TA-2 Concurrente “K-Means”

**Alumno**: Alexander Gonzalo Urbina Mar

**Profesor**: Carlos Alberto Jara Garcia

**Ciclo**: 2020-01



Tabla de contenido

[Introducción 2](#_Toc42818724)

[Algoritmo K-Means 2](#_Toc42818725)

[Desarrollo del Proyecto 3](#_Toc42818726)

[Pruebas de ejecución 3](#_Toc42818727)

[Referencias 6](#_Toc42818728)

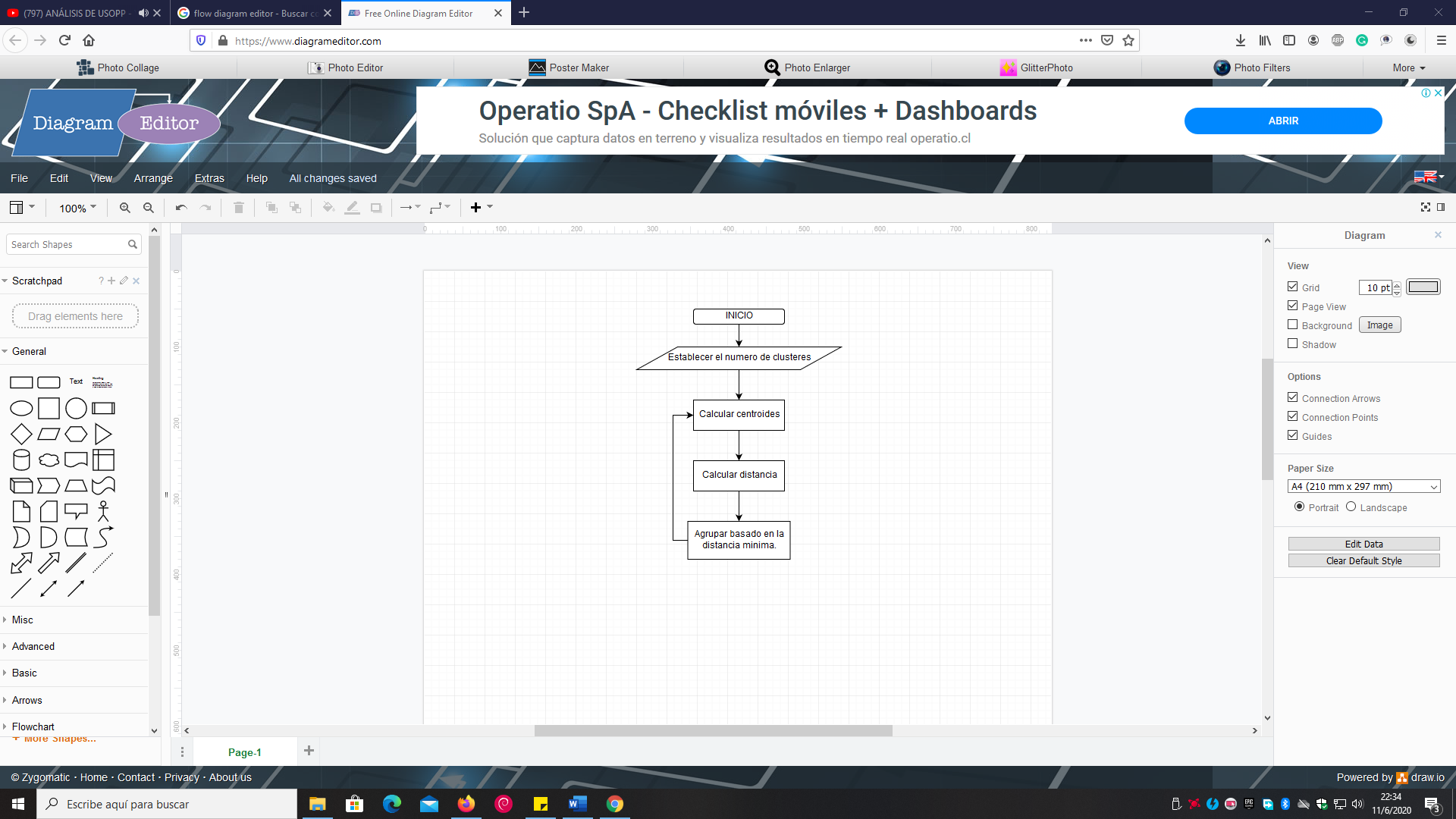
# Introducción

Los algoritmos de agrupamiento, clustering en inglés, son algoritmos que mediante un criterio de agrupación logran clasificar varios datos en clústeres. Estos pueden ser utiles en medicina para claficiar pacientes por sus síntomas, biología para clasificar especie o inclusive publicidad. En este documento veremos una implementación de K – means(medias) el cual es un método de agrupamiento común.

# Algoritmo K-Means

El algoritmo estándar de k means es el siguiente:

1. Seleccionar el número de clústeres que se quiere identificar en el set de datos los cuales suelen ser representados como puntos. Estos vendrían a ser k clústeres.
2. Seleccionar aleatoriamente k puntos de nuestro set de datos que se volverán los centroides iniciales.
3. Comparar las distancias de todos los datos con los k centroides. Esta es una distancia euclidiana.
4. Asignar a los puntos el clúster del centroide con la menor distancia hacia este.
5. Calcular la media de los puntos de cada clúster, este nuevo punto se volverá el centroide actual.
6. Repetir pasos anteriores hasta que haya pocos cambios o sean aceptables.



En la implementación que se realizo en este proyecto el proceso del calculo de la distancia de los puntos a los centroides se hizo paralelo con el lenguaje go para mejorar el rendimiento del cálculo.

# Desarrollo del Proyecto

Se uso el lenguage de programación go para hacer un Web API que será el backend demuestra aplicación web. También se usara el microframework gin para poder mandar y recibir las peticiones HTTP y se usara gorm que es un ORM ligero. Se necesitará tener una versión de gcc en Windows.

Para el frontend que consumirá el backend se utilizó html, css y typescript junto con el framework Angular para su desarrollo.

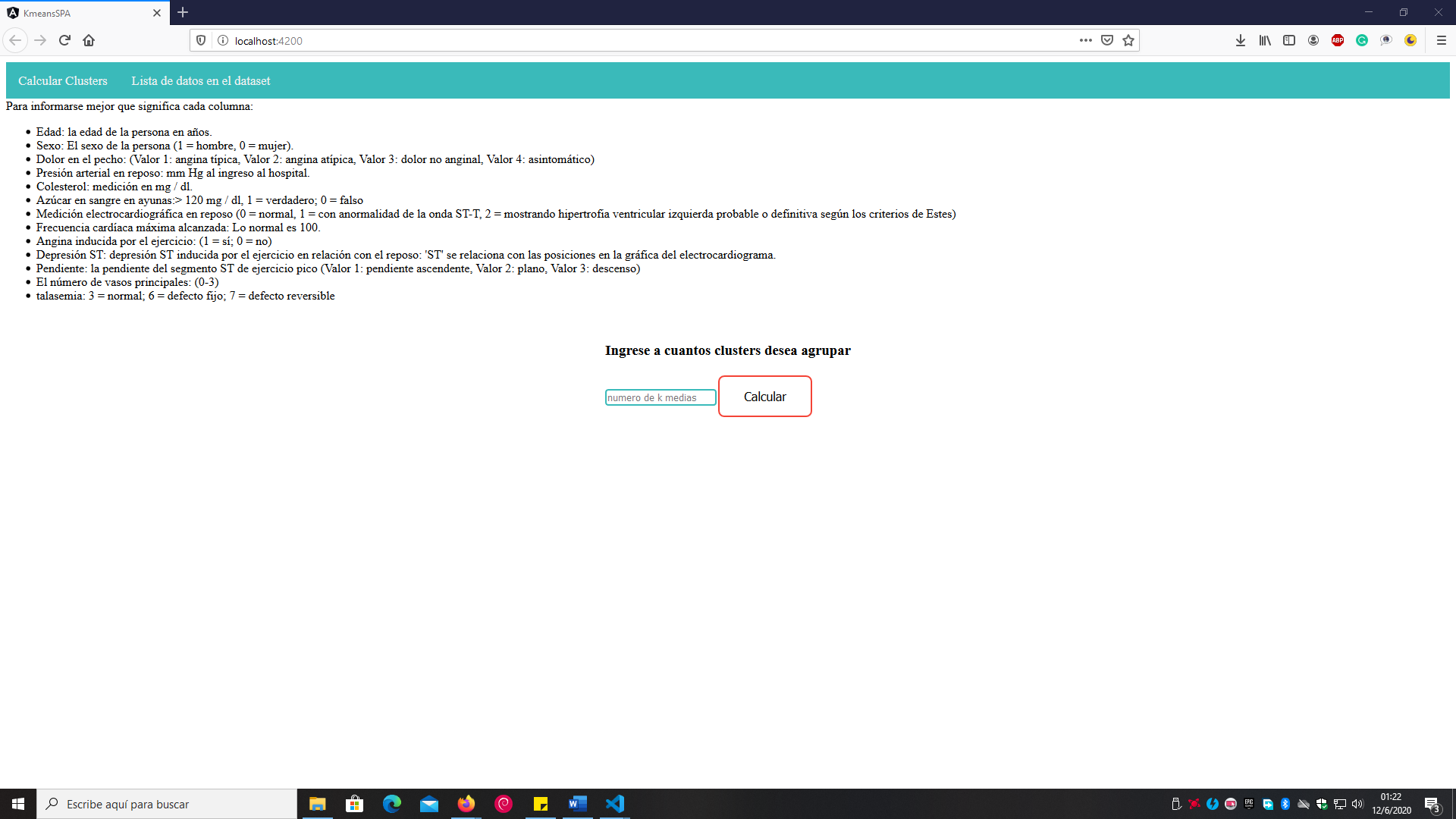
El dataset es una tabla de 13 columnas que tiene diversos datos relacionados a las varias personas diagnosticadas con una enfermedad cardiaca. Originalmente era un archivo csv, pero fue convertido a una tabla de una base de datos sqlite.

Los colmunas del dataset son:

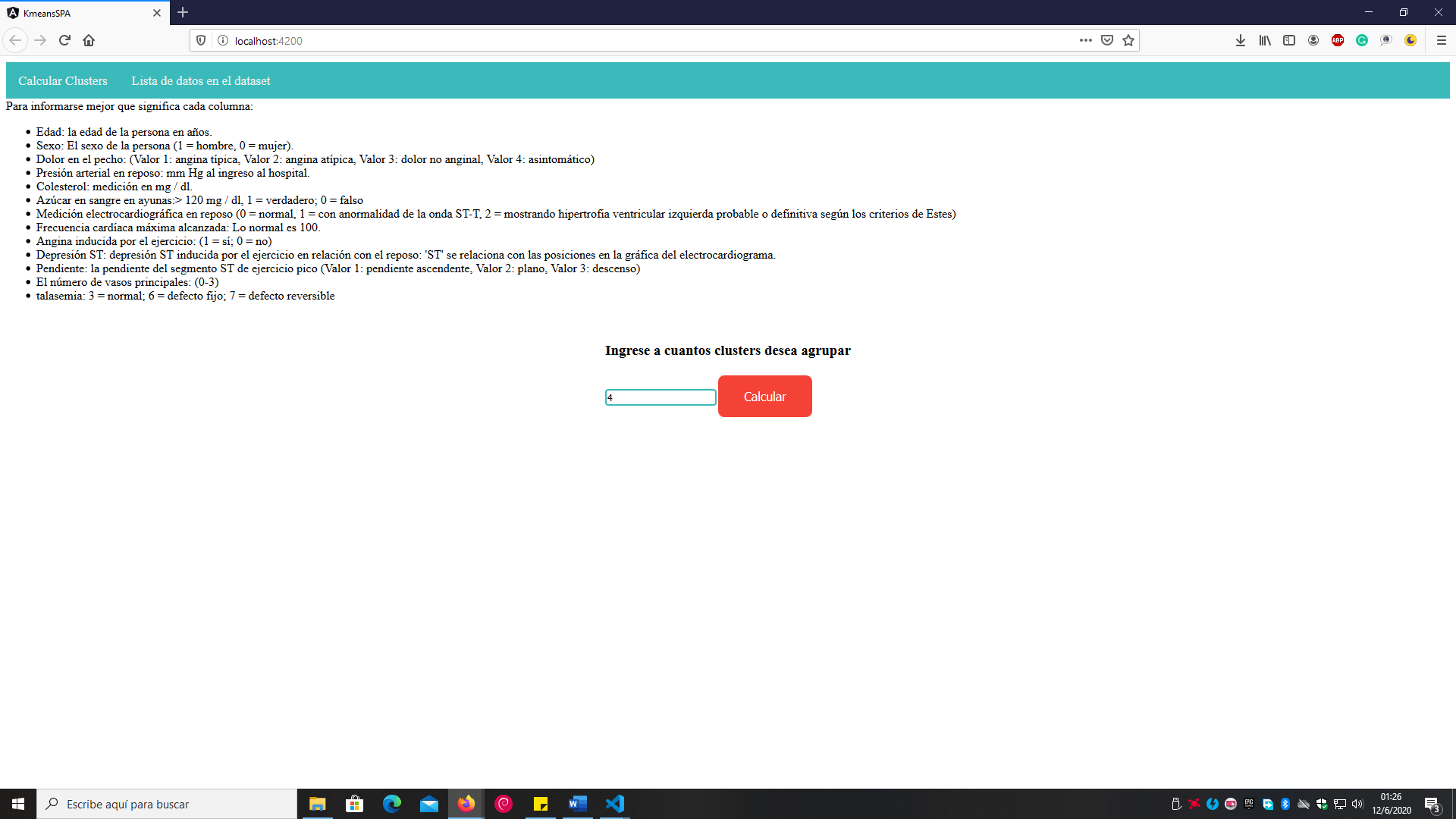
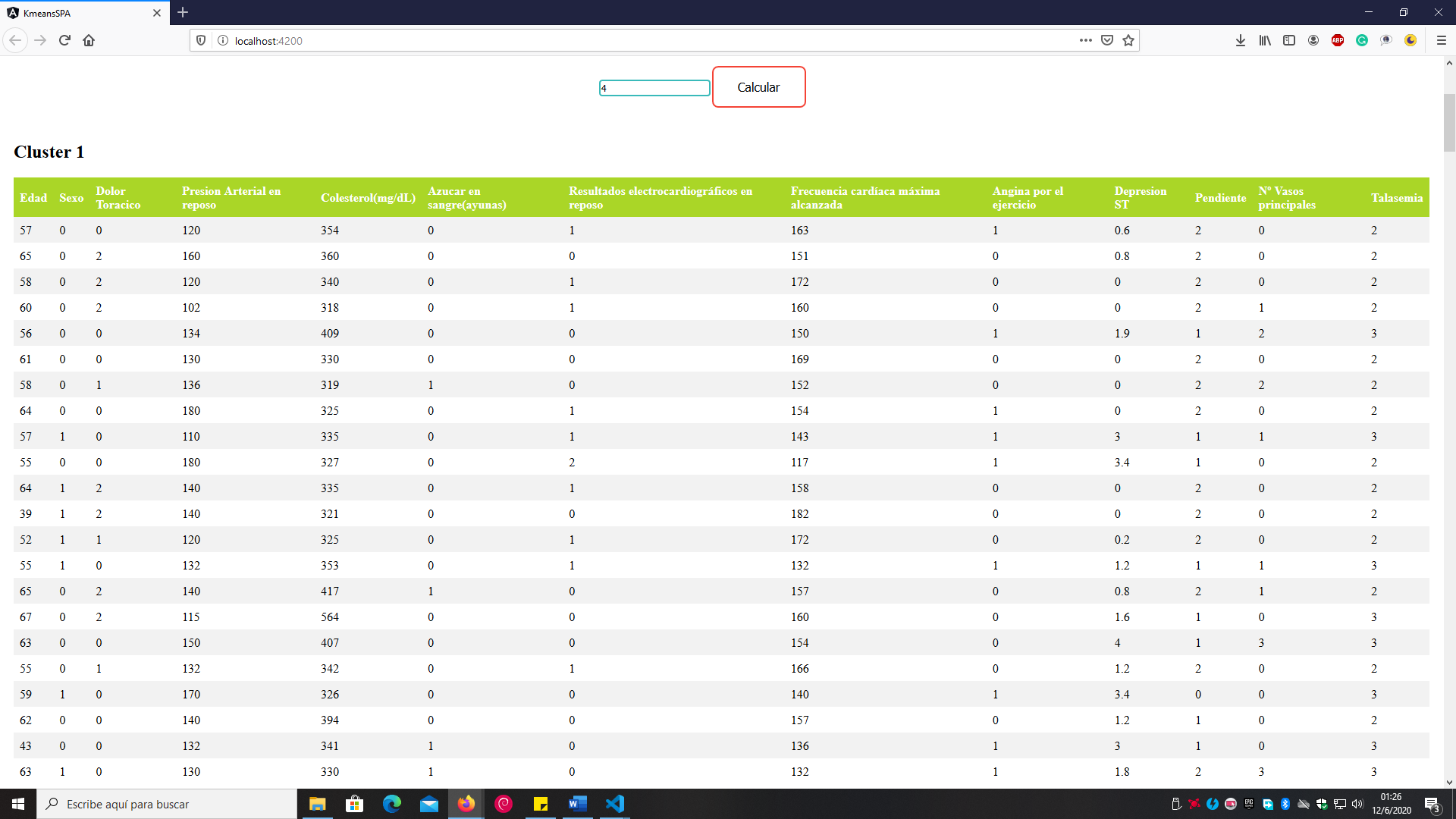
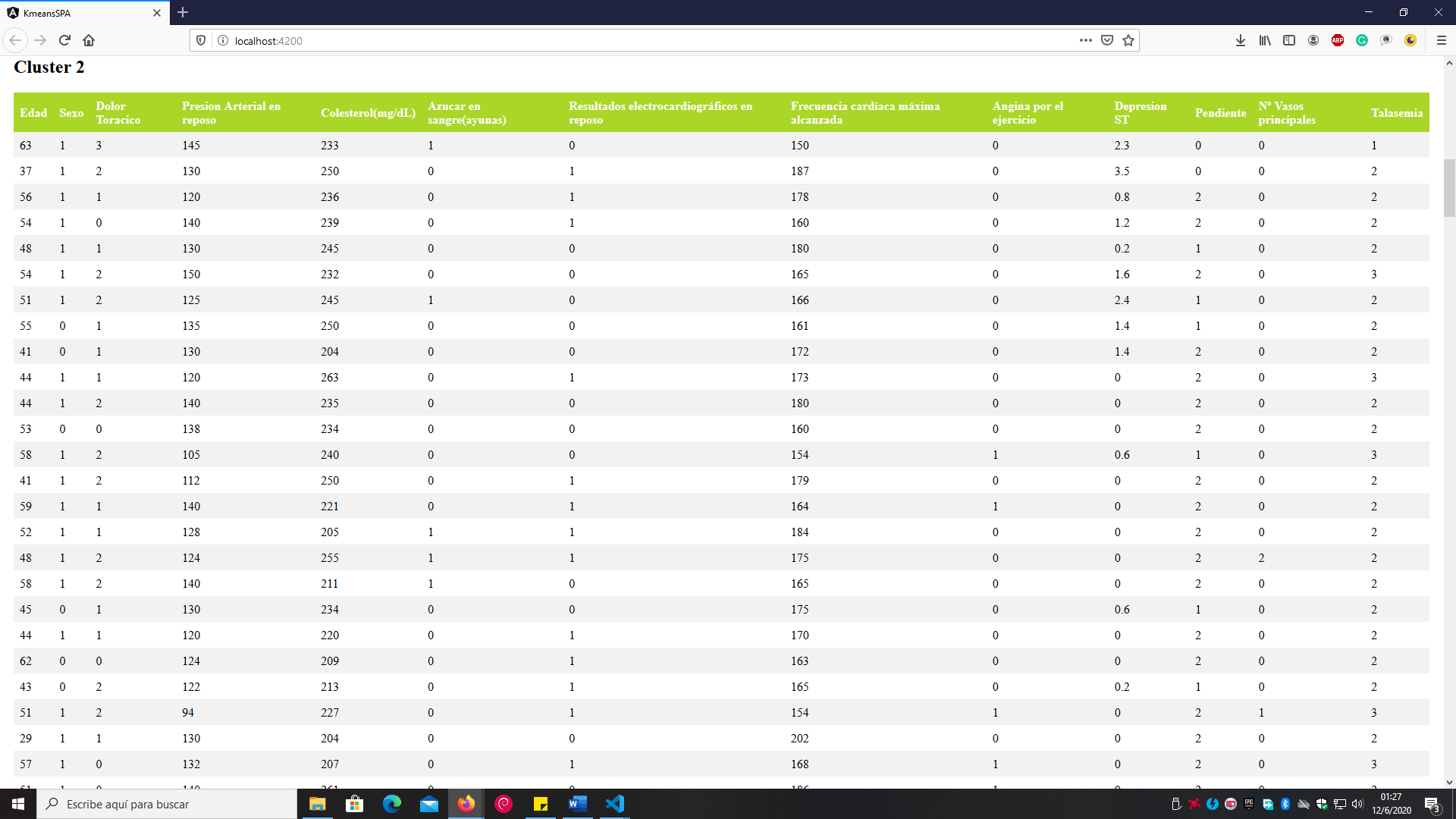
1. Edad: la edad de la persona en años.
2. Sexo: El sexo de la persona (1 = hombre, 0 = mujer).
3. Dolor en el pecho: (Valor 1: angina típica, Valor 2: angina atípica, Valor 3: dolor no anginal, Valor 4: asintomático)
4. Presión arterial en reposo: mm Hg al ingreso al hospital.
5. Colesterol: medición en mg / dl.
6. Azúcar en sangre en ayunas:> 120 mg / dl, 1 = verdadero; 0 = falso
7. Medición electrocardiográfica en reposo (0 = normal, 1 = con anormalidad de la onda ST-T, 2 = mostrando hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)
8. Frecuencia cardíaca máxima alcanzada: Lo normal es 100.
9. Angina inducida por el ejercicio: (1 = sí; 0 = no)
10. Depresión ST: depresión ST inducida por el ejercicio en relación con el reposo: 'ST' se relaciona con las posiciones en la gráfica del electrocardiograma.
11. Pendiente: la pendiente del segmento ST de ejercicio pico (Valor 1: pendiente ascendente, Valor 2: plano, Valor 3: descenso)
12. El número de vasos principales: (0-3)
13. talasemia: 3 = normal; 6 = defecto fijo; 7 = defecto reversible

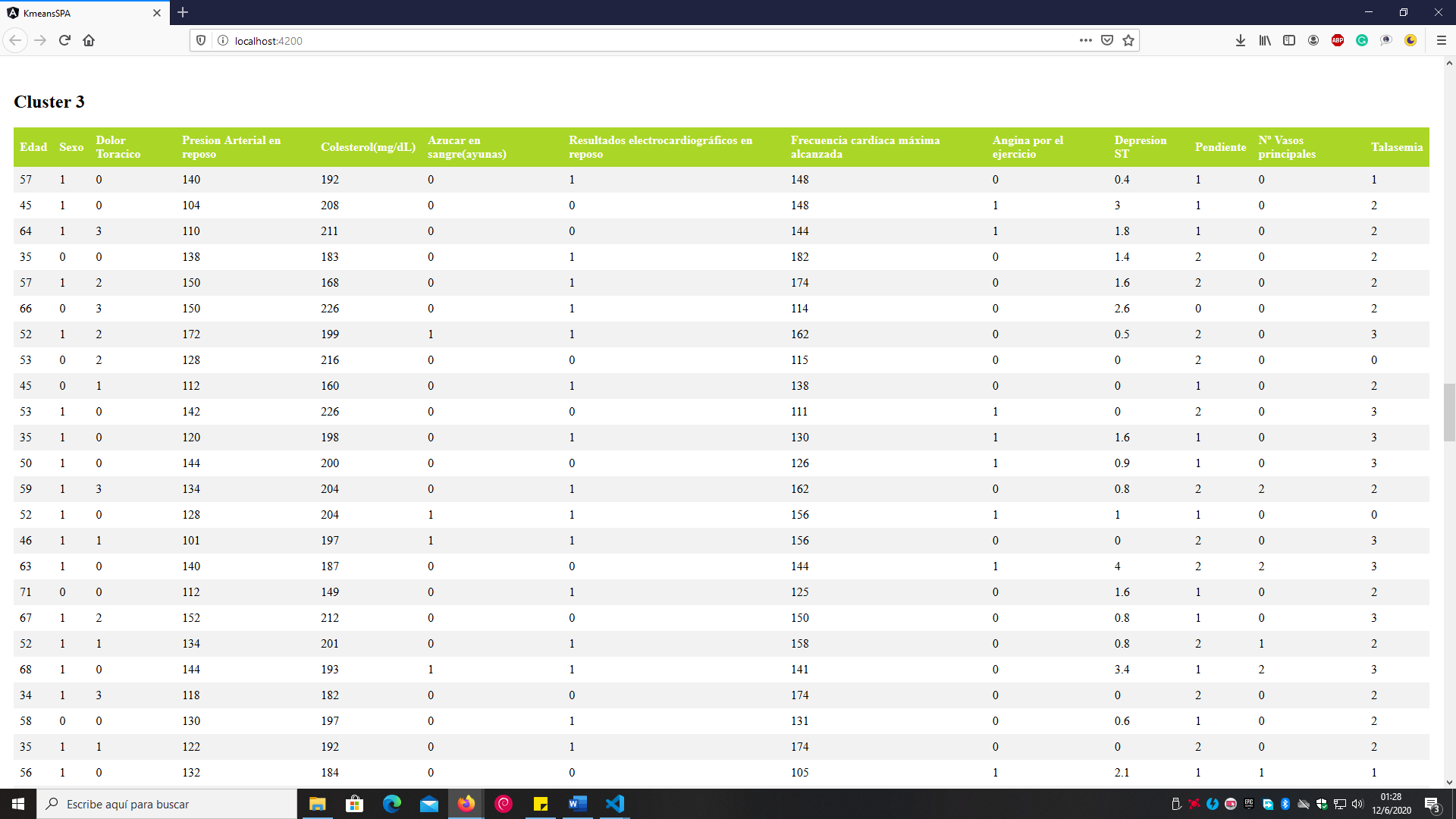
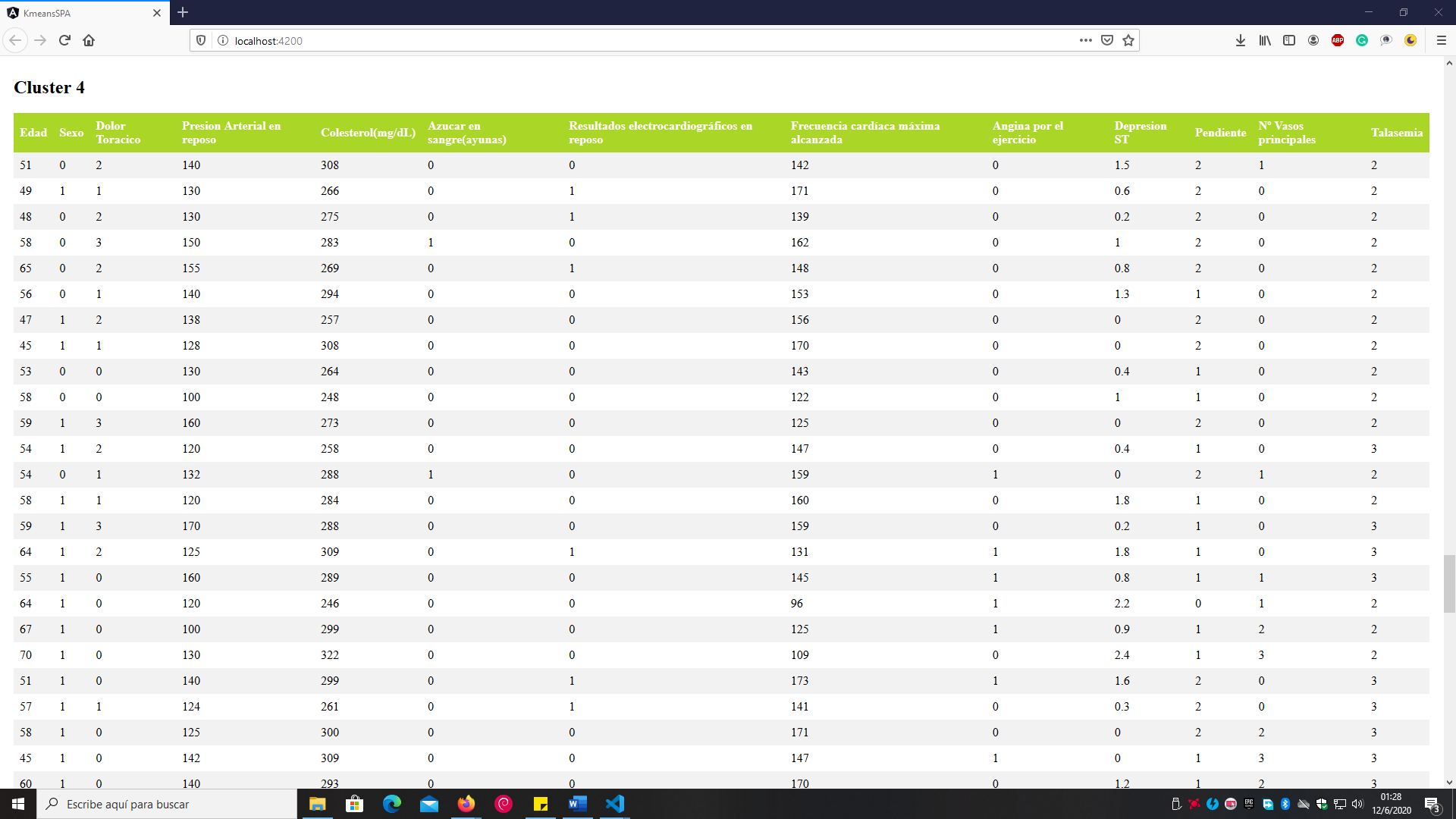
# Pruebas de ejecución

La interfaz por defecto nos muestra un pequeño listado de información que nos indica que es cada valor. Nos muestra un input donde colocar cuantos en cuantos clústeres se desea agrupar.

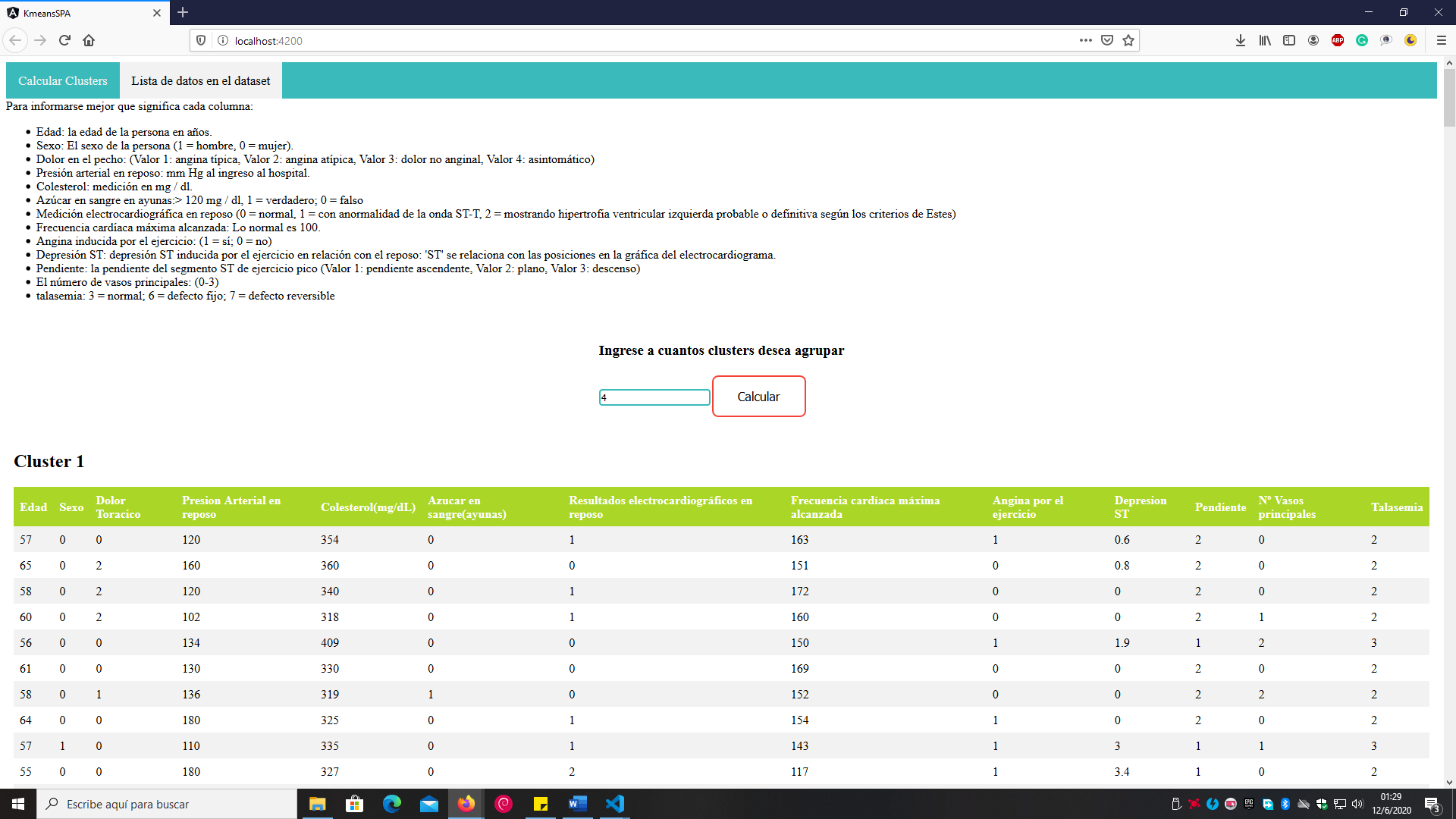


Seleccionando uno y presionando calcular nos mostrara los agrupamientos.





También nos permite ver el dataset sin agrupar si presionamos Lista de datos en el dataset.





# Referencias

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1 de Julio de 1988). *Heart Disease Data Set .* Obtenido de UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Schott, M. (23 de Abril de 2019). *K-Means Clustering Algorithm for Machine Learning.* Obtenido de Medium: https://medium.com/capital-one-tech/k-means-clustering-algorithm-for-machine-learning-d1d7dc5de882

Sharma, N. (Febrero de 21 de 2020). *Understanding the Mathematics behind K-Means Clustering.* Obtenido de Heartbeat: https://heartbeat.fritz.ai/understanding-the-mathematics-behind-k-means-clustering-40e1d55e2f4c