“Tarea 2”

“Minería de datos”

“Naïve Bayes”

Issam Silverio Jimenez Ortega

Ruddy Miranda Marez

Víctor Hugo García Méndez

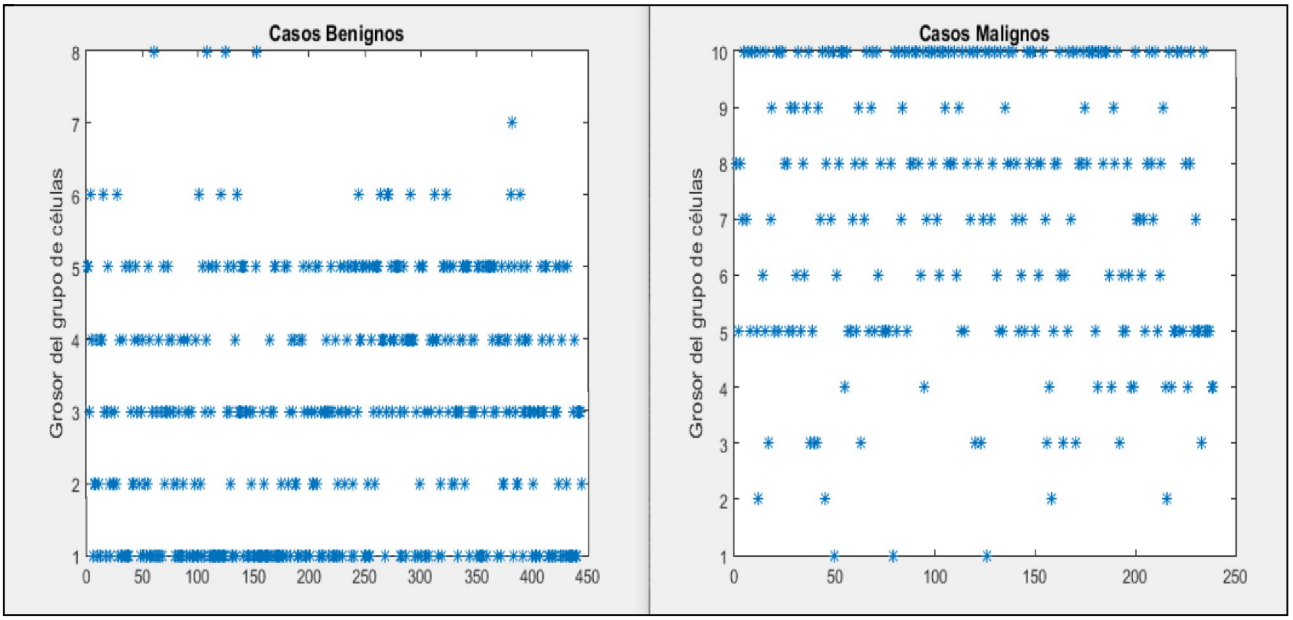
Índice

Análisis de las variables.

Grosor del grupo de células.

Rango 1-10

	Benignos	Malignos
Media	2.9640	7.1883
Desv.Est	1.6727	2.4379

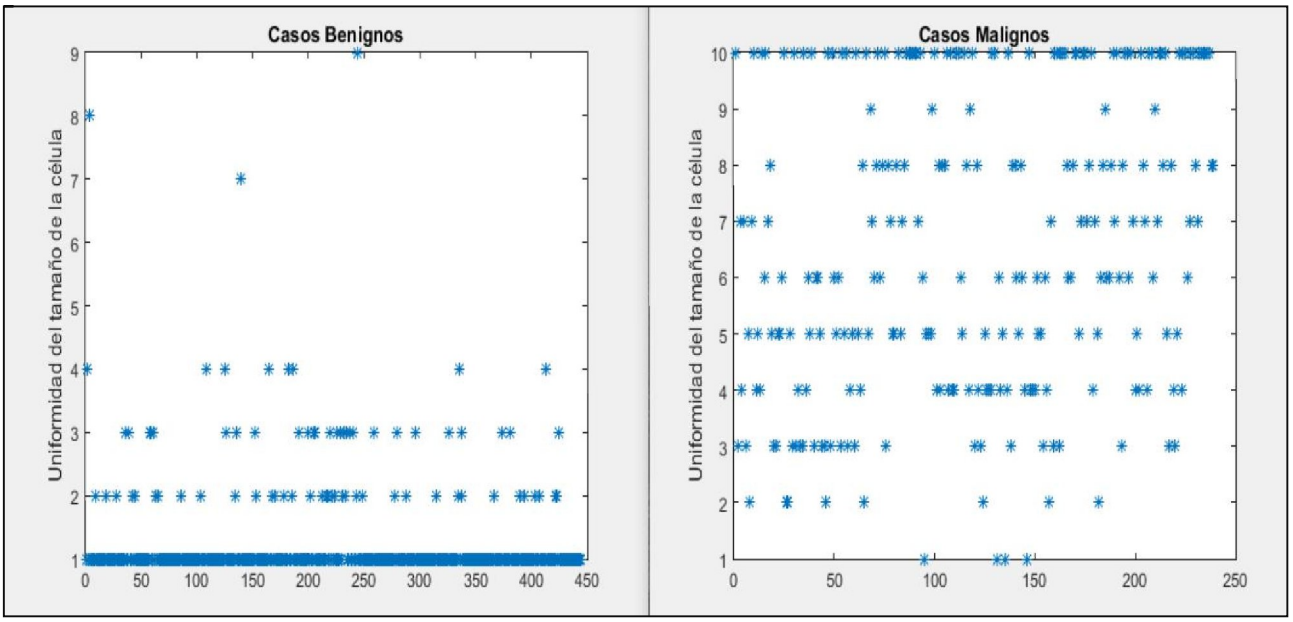


En esta primera característica se puede observar la diferencia entre los casos que son benignos y malignos, aunque algunas pruebas quedan agrupadas en el centro en ambos grupos, es visible tanto en la gráfica como en su media y desviación estándar que la mayoría de los tumores benignos son de tamaño pequeño y la de los malignos normalmente son de tamaño grande.

Uniformidad del tamaño de la célula.

Rango 1-10

	Benignos	Malignos
Media	1.3063	6.5774
Desv.Est	0.8557	2.7242

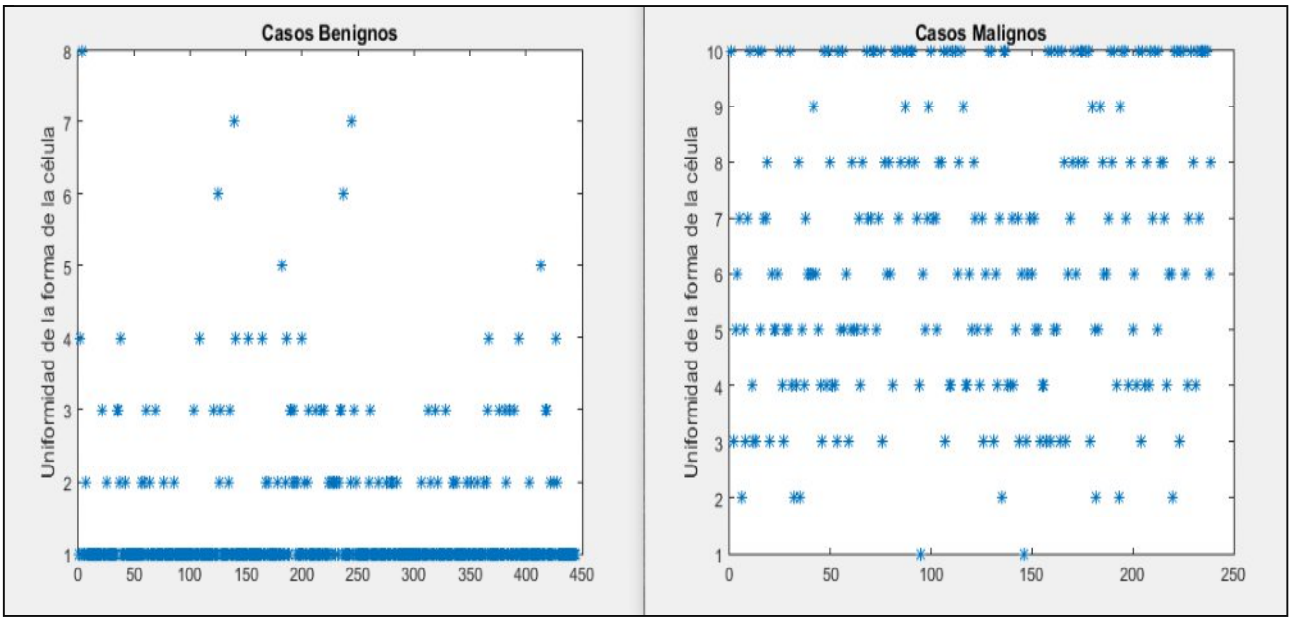


En esta característica se habla de la uniformidad del tamaño del tumor donde vemos que la gran mayoría de los tumores benignos tienen una uniformidad de 1, mientras que los tumores malignos, aunque vemos que la mayoría tienen uniformidad de 10 estos se encuentran dispersos entre los diferentes tamaños.

Uniformidad de la forma de la célula.

Rango 1-10

	Benignos	Malignos
Media	1.4144	6.5607
Desv.Est	0.9570	2.5691

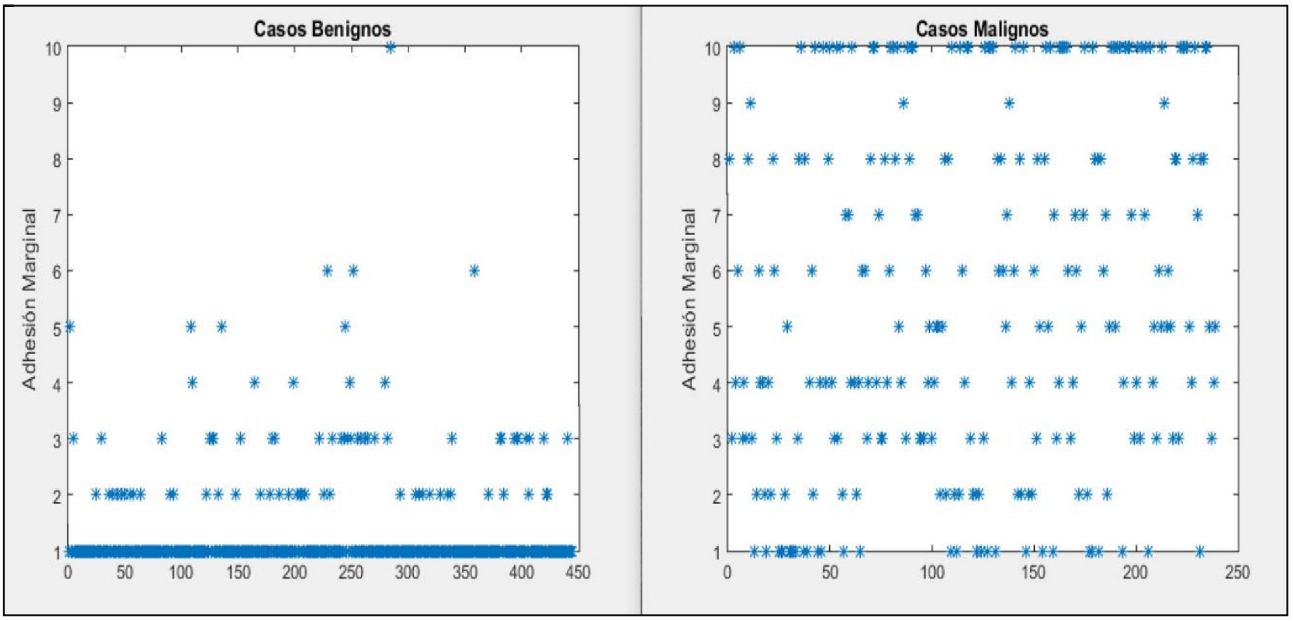


En esta característica podemos observar que prácticamente va de la mano con la pasada que es uniformidad del tamaño de la célula, viendo que la uniformidad de la forma de las células benignas es normalmente más uniforme, mientras que la de las células malignas tienen una tendencia a ser más irregulares y a comparación de los casos benignos los casos están más dispersos en uniformidad de tamaño y forma.

Adhesión Marginal

Rango 1-10

	Benignos	Malignos
Media	1.3468	5.5858
Desv.Est	0.9171	3.1966

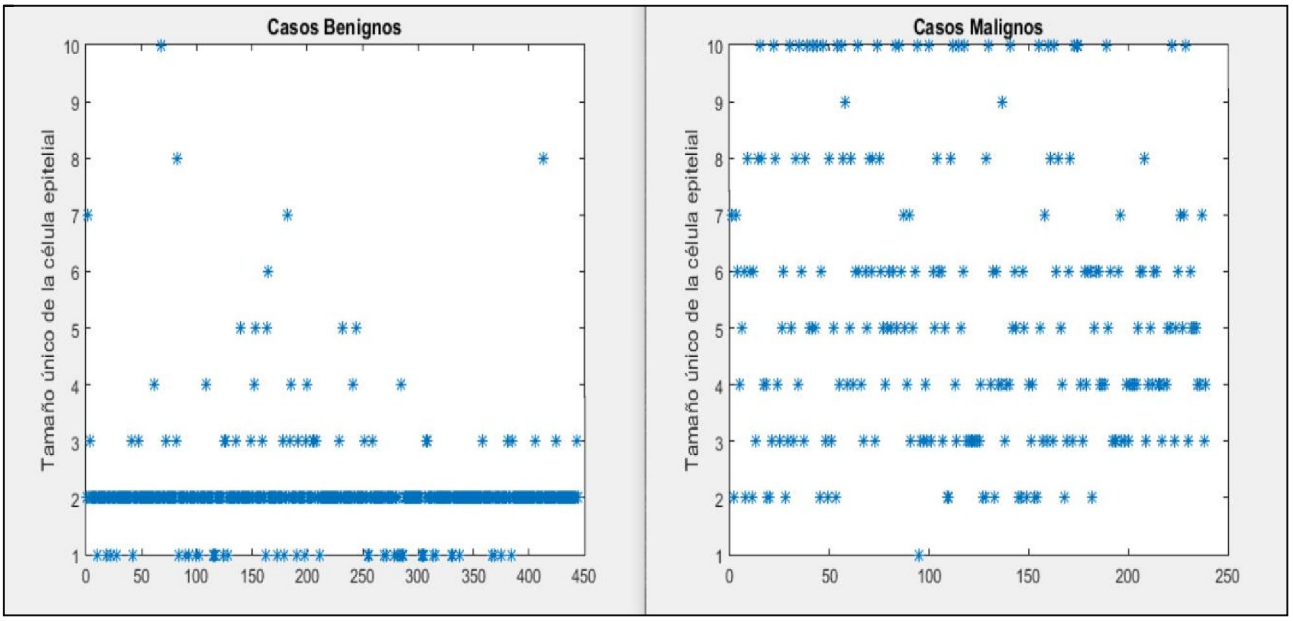


Observamos que la adhesión marginal en los casos benignos es casi nula a excepción de algunos datos que aun así no tienen un número tan alto, mientras que las células de los casos malignos están muy dispersas entre toda la gráfica, aunque la mayoría está en el tope.

Tamaño único de la célula epitelial.

Rango 1-10

	Benignos	Malignos
Media	2.1081	5.3264
Desv.Est	0.8771	2.4431

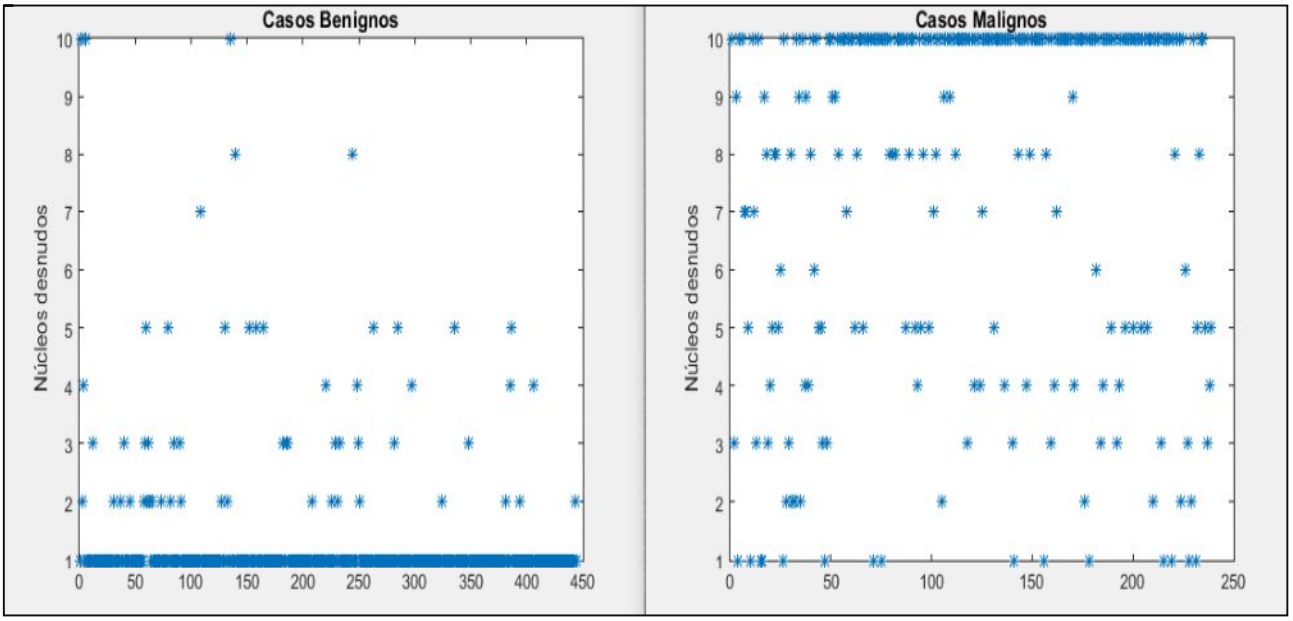


Como se observa es normal tener un tamaño único de célula epitelial bajo, aunque con algunas pruebas teniendo un tamaño mayor estas se mantienen pequeñas, mientras que en los casos de células malignas se ve una media de 5 un tamaño medio, pero están también muy dispersas en la gráfica llegando a tener un tamaño grande.

Núcleos desnudos.

Rango 1-10

	Benignos	Malignos
Media	1.3468	7.6276
Desv.Est	1.1778	3.1167

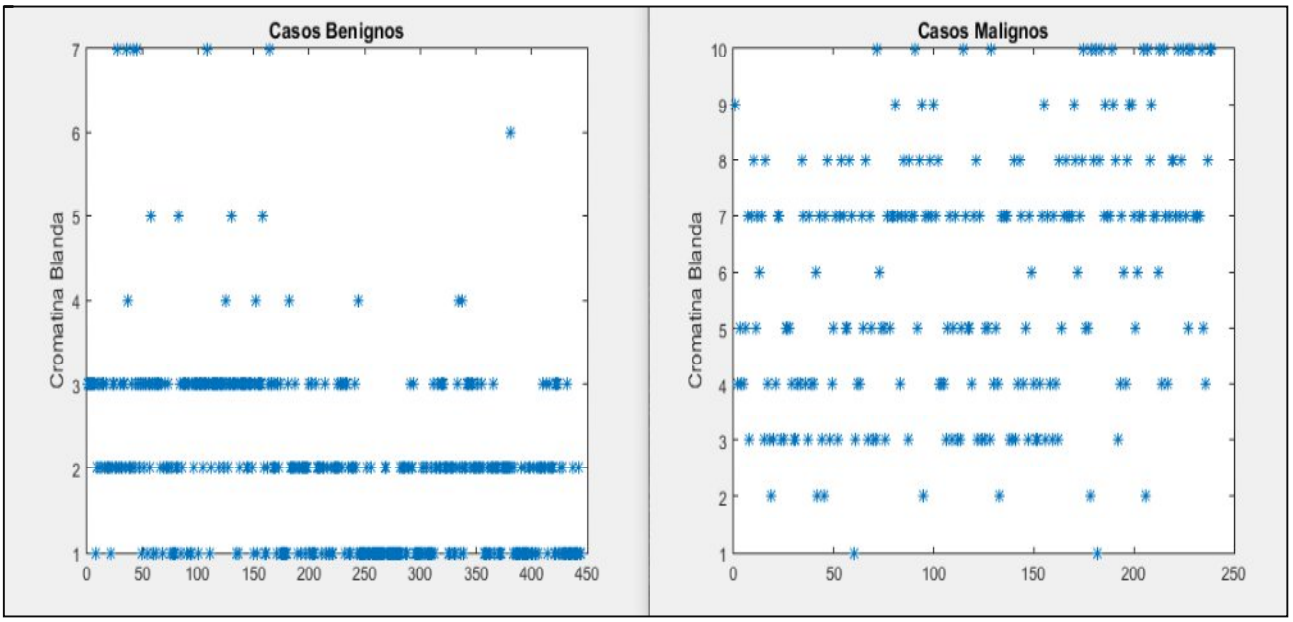


Observamos que los núcleos desnudos en los casos benignos son pocos si no es que nulos, mientras que en los casos malignos la mayoría de ellos tiene gran cantidad de núcleos desnudos mostrando poca dispersión entre los casos presentados.

Cromatina Blanda.

Rango 1-10

	Benignos	Malignos
Media	2.0833	5.9749
Desv.Est	1.0623	2.2824

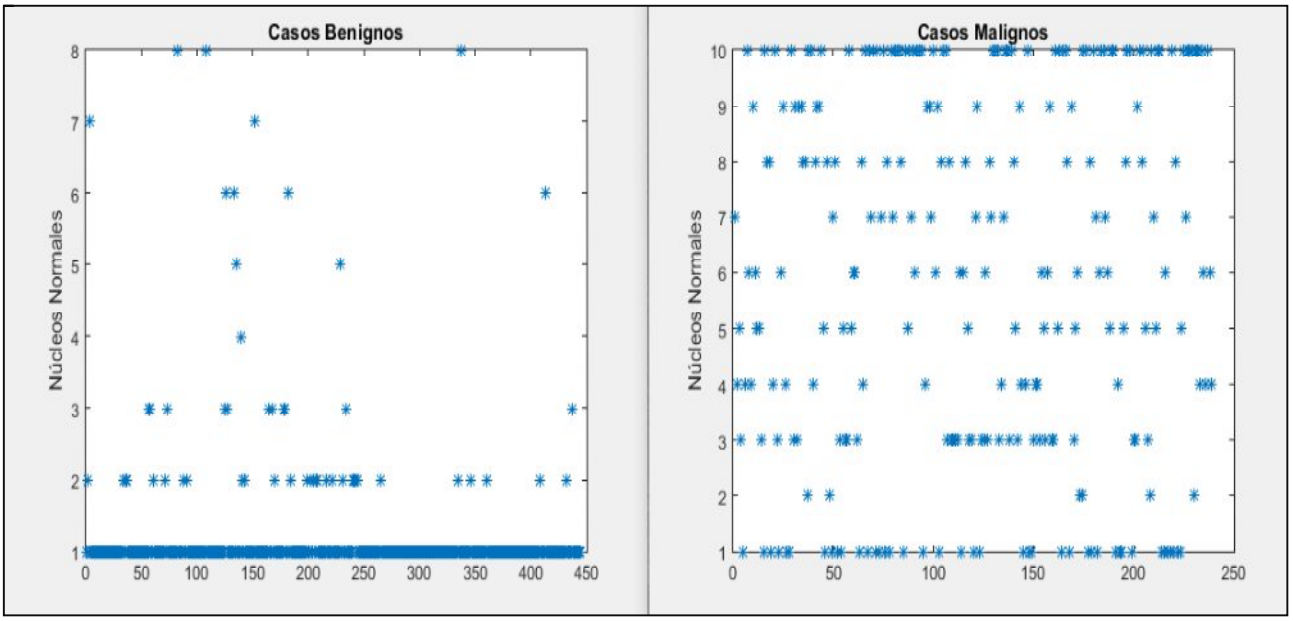


Observamos que en los casos benignos la cromatina blanda se separa su mayoría en los primeros 3 niveles mientras que, en los casos malignos, aunque la media marca en 6 los datos están bastante dispersos en la gráfica.

Núcleos Normales.

Rango 1-10

	Benignos	Malignos
Media	1.2613	5.8577
Desv.Est	0.9546	3.3489

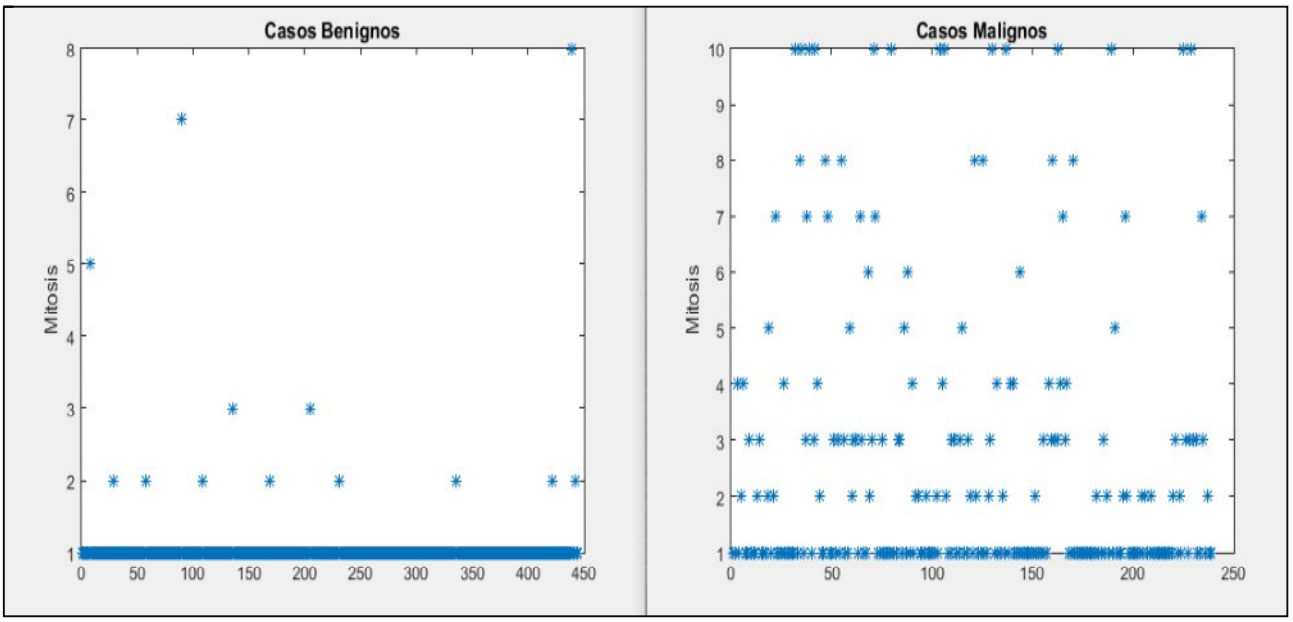


En esta característica la mayoría de datos en los casos benignos se encuentran en 1, en los casos malignos la media de los datos queda arriba de la mitad llegando a tener muchos datos al tope de la gráfica.

Mitosis.

Rango 1-10

	Benignos	Malignos
Media	1.0653	2.6025
Desv.Est	0.5097	2.5645



Aquí se observa que en los casos benignos casi la totalidad de las pruebas quedaron en 1, mientras que en el caso de los tumores malignos también se observa baja mitosis es normal ver que esta es más irregular y puede llegar hasta los niveles más altos de la gráfica.

Training

Matrices de conteos de Malignos y Benignos

Conteo de Malignos

```
10 -  Map_Maligno = Train(:,11)==4;     %Filtra los casos que sean igual a 4 osea malignos
11 -  Malignos=Train((Map_Maligno), :); %Crea la matriz con puros casos malignos
12
13 -  Map_Benigno = Train(:,11)==2;
14 -  Benignos=Train((Map_Benigno), :); %Crea la matriz con puros casos benignos.
15
16 -  Num_malignos=size(Malignos, 1); %Numero de casos malignos
17 -  Num_Benignos=size(Benignos, 1); %Numero de casos benignos
18
19 -  A=unique(Train(:,2:10)); %Saca los valores unicos de la columna 2 a la 10 de la matriz de train
20 -  SA=size(A,1);
21 -  col=(2:10);
22 -  ind=(1:10)';
```

```
24     %Matriz de Malignos
25 -   Mtx=Malignos;
26 -   for c=1:SA
27 -       A(c,col)= sum(Mtx(:, col) == A(c,1));
28 -   end
29
30 -   CMal=A(:,2:10)+1; %Creamos el conteo de datos unicos para cada campo en la matriz de los malignos
31 -   ConMal=[ind,CMal];%Agregamos indice a la matriz
```

Conteo | ConteoMaligno | ConteoBenigno | Norm_Mal | Norm_Ben | ConMal | ConBen

10x10 double

1	2	3	4	5	6	7	8	9	10
1	3	4	3	20	2	8	3	29	87
2	3	6	6	14	14	8	6	6	17
3	9	19	13	21	29	12	26	20	24
4	8	22	22	19	25	9	21	11	7
5	34	15	22	13	23	11	20	12	3
6	14	17	17	14	24	5	6	12	2
7	11	17	19	8	5	5	43	8	5
8	26	19	20	19	15	12	21	10	6
9	10	3	6	4	3	7	8	11	1
10	47	43	37	33	25	88	11	46	13

Hacemos lo mismo para la matriz de benignos.

```
38      %Matriz de Benignos
39 -    Mtz=Benignos;
40 -    AB = A;
41
42 -    for c=1:SA
43 -        AB(c,col)= sum(Mtz(:, col) == A(c,1));
44 -    end
45
46 -    CBen=AB(:,2:10)+1;
47 -    ConBen=[ind,CBen];
```

Conteo | ConteoMaligno | ConteoBenigno | Norm_Mal | Norm_Ben | ConMal | ConBen

10x10 double

1	2	3	4	5	6	7	8	9	10
	91	238	227	239	30	252	97	252	281
2	30	27	33	24	231	17	101	22	6
3	56	20	19	19	20	11	84	10	2
4	53	7	11	6	5	4	3	1	1
5	52	1	2	4	4	7	2	3	2
6	11	1	2	3	2	1	2	4	1
7	1	1	1	1	2	2	7	3	2
8	3	2	2	1	2	1	1	2	2
9	1	1	1	1	1	1	1	1	1
10	1	1	1	1	2	3	1	1	1

Matrices normalizadas de Benignos y Malignos

Tabla normalizada de casos malignos

```
33 -    SMal=sum(ConMal); %Sumamos los datos unicos de la matriz de conteo para obtener el factor de normalizacion
34 -    FNM=SMal(2);      %Apuntamos al numero
35 -    Norm_Mal=ConMal(:,2:10)/FNM; %Normalizamos la matriz para sacar la probabilidad para cada dato
36 -    Norm_Mal=[ind,Norm_Mal]; %Agregamos indice a la matriz anterior
```

Conteo | ConteoMaligno | ConteoBenigno | Norm_Mal | Norm_Ben | ConMal | ConBen

10x10 double

1	2	3	4	5	6	7	8	9	10
1	0.0182	0.0242	0.0182	0.1212	0.0121	0.0485	0.0182	0.1758	0.5273
2	0.0182	0.0364	0.0364	0.0848	0.0848	0.0485	0.0364	0.0364	0.1030
3	0.0545	0.1152	0.0788	0.1273	0.1758	0.0727	0.1576	0.1212	0.1455
4	0.0485	0.1333	0.1333	0.1152	0.1515	0.0545	0.1273	0.0667	0.0424
5	0.2061	0.0909	0.1333	0.0788	0.1394	0.0667	0.1212	0.0727	0.0182
6	0.0848	0.1030	0.1030	0.0848	0.1455	0.0303	0.0364	0.0727	0.0121
7	0.0667	0.1030	0.1152	0.0485	0.0303	0.0303	0.2606	0.0485	0.0303
8	0.1576	0.1152	0.1212	0.1152	0.0909	0.0727	0.1273	0.0606	0.0364
9	0.0606	0.0182	0.0364	0.0242	0.0182	0.0424	0.0485	0.0667	0.0061
10	0.2848	0.2606	0.2242	0.2000	0.1515	0.5333	0.0667	0.2788	0.0788

Ahora hacemos lo mismo para la matriz de benignos.

```
49 -    SBen=sum(ConBen);
50 -    FNB=SBen(2);
51 -    Norm_Ben=ConBen(:,2:10)/FNB;
52 -    Norm_Ben=[ind,Norm_Ben];
```

Conteo | ConteoMaligno | ConteoBenigno | Norm_Mal | Norm_Ben | ConMal | ConBen

10x10 double

1	2	3	4	5	6	7	8	9	10
1	0.3043	0.7960	0.7592	0.7993	0.1003	0.8428	0.3244	0.8428	0.9398
2	0.1003	0.0903	0.1104	0.0803	0.7726	0.0569	0.3378	0.0736	0.0201
3	0.1873	0.0669	0.0635	0.0635	0.0669	0.0368	0.2809	0.0334	0.0067
4	0.1773	0.0234	0.0368	0.0201	0.0167	0.0134	0.0100	0.0033	0.0033
5	0.1739	0.0033	0.0067	0.0134	0.0134	0.0234	0.0067	0.0100	0.0067
6	0.0368	0.0033	0.0067	0.0100	0.0067	0.0033	0.0067	0.0134	0.0033
7	0.0033	0.0033	0.0033	0.0033	0.0067	0.0067	0.0234	0.0100	0.0067
8	0.0100	0.0067	0.0067	0.0033	0.0067	0.0033	0.0033	0.0067	0.0067
9	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033
10	0.0033	0.0033	0.0033	0.0033	0.0067	0.0100	0.0033	0.0033	0.0033

Probabilidades de Malignos y Benignos

Para sacar las probabilidades a priori de que el caso sea benigno o maligno hicimos uso de una función para calcularla.

Clasificador.m | Probabilidad.m

```
function [Prob] = Probabilidad( Matrix_Casos, Columna, valor_col )
%Calcula la probabilidad de tener  el valor_col en la columna indicada
%En este caso Matrix_Casos será Train
%Calcular num de registros de Matrix_casos
%Probabilidad = Num eventos / N
%Prob = Num de coincidencias del valor_col en Columna / Total de registros en Matrix_Casos

num = size(Matrix_Casos, 1);
Mapa = Matrix_Casos(:, Columna) == valor_col;
num_criterio = sum(Mapa);
Prob = num_criterio/num;


end
```

```
%Paso 1
Prob_M = Probabilidad( Train, 11, 4 ); %Probabilidad de  que los casos sean malignos
Prob_B = 1-Prob_M;                     %Probabilidad de que los casos sean benignos

```

Precision	1
Prob_B	0.6509
Prob_M	0.3491
Recall	0.9742

Test

Porcentaje de acierto del clasificador en general

```
54                                          %T E S T I N G
55
56 -   SAT=size(Test); %Obtenemos las dimenciones de la matriz Test
57 -   Conteo=zeros(SAT(1),3); %Creamos matriz de conteos
58
59 -   for f=1:SAT(1)
60 -   x=Test(f,2:11); %Iteramos caso por caso de la matriz Test
61 -   k=Prob_B; %Probabilidad a priori de los casos Benignos
62
63 -       for c=1:9
64 -           k=k*Norm_Ben(x(c),c+1); %Iteramos por cada feature de cada registro de la matriz Test
65 -       end                         %y multiplicamos por su respectiva probabilidad
66 -       Conteo(f)=k; %Agregamos la probabilidad final de que sea benigno para cada caso en una matriz de conteo
67 -   end
69 -   for f=1:SAT(1)
70 -   x=Test(f,2:11);
71 -   k=Prob_M;
72
73 -       for c=1:9
74 -           k=k*Norm_Mal(x(c),c+1); %Ahora sacamos las probabilidades de que el caso sea maligno
75 -       end
76 -       Conteo(f,2)=k; %Agregamos el caso en la siguiente columna de la matriz de conteo para coomparar
77 -   end
78
79 -   for s=1:SAT(1)
80 -      if Conteo(s)>Conteo(s,2);  %Dependiendo de la mayor probabilidad agregamos 2 si es benigno o 4 si
81 -          Conteo(s,3)=2;
82 -      else
83 -          Conteo(s,3)=4;
84 -      end
85 -   end
86
87 -   Coomp=Test(:,11)==Conteo(:,3); %Coomparamos los resultados obtenidos en la matriz de conteo con los
88 -   Porcentaje=sum(Coomp)/SAT(1);%Sacamos el porcentaje de aciertos obtenidos para toda la matriz Test
```

k	0.0053
Malignos	155x11 double
Map_Benigno	444x1 logical
Map_Maligno	444x1 logical
Mtx	155x11 double
Mtz	289x11 double
Norm_Ben	10x10 double
Norm_Mal	10x10 double
Num_Benignos	289
Num_malignos	155
Porcentaje	0.9833
PorcentajeB	0.9742
PorcentajeM	1
Precision	1
Prob_B	0.6509
Prob_M	0.3491
Recall	0.9742
s	155
SA	10
SAT	[239,11]

El porcentaje de aciertos para la matriz completa de Test nos salió 0.983263598326360.

Nuestro algoritmo acertó en un 98% de los casos.

Fragmento de la tabla de comparaciones de la matriz Test

1	2	3
0.0079	2.3088e-10	2
5.1710e-19	7.4155e-09	4
1.6506e-04	2.2409e-10	2
6.6448e-04	4.0744e-11	2
0.0018	2.9103e-12	2
0.0055	4.7012e-12	2
6.0521e-04	3.0558e-11	2
0.0097	1.2537e-11	2
6.9560e-18	1.8643e-08	4
5.3980e-19	4.9675e-10	4
0.0103	1.4104e-11	2
2.3792e-10	3.6872e-10	4
0.0095	5.3281e-11	2
0.0013	1.0186e-11	2
4.5900e-04	1.4260e-11	2
0.0055	4.7012e-12	2
4.8263e-18	2.1113e-08	4
8.7114e-15	2.5487e-10	4
2.8958e-17	5.5181e-08	4

Porcentaje de acierto del clasificador para los casos Malignos.

Aplicamos el mismo procedimiento únicamente para los casos malignos.

```
90                                                  %Test Malignos
91
92                      %Repetimos el proceso ahora separando los casos malignos de la matriz Test
93
94 -    TestMalignos1 = Test(:,11)==4;
95 -    TestMalignos=Test((TestMalignos1), :);  %Tomamos todos los registros malignos de la matriz Test
96
97 -    STM=size(TestMalignos);
98 -    ConteoMaligno=zeros(STM(1),3);
99
100 -   for f=1:STM(1)
101 -   x=TestMalignos(f,2:11);
102 -   k=Prob_M;
103
104 -       for c=1:9
105 -           k=k*Norm_Mal(x(c),c+1);
106 -       end
107 -       ConteoMaligno(f,2)=k;
108 -   end
109
110 -   for f=1:STM(1)
111 -   x=TestMalignos(f,2:11);
112 -   k=Prob_B;
```

```
114 -       for c=1:9
115 -           k=k*Norm_Ben(x(c),c+1);
116 -       end
117 -       ConteoMaligno(f)=k;
118 -   end
119
120 -   for s=1:STM(1)
121 -       if ConteoMaligno(s)>ConteoMaligno(s,2);
122 -           ConteoMaligno(s,3)=2;
123 -       else
124 -           ConteoMaligno(s,3)=4;
125 -       end
126 -   end
127
128 -   CoompM=TestMalignos(:,11) == ConteoMaligno(:,3);
129 -   PorcentajeM=sum(CoompM)/STM(1);
```

Num_Benignos	289
Num_malignos	155
Porcentaje	0.9833
PorcentajeB	0.9742
PorcentajeM	1
Precision	1
Prob_B	0.6509
Prob_M	0.3491
Recall	0.9742
s	155
SA	10
SAT	[239,11]
SBen	[55,299,2
SMal	[55,165,1
STB	[155,11]
STM	[84,11]
Test	239x11 do

El porcentaje para los casos malignos nos salió en 1.

Nuestro algoritmo acertó al 100% de los casos malignos.

Fragmento de la tabla de comparaciones de los casos malignos de la matriz Test.

	1	2	3
1	5.1710e-19	7.4155e-09	4
2	6.9560e-18	1.8643e-08	4
3	5.3980e-19	4.9675e-10	4
4	2.3792e-10	3.6872e-10	4
5	4.8263e-18	2.1113e-08	4
6	8.7114e-15	2.5487e-10	4
7	2.8958e-17	5.5181e-08	4
8	8.2442e-18	2.8390e-10	4
9	2.1070e-12	2.0264e-11	4
10	8.8119e-11	1.9265e-09	4
11	2.0270e-16	1.6330e-08	4
12	1.4630e-14	7.4905e-11	4
13	1.8462e-10	2.1769e-08	4
14	3.8449e-15	1.6632e-08	4
15	4.7128e-13	2.9039e-08	4
16	4.8263e-18	2.3735e-08	4
17	2.6490e-14	1.3230e-09	4
18	5.7251e-21	2.3890e-09	4
19	5.0169e-15	3.5716e-07	4

Porcentaje de acierto del clasificador para los casos Benignos.

Aplicamos el mismo procedimiento únicamente para los casos benignos.

```
                                        %Test Benignos

                    %Repetimos el proceso ahora separando los casos Benignos de la matriz Test

    TestBenignos1 = Test(:,11)==2;
    TestBenignos=Test((TestBenignos1), :); %Tomamos todos los registros benignos de la matriz Test.

    STB=size(TestBenignos);
    ConteoBenigno=zeros(STB(1),3);

  for f=1:STB(1)
    x=TestBenignos(f,2:11);
    k=Prob_M;

        for c=1:9
            k=k*Norm_Mal(x(c),c+1);
        end
        ConteoBenigno(f,2)=k;
    end
```

```
  for f=1:STB(1)
    x=TestBenignos(f,2:11);
    k=Prob_B;

      for c=1:9
          k=k*Norm_Ben(x(c),c+1);
      end
      ConteoBenigno(f)=k;
  end

  for s=1:STB(1)
      if ConteoBenigno(s)>ConteoBenigno(s,2);
          ConteoBenigno(s,3)=2;
      else
          ConteoBenigno(s,3)=4;
      end
  end

  CoompB=TestBenignos(:,11) == ConteoBenigno(:,3);
  PorcentajeB=sum(CoompB)/STB(1);
```

Mtx	155x11 double
Mtz	289x11 double
Norm_Ben	10x10 double
Norm_Mal	10x10 double
Num_Benignos	289
Num_malignos	155
Porcentaje	0.9833
PorcentajeB	0.9742
PorcentajeM	1
Precision	1
Prob_B	0.6509
Prob_M	0.3491
Recall	0.9742
s	155
SA	10
SAT	[239,11]
SBen	[55,299,299,29
SMal	[55,165,165,16
STB	[155,11]
STM	[84,11]
Test	239x11 double

El porcentaje de acierto para los casos benignos nos salió 0.974193548387097.

Nuestro algoritmo acertó en el 97% de los casos benignos.

Fragmento de la tabla de comparaciones para los casos benignos de la matriz Test.

	1	2	3
1	0.0079	2.3088e-10	2
2	1.6506e-04	2.2409e-10	2
3	6.6448e-04	4.0744e-11	2
4	0.0018	2.9103e-12	2
5	0.0055	4.7012e-12	2
6	6.0521e-04	3.0558e-11	2
7	0.0097	1.2537e-11	2
8	0.0103	1.4104e-11	2
9	0.0095	5.3281e-11	2
10	0.0013	1.0186e-11	2
11	4.5900e-04	1.4260e-11	2
12	0.0055	4.7012e-12	2
13	0.0167	4.7012e-12	2
14	0.0081	5.4325e-11	2
15	8.0784e-04	3.4476e-11	2
16	9.7500e-04	8.7756e-12	2
17	8.9880e-04	3.4632e-10	2
18	0.0139	2.0372e-11	2
19	1.8934e-08	7.5189e-10	2

Métricas

```
172                                                 %M e t r i c s
173
174 -    TruePositive(1)=sum(CoompB);
175 -    FalseNegative=STB(1)-TruePositive;
176 -    TrueNegative(1)=sum(CoompM);
177 -    FalsePositive=STM(1)-TrueNegative;
178
179                                                 %P r e c i s i o n
180 -    Precision = TruePositive /(TruePositive + FalsePositive);
181
182                                                 %R e c a l l
183 -    Recall  = TruePositive /(TruePositive + FalseNegative);
184
185                                                 %F-Score
186 -    FScore= 2 * Precision * Recall / (Precision + Recall);
```

Precision: 1

Recall: 0.974193548387097

F-Score: 0.986928104575163

Conclusión

El algoritmo naïve bayes ha demostrado ser bastante confiable para clasificar pruebas que en este caso fueron pruebas de cáncer de mama, demostrando una eficacia en el 98% de los casos en general, este resultado fue obtenido sabiendo interpretar los datos identificando y analizando los datos clasificados en cada prueba obtenida, el tema de estudio en este trabajo es uno muy delicado debido a que estamos tratando con pruebas de casos de cancer de mama, aunque el algoritmo tiene 98% de aciertos en general, para los casos malignos tuvo un acierto del 100% esto quiere decir que nuestro algoritmo no saco casos de falsos positivos que son los mas perjudiciales a la hora de diagnosticar un caso de cancer de mama, al contrario de los casos benignos el algoritmo si tuvo errores al diagnoaticar siertos casos teniendo un porcentaje de acierto de 97% aunque los casos en los que salieron falsos negativos no son tan dañinos como un falso positivo ya que es preferible un diagnóstico de un falso maligno a que el resultado sea un falso benigno por que la persona confiaría que no padece de cáncer cuando en realidad necesita llevar tratamiento lo más tempranamente posible, esto demuestra buena confiabilidad en nuestro algoritmo aunque con mas pruebas podria entrenarse mejor para poder clasificar los casos con aun mas eficiencia y tener una mayor confiabilidad.