Lab01: k-Nearest Neighbors

Por:
Ian Gabriel Cañas Fernández, 1092228

.

Profesor: Juan S. Pérez R.,
Asignatura: INL367L, Secc 01

Resumen:

En este laboratorio estaremos implementando el algoritmo de k Nearest Neigbors (kNN), algoritmo que consiste en la clasificación automática de varios puestos de muestra respecto a los puntos conocidos más cercanos, para ello se ha iniciado haciendo un proceso analítico en el que se simulan los cálculos llevados a cabo y se concluye con la prueba de este mediante un set de 70 000 datos de entrenamiento y siendo probado con 20 000 datos de prueba.

Ejercicios previos:

P1.1 Predicción analítica de género.

En la Ilustración 1, Ilustración 2, Ilustración 3 y Ilustración 4, se presenta una síntesis de lo que se lleva a cabo en el algoritmo y los resultados obtenidos, en tal caso se ha optado por utilizar una distancia euclideana, pues estamos asumiendo un peso equitativo entre todos los datos.

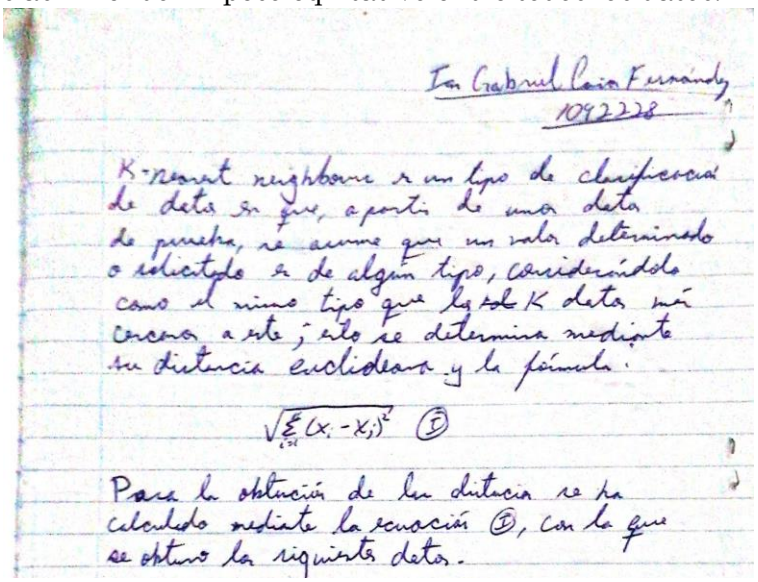


Ilustración 1. Proceso analítico llevado a cabo.

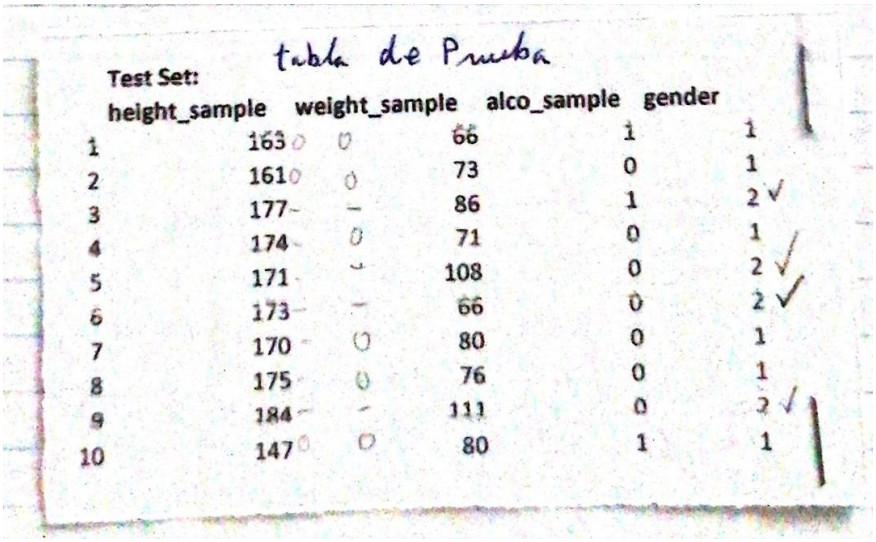
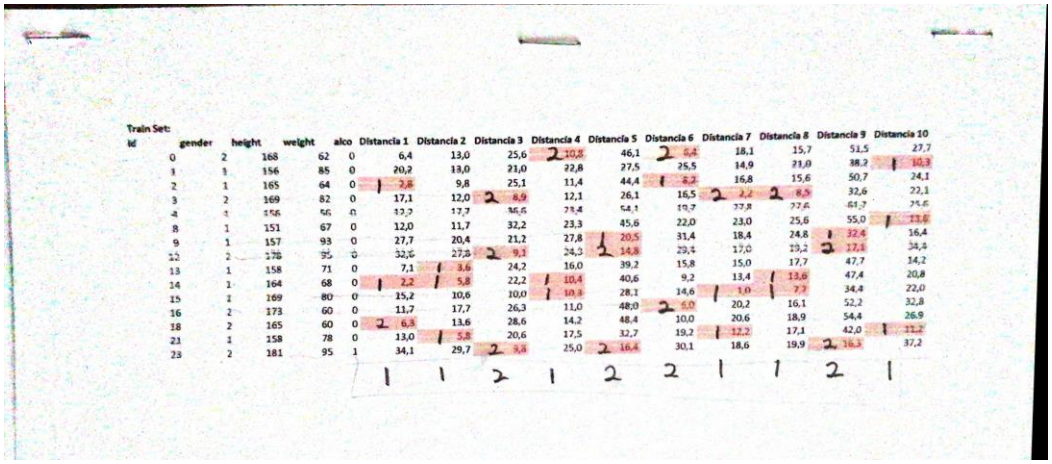


Ilustración 2. Tabla de datos de prueba.



Train Set:

id	gender	height	weight	alco	Distancia 1	Distancia 2	Distancia 3	Distancia 4	Distancia 5	Distancia 6	Distancia 7	Distancia 8	Distancia 9	Distancia 10
0	2	168	62	0	6,4	13,0	25,6	2 10,8	46,1	2 6,4	18,1	15,7	51,5	27,7
1	1	156	85	0	20,2	13,0	21,0	22,8	27,5	25,5	14,9	21,0	38,2	1 10,3
2	1	165	64	0	1 2,8	9,8	25,1	11,4	44,4	1 8,2	16,8	15,6	50,7	24,1
3	2	169	82	0	17,1	12,0	2 8,9	12,1	26,1	16,5	2 2,2	2 8,5	32,6	22,1
4	1	156	56	0	13,9	17,7	36,6	23,4	54,1	19,7	27,8	22,6	61,7	35,6
8	1	151	67	0	12,0	11,7	32,2	23,3	45,6	22,0	23,0	25,6	55,0	1 13,6
9	1	157	93	0	27,7	20,4	21,2	27,8	1 20,5	31,4	18,4	24,8	1 32,4	16,4
22	2	178	95	0	32,6	27,8	2 9,1	24,3	2 14,8	29,4	17,0	19,2	2 17,1	34,4
13	1	158	71	0	7,1	1 3,6	24,2	16,0	39,2	15,8	15,0	17,7	47,7	34,2
14	1	164	68	0	1 2,2	1 5,8	22,2	1 10,4	40,6	9,2	13,4	1 13,6	47,4	20,8
15	1	169	80	0	15,2	10,6	10,0	1 10,3	28,1	14,6	1 1,0	1 7,7	34,4	22,0
16	2	173	60	0	11,7	17,7	26,3	11,0	48,0	2 6,0	20,2	16,1	52,2	32,8
18	2	165	60	0	2 6,3	13,6	28,6	14,2	48,4	10,0	20,6	18,9	54,4	26,9
21	1	158	78	0	13,0	1 5,8	20,6	17,5	32,7	19,2	1 12,2	17,1	42,0	1 11,2
23	2	181	95	1	34,1	29,7	2 9,8	25,0	2 16,4	30,1	18,6	19,9	2 16,3	37,2
					1	1	2	1	2	2	1	1	2	1

Ilustración 3. Comparación con distancias menores.

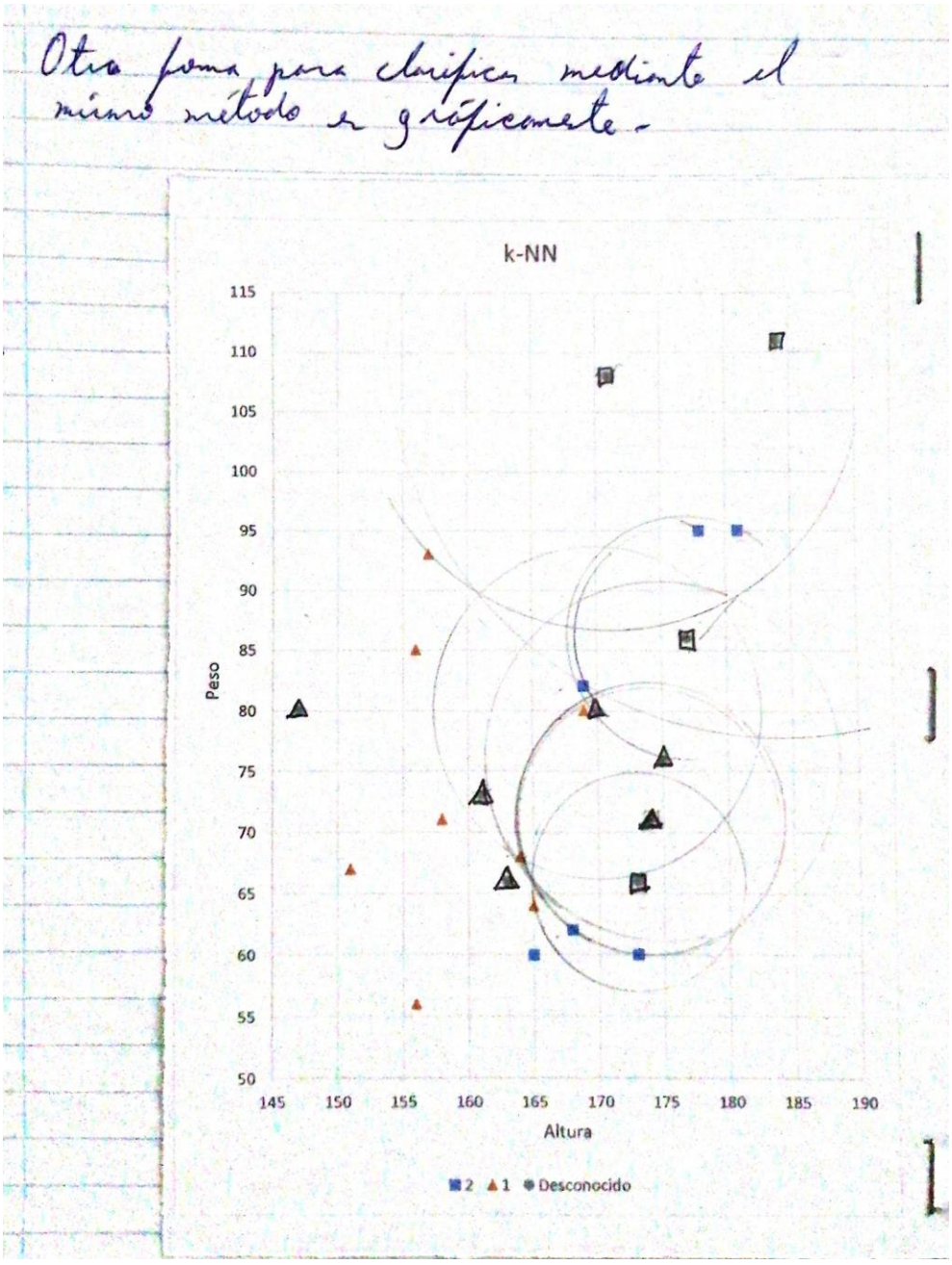


Ilustración 4. Resultados analíticos.

E1.1 Implementación en Matlab de kNN

Para el proceso computarizado de la implementación de kNN, se ha empezado importando los datos en dos tablas, luego, se ha identificado los datos para próximamente ser graficados y habiéndose solicitado la cantidad de vecinos con los que se solicita trabajar, véase la Ilustración 5.

```
clear, clc

train_set = readtable("Train set.xlsx"); %readtable("Train2.xlsx");
test_set = readtable("Test set.xlsx"); %readtable("Test2.xlsx");

g1 = train_set(train_set.gender==1, :);
g2 = train_set(train_set.gender==2, :);

plot(g1.height, g1.weight, '.b')
hold on
plot(g2.height, g2.weight, '.y')

k = input('Introdutrazca la cantidad de vecinos: ');
```

Ilustración 5. Clasificación de datos.

A continuación, se obtuvo la distancia de los puntos de prueba respecto al modelo, desde donde se comparó los k valores más pequeños y se obtuvo sus respectivos índices, mediante los que se podrá accesar y comparar con los resultados obtenidos, en la Ilustración 6 se puede observar dicho extracto de código.

```
nTrainData = length(train_set.height) %size(train_set);

for i = 1:length(test_set.height)
    %i

    Rep = repmat([test_set.height(i), test_set.weight(i)], nTrainData, 1);
    d = ((Rep - [train_set.height(:) train_set.weight(:)]).^2);
    d = sqrt(d(:,1)+d(:,2));

    [dis pos] = sort(d,'ascend');
    kNN=pos(1:k);
    kND=dis(1:k);
```

Ilustración 6. Obtención de distancias.

Finalmente se analizan los datos obtenidos por saber la etiqueta que le tocaría a cada valor y finalmente se grafican, este extracto de código se presenta en la Ilustración 7:

```
    c1 = 0;
    c2 = 0;
    for m = 1:k
        if (train_set.gender(pos(m))==1)
            c1 = c1+1;
        elseif (train_set.gender(pos(m))==2)
            c2 = c2+1;
        end
    end

    if c1>c2
        test_set.gender(i)=1;
    elseif c2>c1
        test_set.gender(i)=2;
    end

end

p1 = test_set(test_set.gender==1, :);
p2 = test_set(test_set.gender==2, :);

plot(p1.height, p1.weight, 'c^')
plot(p2.height, p2.weight, 'rs')
```

Ilustración 7. Análisis de datos.

Antes de la toma de resultados, se le ha dado mayor fiabilidad al algoritmo mediante la comprobación con el modelo ideado para los cálculos analíticos, este se presenta en la .

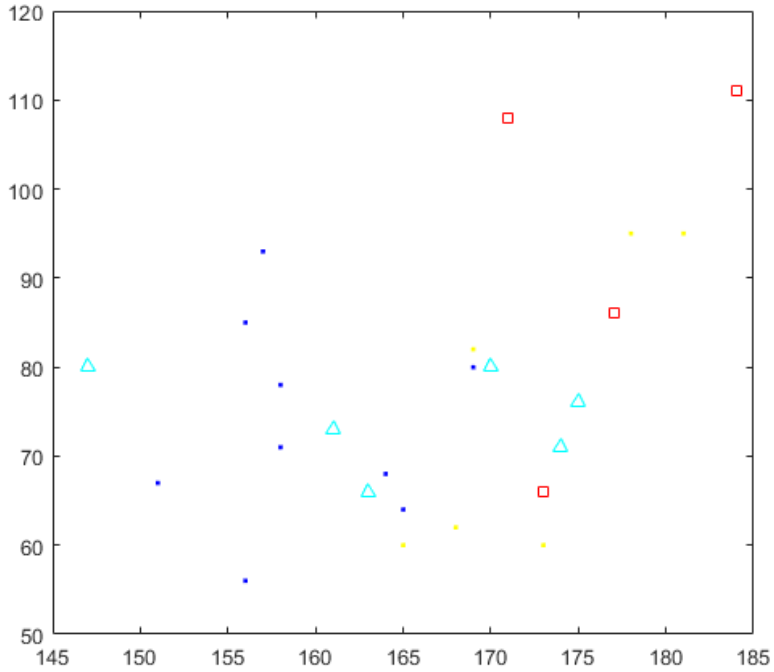


Ilustración 8. Prueba de fiabilidad de código a implementar.

R1.1 Resultados:

Antes de presentar los datos obtenidos para cada valor de k, se presenta en la Ilustración 9 el dataset de entrenamiento.

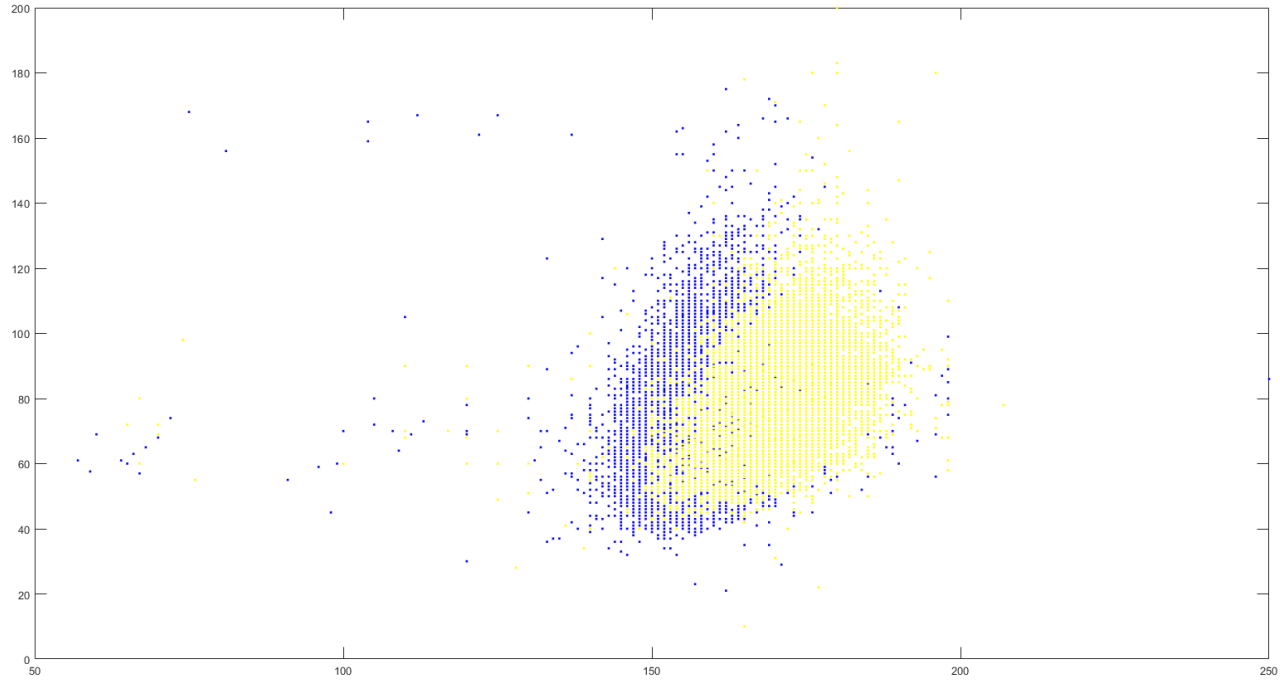


Ilustración 9. Datos de entrenamiento.

A continuación, se presentan todos los gráficos obtenidos. Empezando con $k = 1$ (ver Ilustración 10).

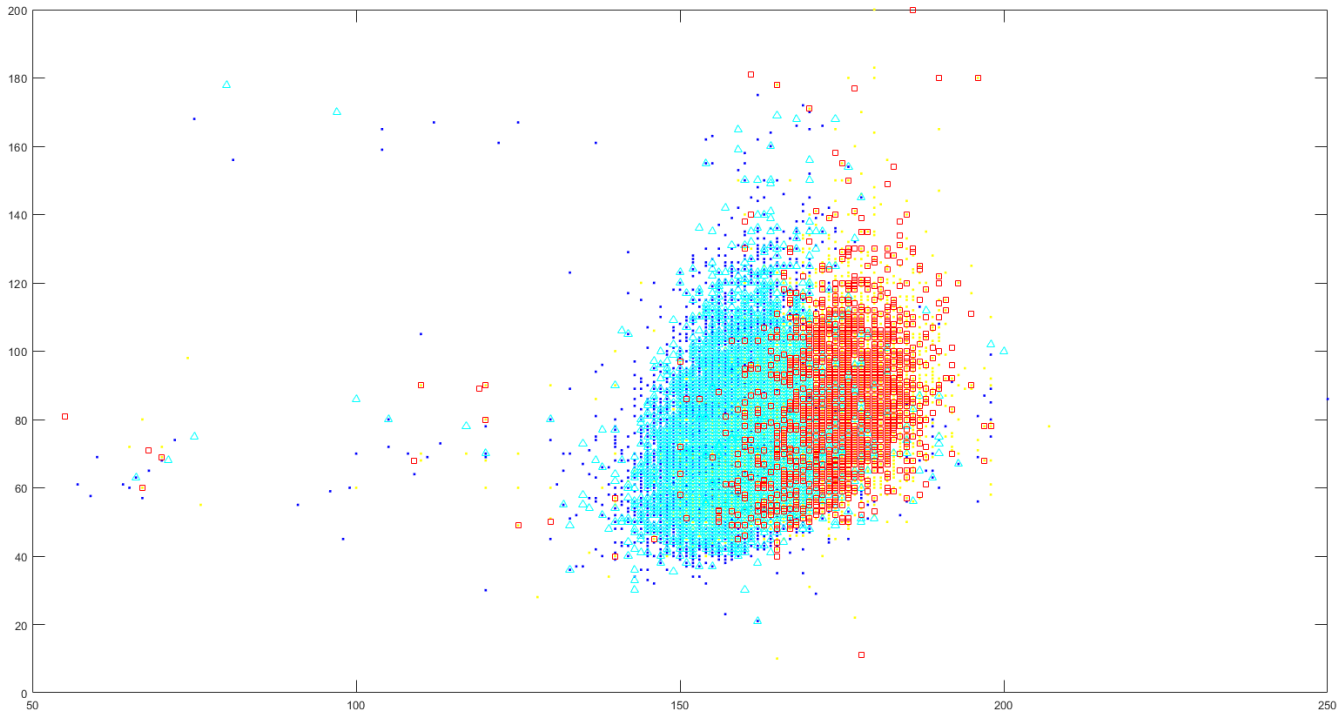


Ilustración 10. resultados para k=1.

Para $k = 3$ (ver Ilustración 11).

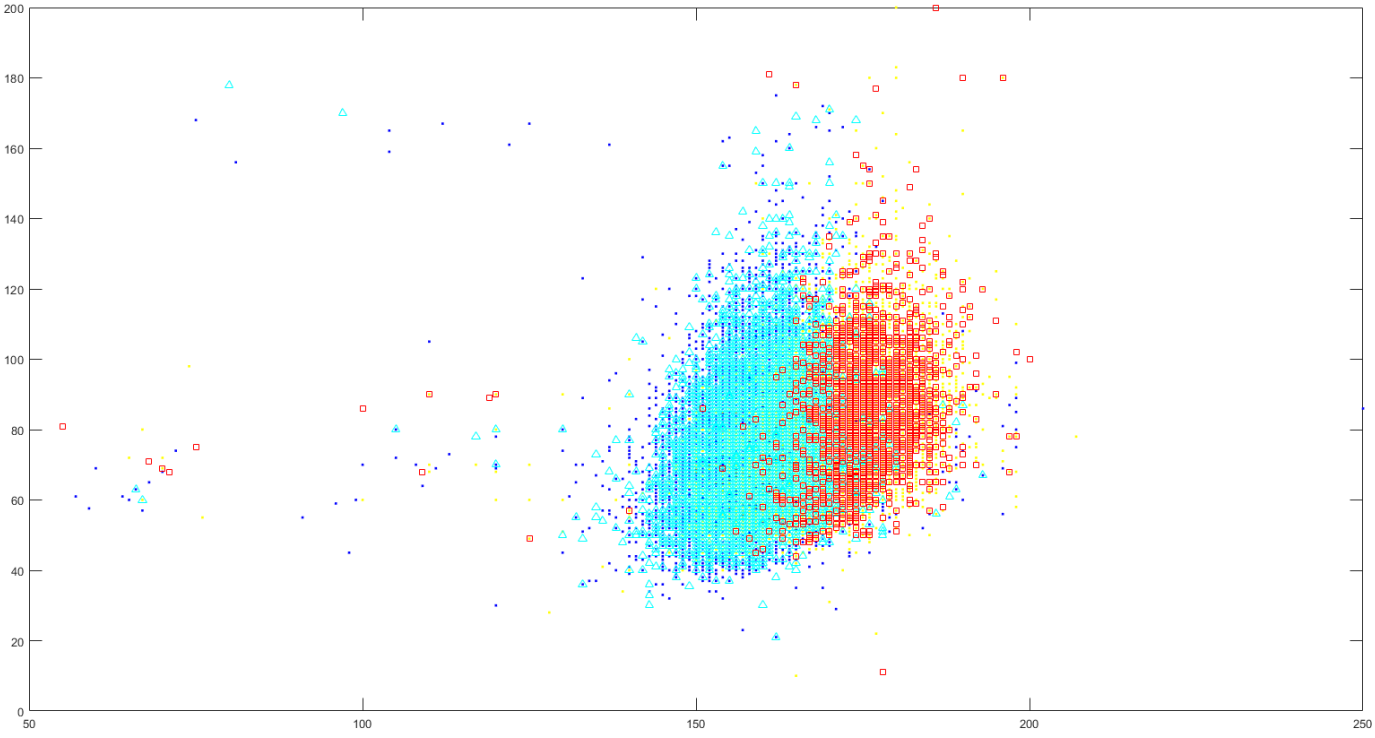


Ilustración 11. Resultados para k=3.

Para $k = 5$ (ver Ilustración 12).

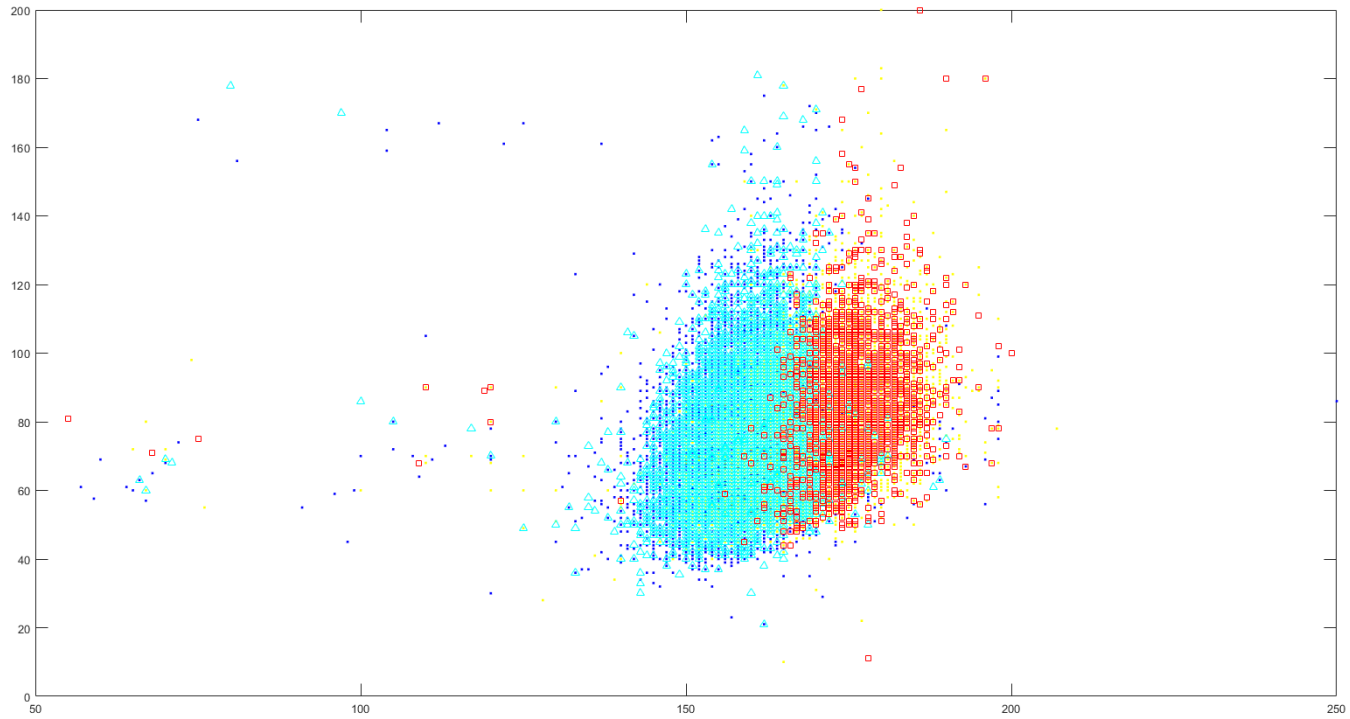


Ilustración 12. Resultados para k=5.

Para $k = 9$ (ver Ilustración 13).

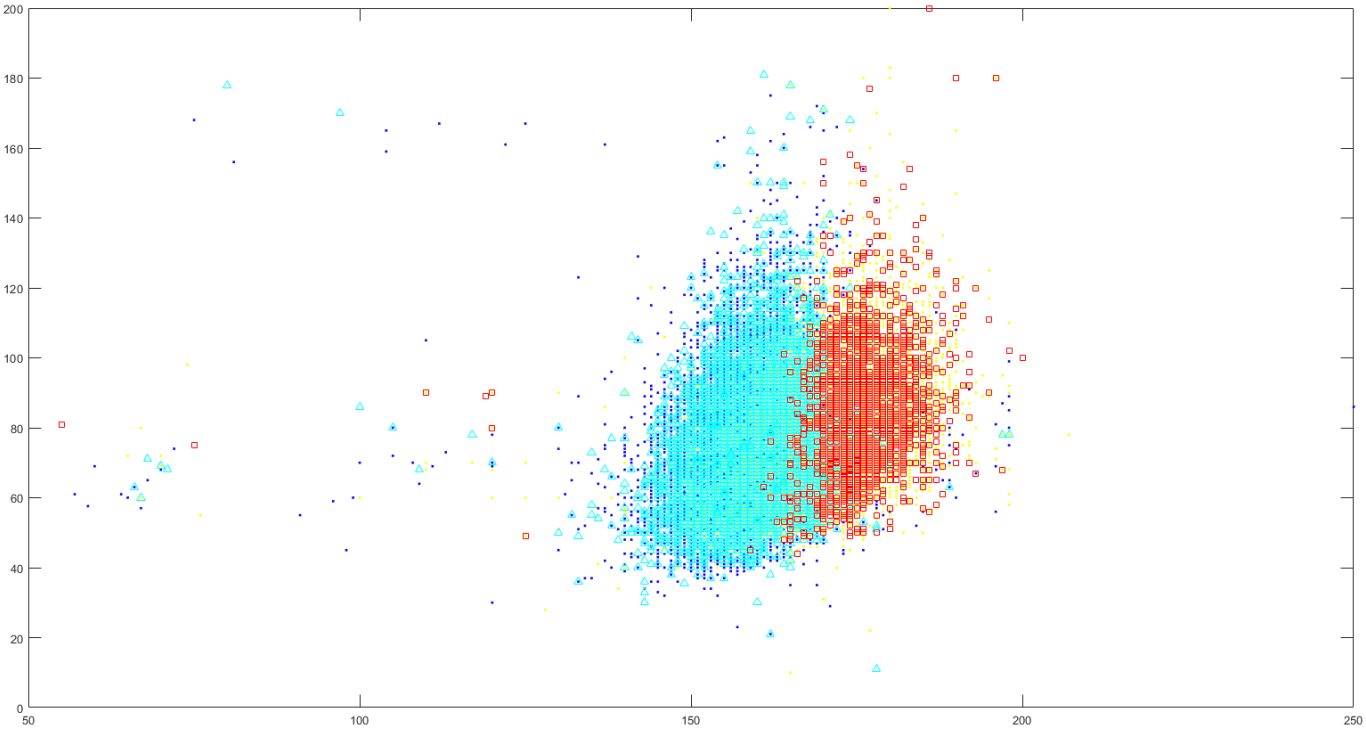


Ilustración 13. Resultados para k=9.

Para $k = 15$ (ver Ilustración 14).

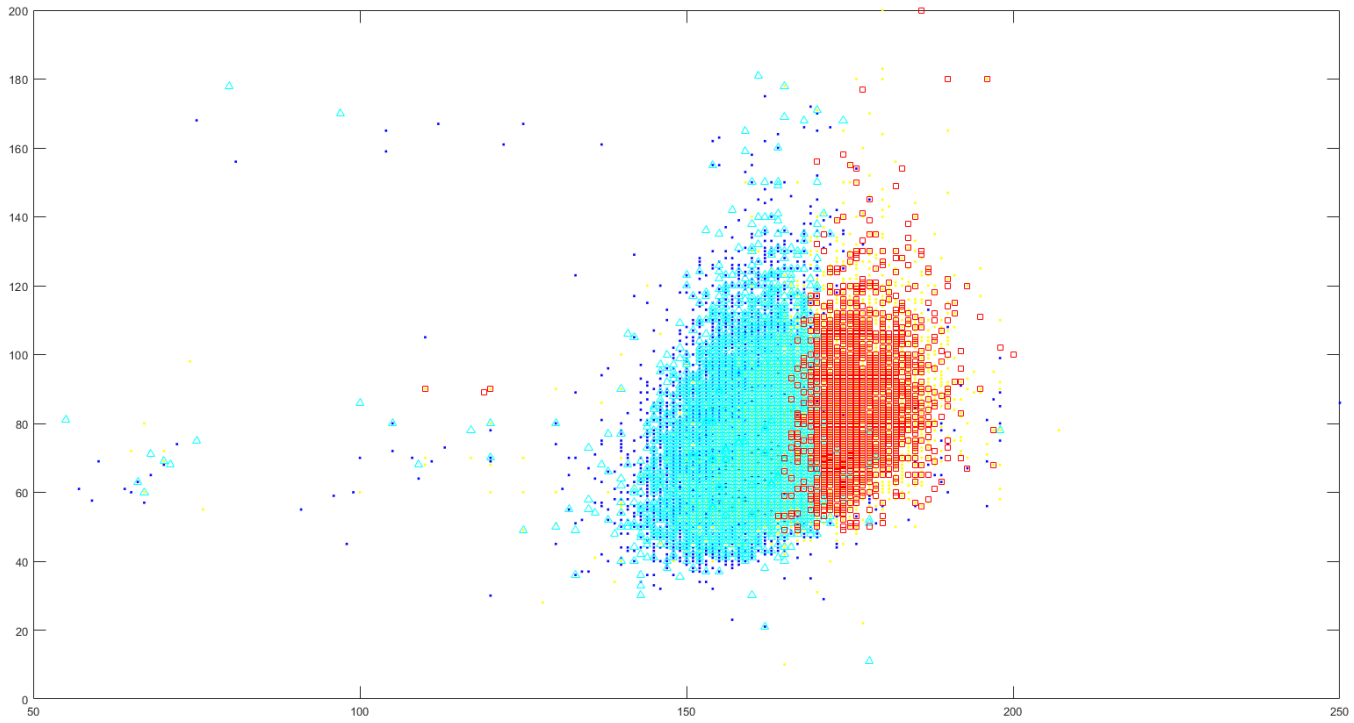


Ilustración 14. Resultados para k=15.

A.1.1 Análisis:

Se ha podido observar el comportamiento del algoritmo de los k vecinos más cercanos, proceso de clasificación mediante el cual se hace una comparación entre los datos a ser analizados y su posición en el plano, tomando en consideración los datos ya reconocidos y clasificados que se encuentren en su entorno.

El algoritmo implementado ha sido probado mediante varios valores de k, de donde se interpreta que, para valores de k mayores, tiende a tenerse mayor nitidez en los cúmulos de un valor determinado. Confirmando este algoritmo con la clasificación de los datos calculados analíticamente para tener garantía de su correcto funcionamiento. Se ha seleccionado para todo caso una distancia euclideana.