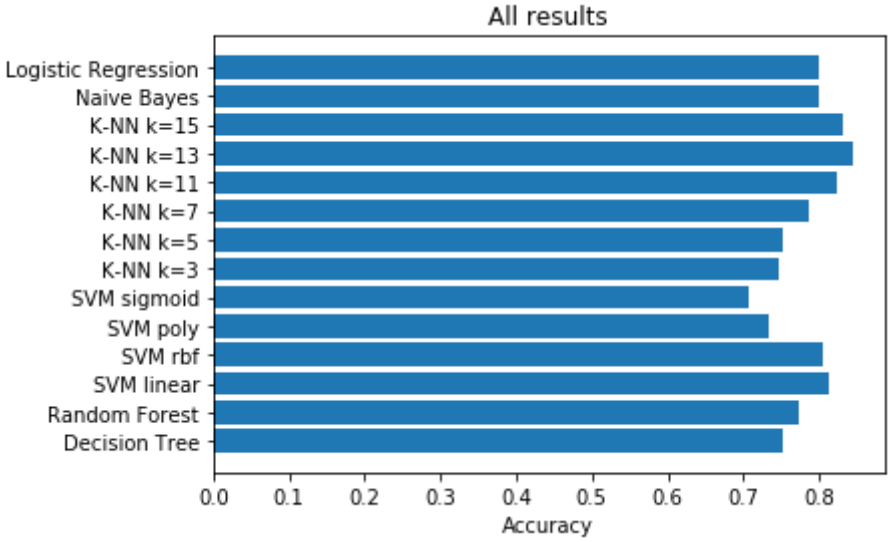

Figura 3. Exactitud de cada algoritmo de clasificación para el set de datos utilizado

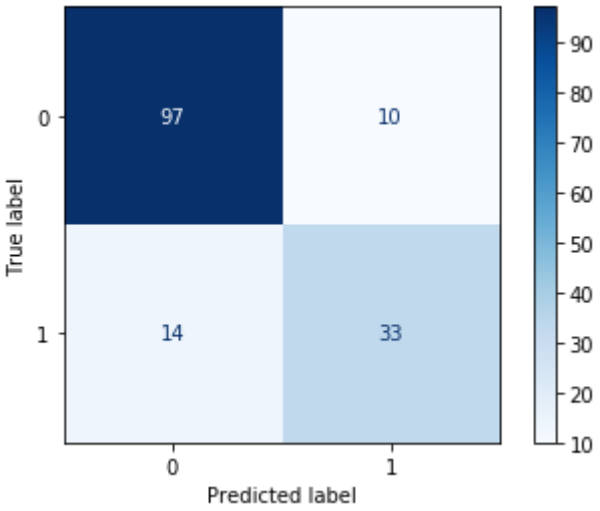

Figura 4. Matriz de confusión que determina el porcentaje de exactitud para el algoritmo KNN con K=13

Algoritmos a utilizar

Se utilizó un módulo llamado ‘’Klas” (hecho por el grupo) que permite la comparación de los distintos algoritmos de clasificación y selecciona el algoritmo que muestre un mayor porcentaje de exactitud. En base a esta comparación se decidieron utilizar los siguientes algoritmos para el set de datos:

KNN

Este algoritmo es un método de clasificación supervisado que permite calcular la probabilidad de que un elemento pertenezca a una clase, basándose en la información proporcionada por elementos asignado a una clase previamente. Así un elemento es asignado a la clase con mayor frecuencia de los ejemplos de entrenamientos que se encuentre más cercano al elemento. La distancia euclidiana es la que se utiliza para calcular las distancias entre cada elemento. Al utilizar un valor más grande de K se promueve la reducción del ruido en la clasificación (Srivastava, 2018).

- KNN con K=15
- KNN con K=13
- KNN con K=11

SVM

Es un algoritmo de clasificación que predice la clase de un elemento en base a clases previamente proporcionadas al modelo en la fase de entrenamiento. La forma en la que predice estas clasificaciones se basa en representar a los puntos de muestra en el espacio, separando las clases a espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los puntos, de las n clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase (Numerentur, 2019).

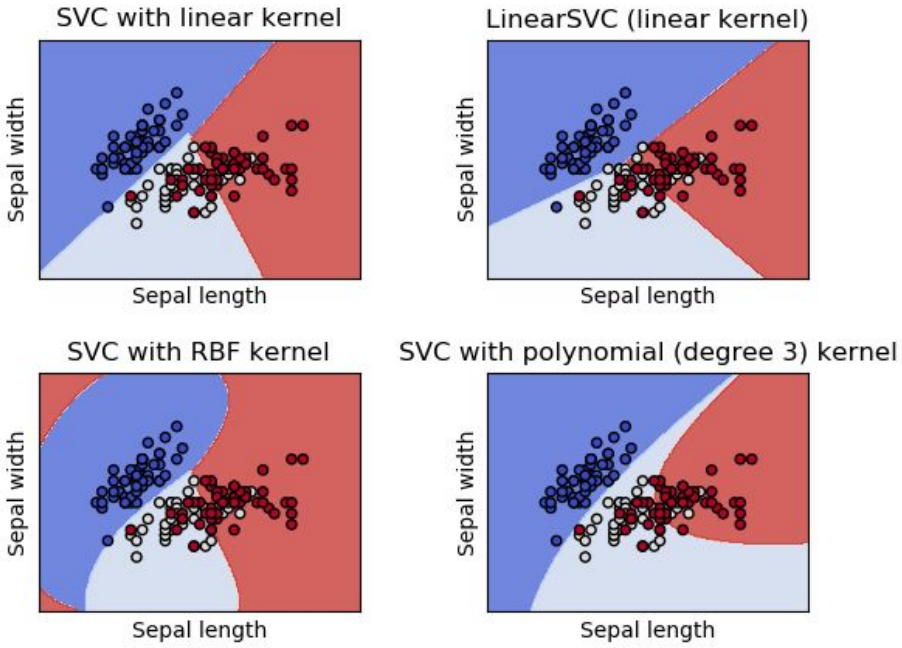

Figura 5. Ilustración de la clasificación de datos utilizando el algoritmo SVM

La diferencia entre el algoritmo SVM lineal y SVM RBF se basa en la forma de la delimitación del hiperplano. Como se observa en la figura 5, el algoritmo lineal utiliza limitaciones lineales para cada clase. Mientras que el SVM RBF utiliza limitaciones radiales para la separación de clases (Sckit-learn, 2020).

- SVM lineal
- SVM RBF

Algoritmo de clustering seleccionado para utilizar

En base a la figura 1 se observó que existen posibles agrupaciones que pueden ser descritas utilizando un algoritmo de clustering. Al obtener estas agrupaciones sería posible describir distintos fenómenos de la enfermedad en relación con las variables involucradas.

- KMeans Clustering: es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

Referencias

Numeretur. (2019). Máquina de soporte vectorial SVM. Extraído de: http://numerentur.org/svm/

Sckit-learn. (2020). Support Vector Machines. Extraído de: https://scikit-learn.org/stable/modules/svm.html

Srivastava. (26 de marzo del 2018). K nearest neighbor. Extraído de: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/