In [1]:
```
import time
inicio = time.time()
```

In [2]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from imblearn.under_sampling import ClusterCentroids
# el muestreo aleatorio simple sin reemplazamiento no sirve porque no capta la var
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import LeaveOneOut

from sklearn.naive_bayes import GaussianNB

from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
from sklearn.metrics import average_precision_score
from sklearn.metrics import precision_recall_curve
```

Se importan los datos

In [3]:
```
df = pd.read_excel("..//Base_datos_Clasificacion binaria.xlsx", index_col = 0)
```

In [4]:
```
df
```

Out[4]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	MP	Empresa sancionada	UTE
0	0	1738093.21	1484428.72	6	1	0
1	0	469670.24	272492.00	4	0	0
2	0	1025088.19	707310.85	1	0	0
3	0	999890.00	497621.36	6	0	0
4	0	72598.27	47508.92	1	1	0
...	...	...	...	...	...	...
2211	1	5542028.88	5514320.00	6	1	0
2212	1	6095782.00	5100555.00	6	1	0
2213	1	3752906.00	3744400.00	6	1	0
2214	1	23896564.00	19547338.00	6	0	1
2215	1	33415098.00	33349532.00	6	1	0

2216 rows × 6 columns

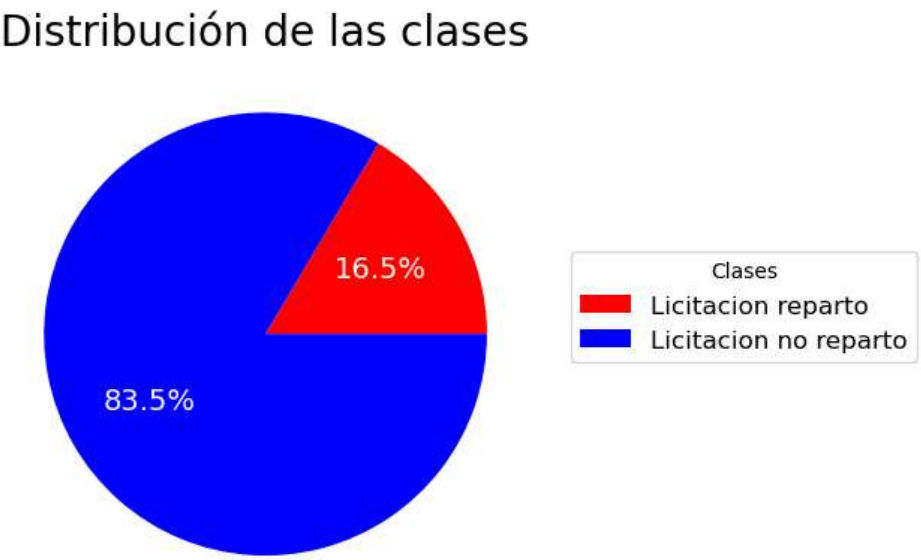
Se trata el desbalanceo de clases mediante una tecnica de Infra - Muestreo. Hay que

instalar la libreria imbalanced - learn

In [5]:
```
y = df["Licitacion reparto"]
X = df[df.columns[1:]]
```

In [6]:
```
datos = [len(y[y == 1]), len(y[y == 0])]
etiquetas = ['Licitacion reparto', 'Licitacion no reparto']
colores = ['red', 'blue']

plt.pie(datos, labels = etiquetas, colors = colores, autopct = '%1.1f%%', textprop
plt.title('Distribución de las clases', fontsize = 20)
plt.legend(title = "Clases", bbox_to_anchor=(1.7, 0.55), loc='right', fontsize = 1
plt.show()
```



Codificacion OneHotEncoder

In [7]:
```
codificador = OneHotEncoder()
```

In [8]:
```
df[["MP", "Empresa sancionada", "UTE"]] = df[["MP", "Empresa sancionada", "UTE"]].
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2216 entries, 0 to 2215
Data columns (total 6 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Licitacion reparto    2216 non-null   int64  
 1   Importe presupuestado 2216 non-null   float64
 2   Importe adjudicado    2216 non-null   float64
 3   MP                    2216 non-null   float64
 4   Empresa sancionada    2216 non-null   float64
 5   UTE                   2216 non-null   float64
dtypes: float64(5), int64(1)
memory usage: 121.2 KB
```

In [9]:
```
df
```

Out[9]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	MP	Empresa sancionada	UTE
0	0	1738093.21	1484428.72	6.0	1.0	0.0
1	0	469670.24	272492.00	4.0	0.0	0.0
2	0	1025088.19	707310.85	1.0	0.0	0.0
3	0	999890.00	497621.36	6.0	0.0	0.0
4	0	72598.27	47508.92	1.0	1.0	0.0
...	...	...	...	...	...	...
2211	1	5542028.88	5514320.00	6.0	1.0	0.0
2212	1	6095782.00	5100555.00	6.0	1.0	0.0
2213	1	3752906.00	3744400.00	6.0	1.0	0.0
2214	1	23896564.00	19547338.00	6.0	0.0	1.0
2215	1	33415098.00	33349532.00	6.0	1.0	0.0

2216 rows × 6 columns

In [10]:
```
codificacion = codificador.fit_transform(df[["MP"]])

mercados = pd.DataFrame(codificacion.toarray(),
                        columns = ["MP1", "MP4", "MP6"])
df.drop("MP", axis = 1, inplace = True)
df.reset_index(inplace = True, drop = True)
df = pd.concat([df, mercados], axis = 1)
df
```

Out[10]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	Empresa sancionada	UTE	MP1	MP4	MP6
0	0	1738093.21	1484428.72	1.0	0.0	0.0	0.0	1.0
1	0	469670.24	272492.00	0.0	0.0	0.0	1.0	0.0
2	0	1025088.19	707310.85	0.0	0.0	1.0	0.0	0.0
3	0	999890.00	497621.36	0.0	0.0	0.0	0.0	1.0
4	0	72598.27	47508.92	1.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...
2211	1	5542028.88	5514320.00	1.0	0.0	0.0	0.0	1.0
2212	1	6095782.00	5100555.00	1.0	0.0	0.0	0.0	1.0
2213	1	3752906.00	3744400.00	1.0	0.0	0.0	0.0	1.0
2214	1	23896564.00	19547338.00	0.0	1.0	0.0	0.0	1.0
2215	1	33415098.00	33349532.00	1.0	0.0	0.0	0.0	1.0

2216 rows × 8 columns

In [11]:
```
y = df["Licitacion reparto"]
X = df[df.columns[1:]]
```

In [12]:
```
nombre_clases = ["Licitacion no reparto", "Licitacion reparto"]
nombre_predicciones = list(map(lambda x : x + " predicho", nombre_clases))
```

Se utiliza la técnica "ClusterCentroids" que reemplaza las observaciones por un centroide dado por los k vecinos más cercanos.

In [13]:
```
# conda install -c conda-forge imbalanced-learn
cc = ClusterCentroids(random_state=123)
X_res, y_res = cc.fit_resample(X, y)
df = pd.concat([y_res, X_res], axis = 1)
df
```

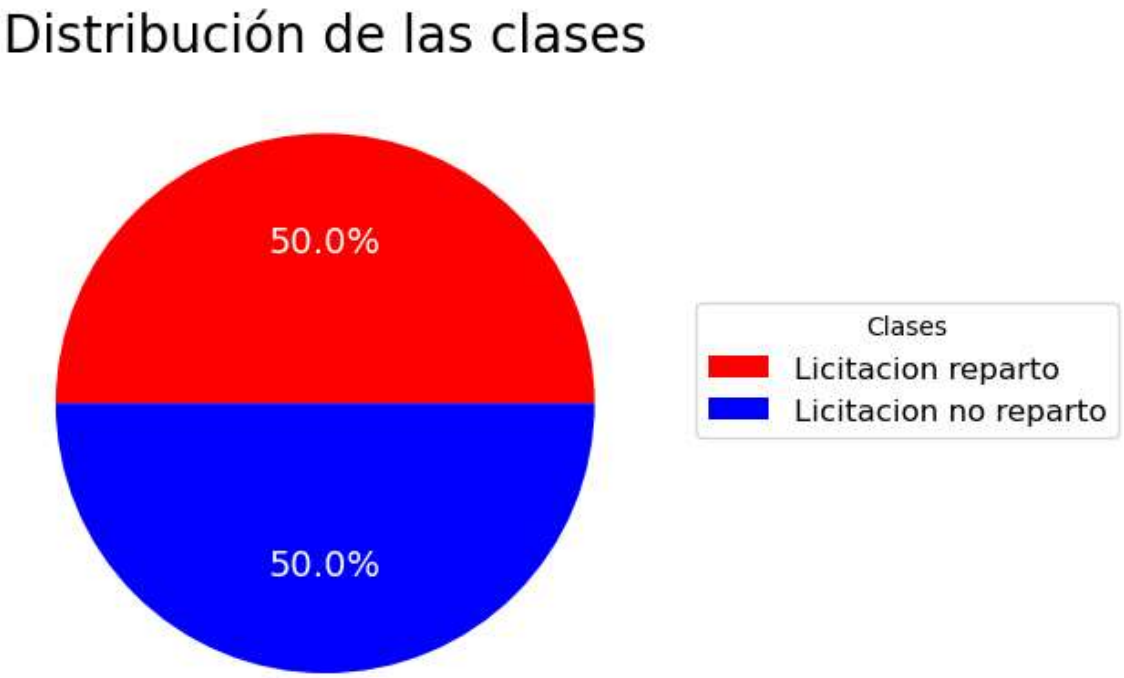
Out[13]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	Empresa sancionada	UTE	MP1	MP4	MP6
0	0	6.776405e+04	6.312591e+04	0.287037	0.037037	0.083333	0.009259	0.907407
1	0	3.563140e+08	2.915613e+08	0.000000	1.000000	0.000000	0.000000	1.000000
2	0	6.513581e+07	6.513581e+07	0.000000	0.000000	1.000000	0.000000	0.000000
3	0	6.407947e+08	5.119260e+08	0.000000	1.000000	0.000000	0.000000	1.000000
4	0	2.458220e+07	2.428981e+07	1.000000	1.000000	1.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...
725	1	5.542029e+06	5.514320e+06	1.000000	0.000000	0.000000	0.000000	1.000000
726	1	6.095782e+06	5.100555e+06	1.000000	0.000000	0.000000	0.000000	1.000000
727	1	3.752906e+06	3.744400e+06	1.000000	0.000000	0.000000	0.000000	1.000000
728	1	2.389656e+07	1.954734e+07	0.000000	1.000000	0.000000	0.000000	1.000000
729	1	3.341510e+07	3.334953e+07	1.000000	0.000000	0.000000	0.000000	1.000000

730 rows × 8 columns

In [14]:
```
datos = [len(y_res[y_res == 1]), len(y_res[y_res == 0])]
etiquetas = ['Licitacion reparto', 'Licitacion no reparto']
colores = ['red', 'blue']

plt.pie(datos, labels = etiquetas, colors = colores, autopct = '%1.1f%%', textprop
plt.title('Distribución de las clases', fontsize = 20)
plt.legend(title = "Clases", bbox_to_anchor=(1.7, 0.55), loc='right', fontsize = 1
plt.show()
```



In [15]:
```
df
```

Out[15]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	Empresa sancionada	UTE	MP1	MP4	MP6
0	0	6.776405e+04	6.312591e+04	0.287037	0.037037	0.083333	0.009259	0.907407
1	0	3.563140e+08	2.915613e+08	0.000000	1.000000	0.000000	0.000000	1.000000
2	0	6.513581e+07	6.513581e+07	0.000000	0.000000	1.000000	0.000000	0.000000
3	0	6.407947e+08	5.119260e+08	0.000000	1.000000	0.000000	0.000000	1.000000
4	0	2.458220e+07	2.428981e+07	1.000000	1.000000	1.000000	0.000000	0.000000
...	...	...	...	...	...	...	...	...
725	1	5.542029e+06	5.514320e+06	1.000000	0.000000	0.000000	0.000000	1.000000
726	1	6.095782e+06	5.100555e+06	1.000000	0.000000	0.000000	0.000000	1.000000
727	1	3.752906e+06	3.744400e+06	1.000000	0.000000	0.000000	0.000000	1.000000
728	1	2.389656e+07	1.954734e+07	0.000000	1.000000	0.000000	0.000000	1.000000
729	1	3.341510e+07	3.334953e+07	1.000000	0.000000	0.000000	0.000000	1.000000

730 rows × 8 columns

In [16]:
```
df[df.columns[4:]] = df[df.columns[4:]].applymap(lambda x : 1 if x >= 0.5 else 0)
df[df.columns[4:]]
```

Out[16]:

	UTE	MP1	MP4	MP6
0	0	0	0	1
1	1	0	0	1
2	0	1	0	0
3	1	0	0	1
4	1	1	0	0
...	...	...	...	...
725	0	0	0	1
726	0	0	0	1
727	0	0	0	1
728	1	0	0	1
729	0	0	0	1

730 rows × 4 columns

In [17]:
```
df
```

Out[17]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	Empresa sancionada	UTE	MP1	MP4	MP6
0	0	6.776405e+04	6.312591e+04	0.287037	0	0	0	1
1	0	3.563140e+08	2.915613e+08	0.000000	1	0	0	1
2	0	6.513581e+07	6.513581e+07	0.000000	0	1	0	0
3	0	6.407947e+08	5.119260e+08	0.000000	1	0	0	1
4	0	2.458220e+07	2.428981e+07	1.000000	1	1	0	0
...	...	...	...	...	...	...	...	...
725	1	5.542029e+06	5.514320e+06	1.000000	0	0	0	1
726	1	6.095782e+06	5.100555e+06	1.000000	0	0	0	1
727	1	3.752906e+06	3.744400e+06	1.000000	0	0	0	1
728	1	2.389656e+07	1.954734e+07	0.000000	1	0	0	1
729	1	3.341510e+07	3.334953e+07	1.000000	0	0	0	1

730 rows × 8 columns

Se estandarizan y normalizan los predictores numericos

In [18]:
```
predictores_numericos = df[["Importe presupuestado", "Importe adjudicado"]]
```

In [19]:
```
scaler = StandardScaler()
stand = scaler.fit_transform(predictores_numericos)
```

In [20]:
```
df[["Importe presupuestado", "Importe adjudicado"]] = stand
predictores_numericos = df[["Importe presupuestado", "Importe adjudicado"]]
predictores_numericos
```

Out[20]:

	Importe presupuestado	Importe adjudicado
0	-0.404502	-0.411394
1	5.691897	5.257115
2	0.709000	0.854017
3	10.560180	9.542354
4	0.015011	0.059721
...	...	...
725	-0.310822	-0.305389
726	-0.301345	-0.313435
727	-0.341439	-0.339807
728	0.003277	-0.032502
729	0.166167	0.235898

730 rows × 2 columns

In [21]:
```
norm = (predictores_numericos - predictores_numericos.min()) / (predictores_numeri
```

In [22]:
```
df[["Importe presupuestado", "Importe adjudicado"]] = norm
predictores_numericos = df[["Importe presupuestado", "Importe adjudicado"]]
predictores_numericos
```

Out[22]:

	Importe presupuestado	Importe adjudicado
0	0.000069	0.000085
1	0.556034	0.569521
2	0.101616	0.127203
3	1.000000	1.000000
4	0.038327	0.047411
...	...	...
725	0.008613	0.010734
726	0.009477	0.009925
727	0.005821	0.007276
728	0.037257	0.038147
729	0.052112	0.065109

730 rows × 2 columns

In [23]:
```
df
```

Out[23]:

	Licitacion reparto	Importe presupuestado	Importe adjudicado	Empresa sancionada	UTE	MP1	MP4	MP6
0	0	0.000069	0.000085	0.287037	0	0	0	1
1	0	0.556034	0.569521	0.000000	1	0	0	1
2	0	0.101616	0.127203	0.000000	0	1	0	0
3	0	1.000000	1.000000	0.000000	1	0	0	1
4	0	0.038327	0.047411	1.000000	1	1	0	0
...	...	...	...	...	...	...	...	...
725	1	0.008613	0.010734	1.000000	0	0	0	1
726	1	0.009477	0.009925	1.000000	0	0	0	1
727	1	0.005821	0.007276	1.000000	0	0	0	1
728	1	0.037257	0.038147	0.000000	1	0	0	1
729	1	0.052112	0.065109	1.000000	0	0	0	1

730 rows × 8 columns

In [24]:
```
nombre_clases = ["Licitacion no reparto", "Licitacion reparto"]
nombre_predicciones = list(map(lambda x : x + " predicho", nombre_clases))
```

Se hace la estimacion por NaiveBayes

In [25]:
```
X = df[df.columns[1:]]
y = df["Licitacion reparto"]


loo = LeaveOneOut()
gnb = GaussianNB()
predicciones = []
prob_predicciones = []

for train_index, test_index in loo.split(X):

    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y[train_index], y[test_index]

    gnb.fit(X_train, y_train)
    y_pred = gnb.predict(X_test) #son 0 o 1 dependiendo de si la probabilidad es m
    y_pred_prob = gnb.predict_proba(X_test)
    predicciones.append(y_pred[0])
    prob_predicciones.append(y_pred_prob[0][1])
```

In [26]:
```
gnb.get_params()
```

Out[26]:
```
{'priors': None, 'var_smoothing': 1e-09}
```

In [27]:
```
predicciones = pd.Series(predicciones, name = "Prediccion licitacion reparto")
prob_predicciones = pd.Series(prob_predicciones, name = "Prob prediccion")
cp = pd.concat([y, predicciones, prob_predicciones], axis = 1)
cp
# Este dataframe no contiene las predicciones calculadas en el punto optimo de la c
# Contiene las predicciones calculadas segun el punto por defecto para determinar l
```

Out[27]:

	Licitacion reparto	Prediccion licitacion reparto	Prob prediccion
0	0	0	0.051588
1	0	1	0.999990
2	0	1	0.606321
3	0	1	1.000000
4	0	1	0.748123
...	...	...	...
725	1	0	0.102566
726	1	0	0.102555
727	1	0	0.103948
728	1	0	0.035372
729	1	0	0.091805

730 rows × 3 columns

CURVA ROC. El punto optimo es = (fpr = 0, tpr = 1)

In [28]:
```
auc_score = roc_auc_score(y, prob_predicciones)
auc_score
```

Out[28]:
```
0.7634528054043911
```

In [29]:
```
fpr, tpr, thresholds = roc_curve(y, prob_predicciones)
#fpr es 1 - especificidad = 1 - (TN / (FP + TN)) = FP / (FP + TN)
#tpr es la sensibilidad = TP / (TP + FN)
```

In [30]:
```
distances = np.linalg.norm(np.column_stack((fpr, tpr)) - np.array([0, 1]), axis=1)
optimal_threshold_index = np.argmin(distances)
optimal_threshold = thresholds[optimal_threshold_index]
optimal_threshold
```

Out[30]:
```
0.6015957129887438
```

In [31]:
```
optimal_fpr = fpr[optimal_threshold_index]
optimal_fpr
```

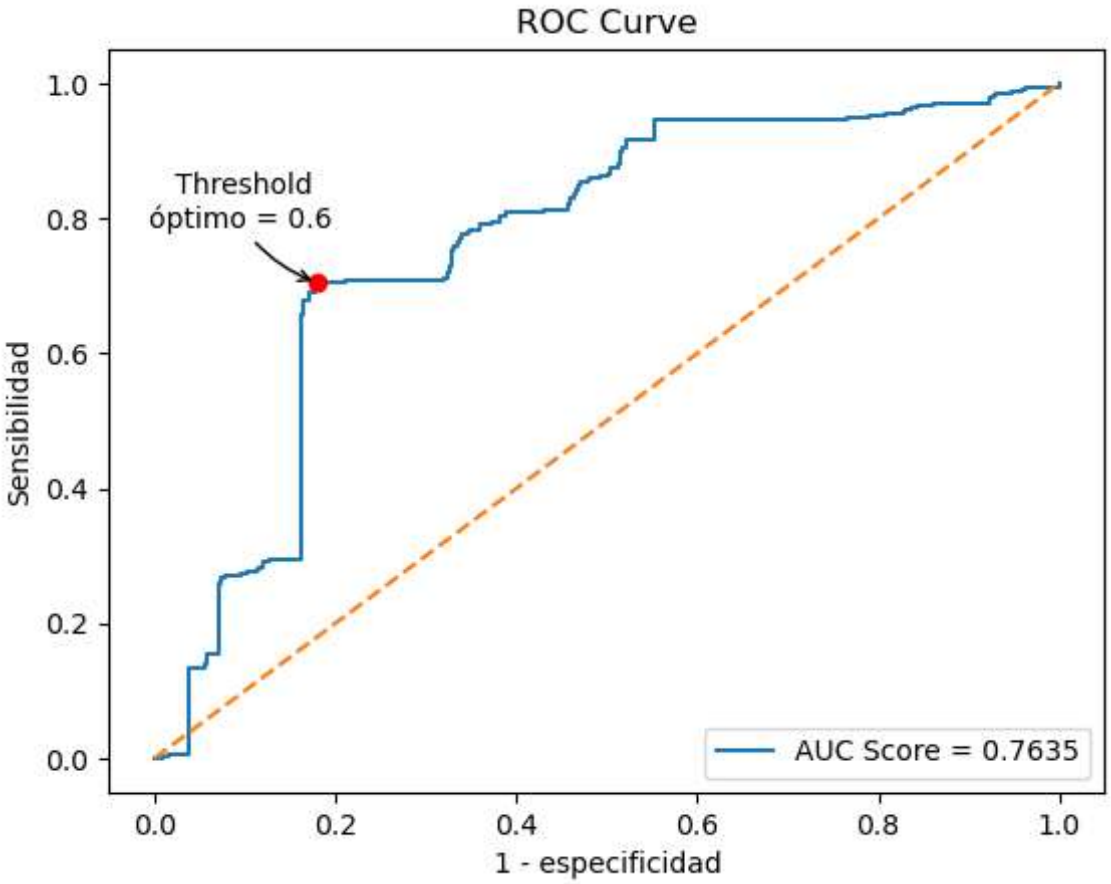
Out[31]:
```
0.18082191780821918
```

In [32]:
```
optimal_tpr = tpr[optimal_threshold_index]
optimal_tpr
```

Out[32]:
```
0.7041095890410959
```

In [59]:
```
plt.plot(fpr, tpr, label = f'AUC Score = {auc_score:.4}')
plt.plot([0,1], linestyle="--")
plt.plot(optimal_fpr, optimal_tpr, 'ro')
plt.annotate(f'      Threshold\n   óptimo = {optimal_threshold:.2}',
```

```
            xy=(optimal_fpr, optimal_tpr), xycoords='data',
            xytext=(-70, +20), textcoords='offset points', fontsize=10,
            arrowprops=dict(arrowstyle="->", connectionstyle = "arc3, rad = .2"))
plt.ylabel('Sensibilidad')
plt.xlabel('1 - especificidad')
plt.title('ROC Curve')
plt.legend(loc = 'lower right')
plt.show()
```



In [34]:
```
cp["Prediccion licitacion reparto en el punto optimo"] = 0
prediccion_final = cp["Prediccion licitacion reparto en el punto optimo"]
prediccion_final = cp.apply(lambda x: 1 if x["Prob prediccion"] >= optimal_thresho
prediccion_final
```

Out[34]:
```
0      0
1      1
2      1
3      1
4      1
      ..
725    0
726    0
727    0
728    0
729    0
Length: 730, dtype: int64
```

In [35]:
```
cp["Prediccion licitacion reparto en el punto optimo"] = prediccion_final
pd.concat([cp, prediccion_final], axis = 1)
cp.drop(columns = "Prediccion licitacion reparto", axis = 1, inplace = True)
cp
```

Out[35]:

	Licitacion reparto	Prob prediccion	Prediccion licitacion reparto en el punto optimo
0	0	0.051588	0
1	0	0.999990	1
2	0	0.606321	1
3	0	1.000000	1
4	0	0.748123	1
...	...	...	...
725	1	0.102566	0
726	1	0.102555	0
727	1	0.103948	0
728	1	0.035372	0
729	1	0.091805	0

730 rows × 3 columns

In [36]:
```
cm = confusion_matrix(y, cp["Prediccion licitacion reparto en el punto optimo"]) #
cm = pd.DataFrame(cm, columns = nombre_clases, index = nombre_predicciones)
cm # esta matriz de confusion esta calculada en el punto optimo de la curva
```

Out[36]:

	Licitacion no reparto	Licitacion reparto
Licitacion no reparto predicho	299	66
Licitacion reparto predicho	108	257

Metricas de la Matriz de Confusion

In [37]:
```
TP = cm.iloc[1,1]
FN = cm.iloc[1,0]
FP = cm.iloc[0,1]
TN = cm.iloc[0,0]
```

In [38]:
```
Tasa_de_aciertos = (TP + TN) / len(df)
Tasa_de_aciertos
```

Out[38]:
```
0.7616438356164383
```

In [39]:
```
Tasa_de_errores = 1 - Tasa_de_aciertos
Tasa_de_errores
```

Out[39]:
```
0.23835616438356166
```

In [40]:
```
Sensibilidad = TP / (TP + FN)
Sensibilidad # se calculo antes. Es optimal_tpr
```

Out[40]:
```
0.7041095890410959
```

In [41]:
```
Tasa_de_falsos_negativos = 1 - Sensibilidad
Tasa_de_falsos_negativos
```

Out[41]:
```
0.2958904109589041
```

In [42]:
```
Especificidad = TN / (FP + TN)
Especificidad # se calculo antes. Es optimal_fpr = 1 - Especificidad
```

Out[42]:
```
0.8191780821917808
```

In [43]:
```
Precision = TP / (TP + FP)
Precision
```

Out[43]:
```
0.7956656346749226
```

In [44]:
```
FP / (FP + TP) # 1 - Precision
```

Out[44]:
```
0.2043343653250774
```

In [45]:
```
F1 = (2 * Sensibilidad * Precision) / (Sensibilidad + Precision)
F1
```

Out[45]:
```
0.747093023255814
```

Métricas de clasificación binaria en el punto óptimo

In [46]:
```
nombre_fichero = "Clasificacion binaria Infra_Muestreo gnb"
```

In [47]:
```
indice_resultados = ["Tasa de aciertos", "Sensibilidad", "Especificidad", "Precisi
resultados_valores = [Tasa_de_aciertos, Sensibilidad, Especificidad, Precision, F1

resultados = pd.DataFrame(data={f"{nombre_fichero}": resultados_valores},
                          index=indice_resultados)
resultados
```

Out[47]:

	Clasificacion binaria Infra_Muestreo gnb
Tasa de aciertos	0.761644
Sensibilidad	0.704110
Especificidad	0.819178
Precision	0.795666
F1	0.747093

Se exportan los datos

In [48]:
```
nombre_columna = list(resultados.columns)[0]
resultados.to_excel(f"Resultados de las métricas de {nombre_fichero}.xlsx")
```

In [49]:
```
average_precision = average_precision_score(y, prob_predicciones)
average_precision
```

Out[49]:
```
0.6904046904998515
```

In [50]:
```
precision, recall, thresholds = precision_recall_curve(y, prob_predicciones)
#precision = TP / (TP + FP)
#recall = sensibilidad = TP / (TP + FN)
```

In [51]:
```
distances = np.linalg.norm(np.column_stack((precision, recall)) - np.array([1, 1]),
optimal_threshold_index = np.argmin(distances)
optimal_threshold = thresholds[optimal_threshold_index]
optimal_threshold
```

Out[51]:
```
0.6015957129887438
```

In [52]:
```
optimal_precision = precision[optimal_threshold_index]
optimal_precision
```

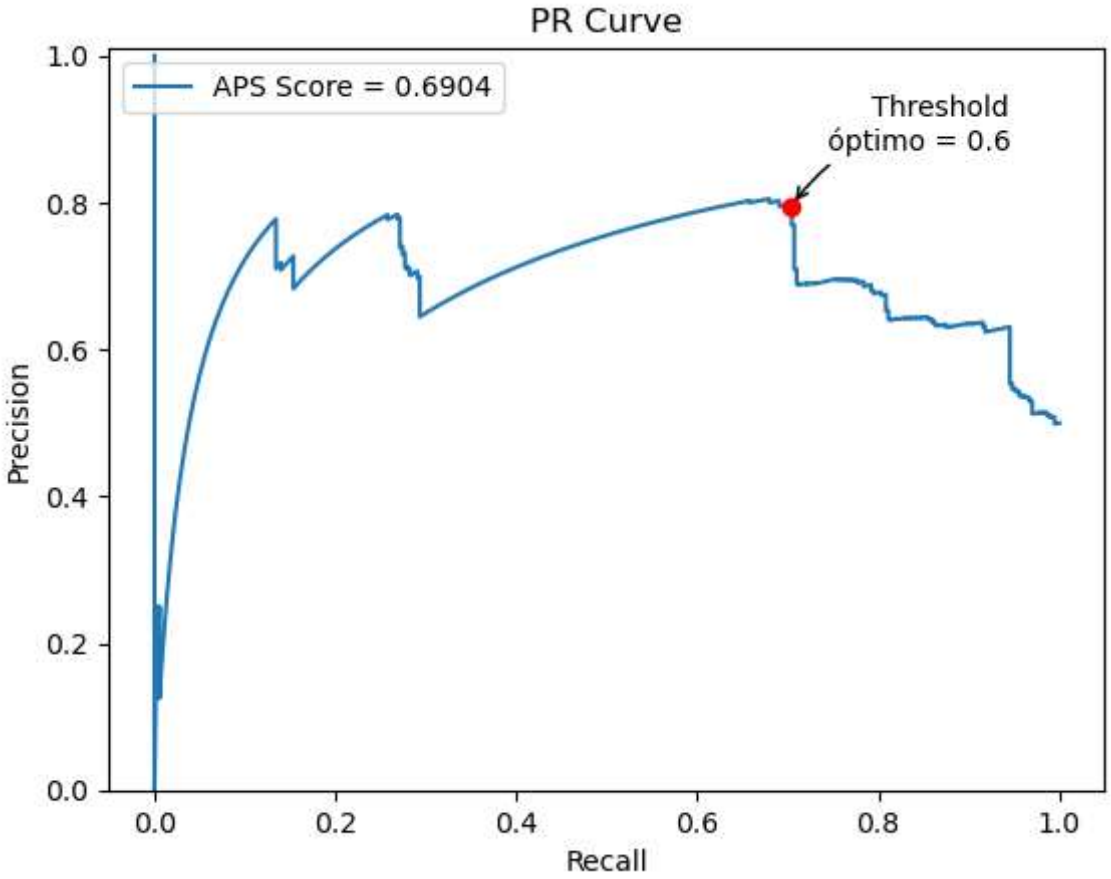
Out[52]:
```
0.7956656346749226
```

In [53]:
```
optimal_recall = recall[optimal_threshold_index]
optimal_recall
```

Out[53]:
```
0.7041095890410959
```

CURVA PR. El punto óptimo es = (precision = 1, recall = 1)

In [63]:
```
plt.plot(recall, precision, label = f'APS Score = {average_precision:.4}')
plt.plot(optimal_recall, optimal_precision, 'ro')
plt.annotate(f'      Threshold\n óptimo = {optimal_threshold:.2}',
             xy=(optimal_recall, optimal_precision), xycoords='data',
             xytext=(10, +20) ,textcoords='offset points', fontsize=10,
             arrowprops=dict(arrowstyle="->", connectionstyle = "arc3, rad = .2"))
plt.ylabel('Precision')
plt.xlabel('Recall')
plt.title('PR Curve')
plt.legend(loc = 'upper left')
plt.ylim(0, 1.01)
plt.show()
```



Se calcula cuanto tiempo ha transcurrido en la ejecucion de todo el notebook y se guarda el tiempo en un fichero de texto para almacenarlo.

In [55]:
```
fin = time.time()
tiempo_transcurrido = fin - inicio
minutos = int((tiempo_transcurrido % 3600) // 60)
segundos = int(tiempo_transcurrido % 60)
with open(f"Tiempo de ejecucion transcurrido en el notebook {nombre_fichero}.txt",
    f.write(f"Tiempo transcurrido: {minutos} minutos, {segundos} segundos")
```

In [56]:
```
print(f"Tiempo transcurrido: {minutos} minutos, {segundos} segundos")
```

```
Tiempo transcurrido: 0 minutos, 6 segundos
```