Proyecto 1

Mario Andrade Vargas (201711906)
María José Ruíz García (202010184)
Juan Camilo Villamarin (201816898)

PROYECTO 1 | 11 OCTUBRE 2021

Agenda

Problema negocio
Descripción del requerimiento
Detalles actividad de minería de datos
Comprensión de los datos y preparación de los datos
Modelado y evaluación
Resultados

Problema del negocio

Se quiere saber cómo es el movimiento del bitcoin en el mercado según la opinión de los tweets en cada día de la semana, teniendo en cuenta palabras clave usadas en redes sociales sobre el bitcoin.

Descripción del requerimiento

Se requiere clasificar tweets sobre bitcoin, según unas palabras claves encontradas en un análisis de sentimientos previamente realizado, para ver cuales son los tweets que afectan positiva y negativamente el precio del bitcoin.

CLASIFICAR LOS TWEETS DE ACUERDO A LAS PALABRAS CLAVES ENCONTRADAS Y EL DÍA DE LA SEMANA DE LA PUBLICACIÓN DE ESTOS.

CLASIFICACIÓN

SUPPORT VECTOR MACHINES

HIPERPARAMETROS:

- KERNEL : LINEAL
- C = 0.1
- GAMMA = 0.1

ÁRBOL DE DECISIÓN

HIPERPARAMETROS:

-CRITERION: ENTROPY

-MAX_DEPTH: 4

-MIN_SAMPLES_SPLIT: 2

KNN

HIPERPARAMETROS:

-BEST P:1

- METRIC 'MANHATTAN'

-BEST N_NEIGHBOURS:9

Comprensión y preparación de datos

1. Se toma como muestra 50.000 datos
2. Se filtran los tweets por idioma para tener solo en inglés. Para esto se usó la librería “langdetect”.
3. Se transforman los datos de la columna Sentiment en una representación numérica donde Positive es 1 y Negative es 0
4. Para la columna Date se filtran las fechas inválidas.
5. Como parte del análisis vamos a tener en cuenta la importancia del día de la semana .Se asigna un valor de 0 a 6.
6. Se hace una columna que cuente las veces que aparece las palabras claves que fueron elegidas a partir de Royal Society Open Science

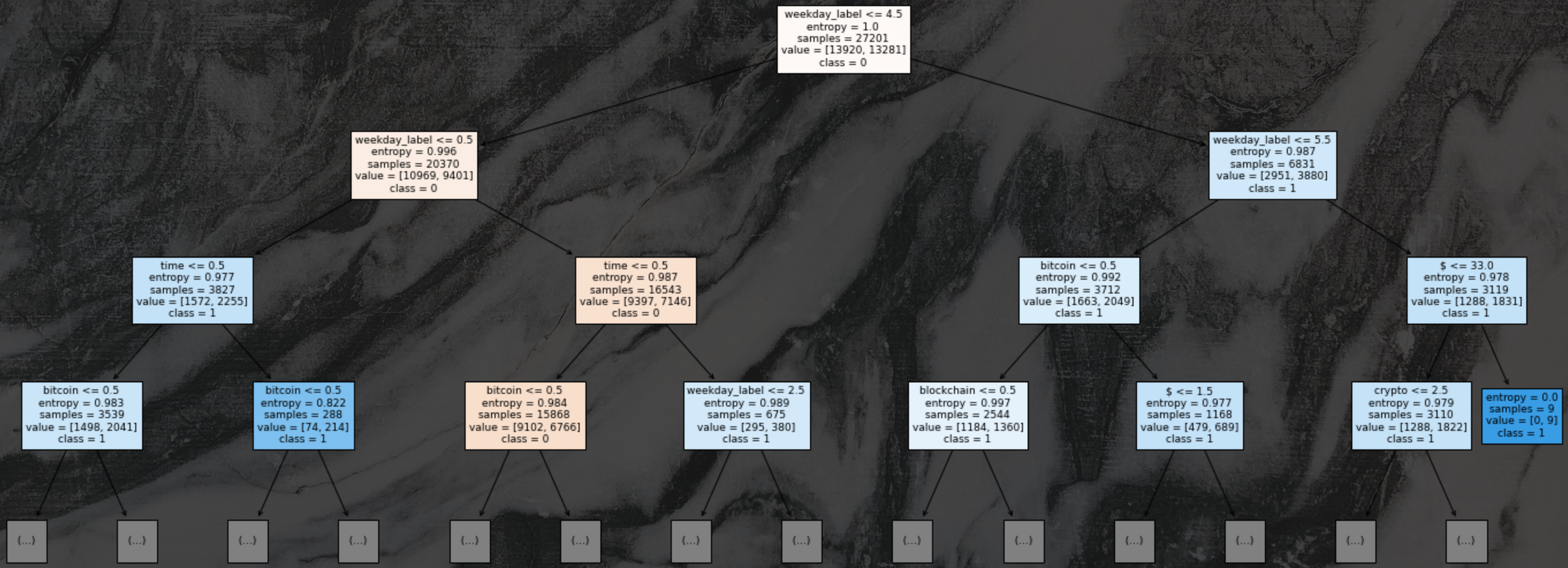
Modelado y evaluación

ÁRBOLES DE DECISIÓN

SUPPORT VECTOR MACHINES

KNN

Árboles de decisión



Support vector machines

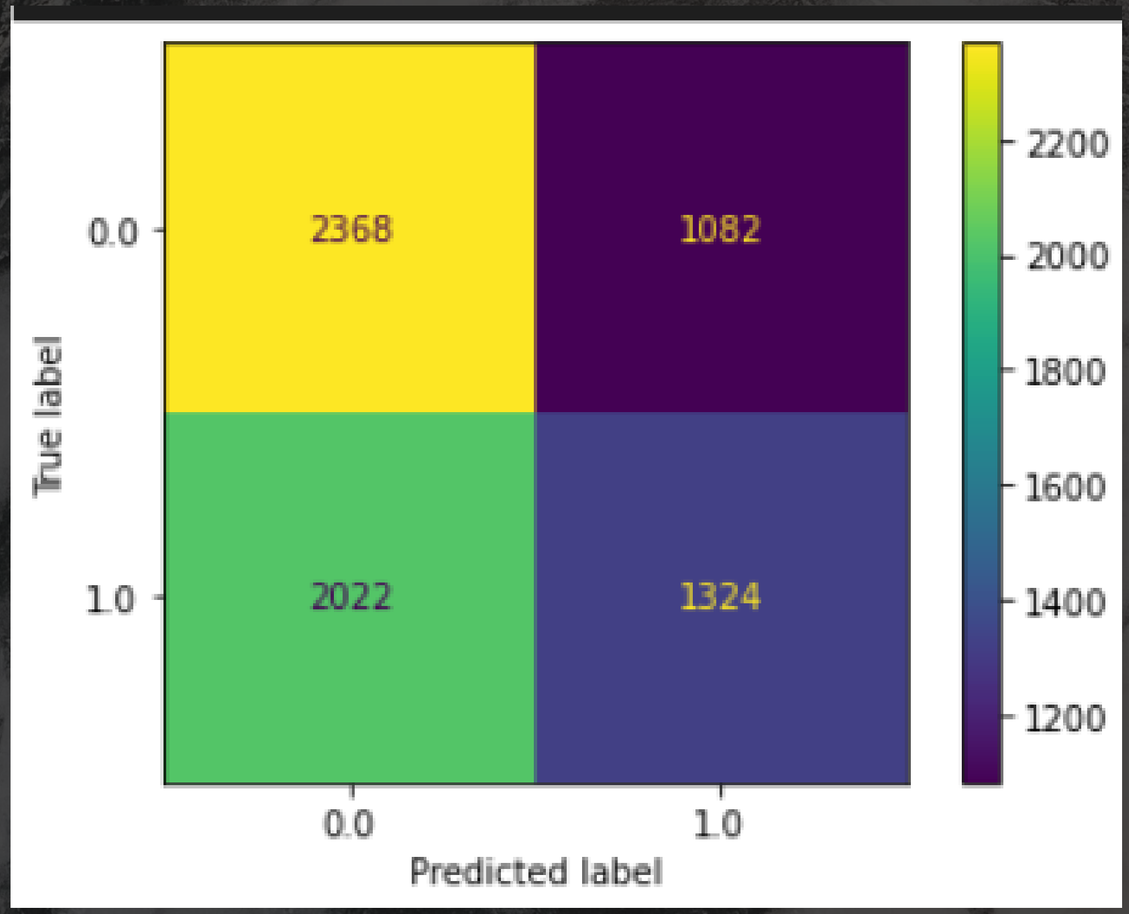
```
#Hiperparametros para probar
param_grid = {'C': [0.1,1, 10],
              'gamma': [0.1,0.01,0.001],
              'kernel': ['linear']}
```

	precision	recall	f1-score	support
0.0	0.50	0.94	0.65	3414
1.0	0.51	0.06	0.11	3397
accuracy			0.50	6811
macro avg	0.51	0.50	0.38	6811
weighted avg	0.51	0.50	0.38	6811

KNN

```
Best p: 1
Best n_neighbors: 9
```

	precision	recall	f1-score	support
0.0	0.54	0.69	0.60	3450
1.0	0.55	0.40	0.46	3346
accuracy			0.54	6796
macro avg	0.54	0.54	0.53	6796
weighted avg	0.54	0.54	0.53	6796



Resultados

El algoritmo de árboles de decisión, encontramos que este modelo tiene una precisión y recall de aproximadamente el 50% para la clasificación positiva.

Resultados

Algoritmo de Support Vector Machines, el resultado no fue tan bueno como en el modelo anterior; la precisión y especialmente el recall fueron muy bajos para la clasificación de tweets positivos

Resultados

El modelo con el algoritmo de K-means, las métricas presentan resultados similares a el modelo de árboles de decisión, tanto como para la clasificación de tweets positivos, como negativos.

Resultados

Teniendo en cuenta todo esto, el modelo que recomendamos para clasificar la influencia del tweet sobre el precio del bitcoin, es el que aplica el algoritmo de árboles de decisión.