Minería de Flujo de Datos

MÁSTER UNIVERSITARIO EN CIBERSEGURIDAD
E INTELIGENCIA DE DATOS

Alba Cruz Torres | Escuela de Doctorado y Estudios de Posgrado | 14/09/2022

INDICE

1. INTRODUCCIÓN

La **Minería de Flujo de Datos** (***Data Stream Mining***) es el proceso de extraer conocimiento un flujo de datos continuo, el cual es una secuencia ordenada de instancias que se puede leer una sola vez o una cantidad muy pequeña de veces, en un sistema limitado de capacidad de memoria y capacidad de almacenamiento.

A lo largo de esta práctica, se hará uso de la herramienta **MOA** (**M**assive **O**nline **A**nalysis), para realizar ejercicios de clasificación, regresión, *clustering*, detección de *outliers* o sistemas de recomendación con *data streams*.

En primer lugar, se ha escogido un conjunto de datos sintéticos, con el fin de entender mejor el comportamiento de la herramienta y aprender cómo interpretar los resultados. Este conjunto de datos sintéticos se denomina ***LED*** y tiene como objetivo predecir el dígito en un *display* de siete segmentos. El *dataset* tiene 7 atributos relacionados con la clase y 17 irrelevantes.

Luego, se ha hecho uso de un conjunto de datos real, para ver cómo cambian los resultados según los algoritmos predictivos que se utilicen e intentar dar respuesta a algunas preguntas de interés. Este *dataset* contiene datos sobre la evolución del precio de la electricidad a lo largo de 24 horas, recogidos por un mercado eléctrico del sureste australiano.

2. DATOS SINTÉTICOS

Como se comentó anteriormente, el primer flujo de datos con el que se va a trabajar es un conjunto de datos sintéticos que representan un *display* de 7 segmentos LED, como el que se observa a continuación.

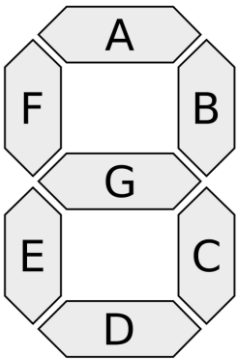

Ilustración 1: Display de 7 segmentos

Lo que se desea conseguir es predecir el número que se va a mostrar en dicho *display* según los atributos de entrada.

Los datos se recopilan en diferentes ficheros, pero para esta práctica se hará uso del fichero *led_w_500_n_0.1_175* que tiene el siguiente formato:

```
27
28  @attribute class {0,1,2,3,4,5,6,7,8,9}
29
30  @data
31  1,0,1,1,0,1,1,0,0,0,0,0,1,0,1,1,0,0,1,1,1,1,1,0,2
32  1,1,0,0,1,1,1,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,1,9
33  1,1,0,1,1,0,1,0,1,1,1,0,0,0,1,0,0,1,1,1,1,0,0,0,5
34  1,1,1,1,1,1,1,1,1,0,0,0,0,1,0,0,0,0,1,1,0,1,1,0,8
```

Ilustración 2: Formato led_w_500_n_0.175.arff

Lo primero que se va a hacer es cargar los datos en la herramienta **MOA** y evaluarlos con diferentes clasificadores, a ver con cual de ellos se consiguen mejores resultados. Como la propia descripción del *dataset* indica, se trata de un problema de clasificación con una variable categórica con 10 etiquetas: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Dentro de los algoritmos de clasificación disponibles en la

herramienta, se van a comparar los resultados de los siguientes: *Naive Bayes*, *K-Nearest Neighbors* y *Hoeffding Tree*.

2.1 NAIVE BAYES

El primero de los algoritmos de predicción que se va a utilizar es ***Naive Bayes***, que es un algoritmo de clasificación que se basa en las probabilidades. Este algoritmo se centra en el Teorema de Bayes, que se fundamenta en las probabilidades condicionales, es decir, la determinación de la probabilidad de que se produzca un acontecimiento en función de un acontecimiento que ya se ha producido.

El Teorema de Bayes se expresa con la siguiente fórmula:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Para poder establecer los requisitos necesarios para entrenar y evaluar el modelo, se han puesto los siguientes parámetros de configuración en **MOA**: en primer lugar, se ha decidido utilizar el evaluador de clasificación ***Intervealed Test Then Train***, el cual evalúa el modelo, que ha sido entrenado con el bloque de datos anterior, con el nuevo bloque de datos que le llega al modelo. Así, el modelo será evaluado por un nuevo bloque de entrada y acto seguido, entrenado por ese mismo bloque, para ser evaluado por el siguiente.

Después, se ha establecido el número de instancias que se van a estudiar que, en este caso, se ha establecido a 100.030, pues el archivo de datos *.arff* introducido está compuesto por ese número de instancias.

Asimismo, se ha establecido que los bloques de datos que van entrando al modelo sean de 1.000 datos cada uno. Dichos parámetros de configuración se pueden ver en la imagen que sigue:

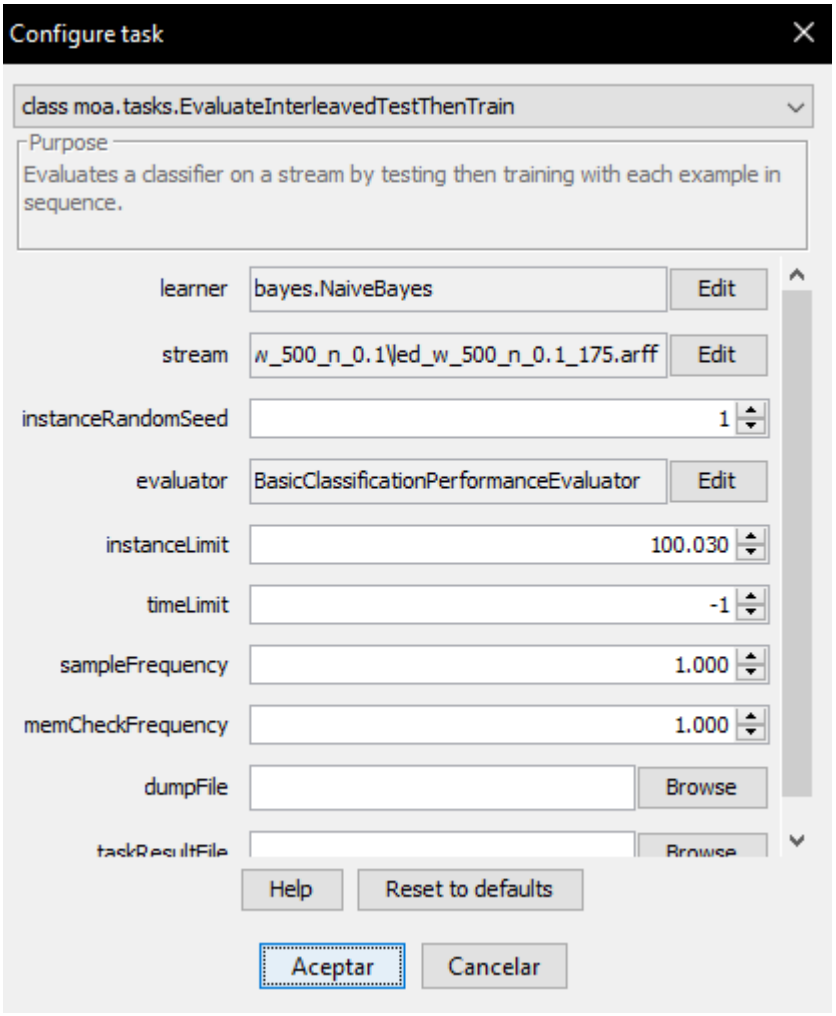

Ilustración 3: Parámetros de configuración de MOA con datos sintéticos

Acto seguido, se ejecuta el modelo que acabamos de establecer y se muestran los resultados de *accuracy*, que es una de las métricas que se utilizan para comprobar el buen funcionamiento de un modelo. Como se puede observar en la imagen a continuación, con el algoritmo de clasificación de *Naives Bayes*, se consigue un resultado final de *accuracy* del 70,33%, el cual es un porcentaje de clasificación bastante alto.

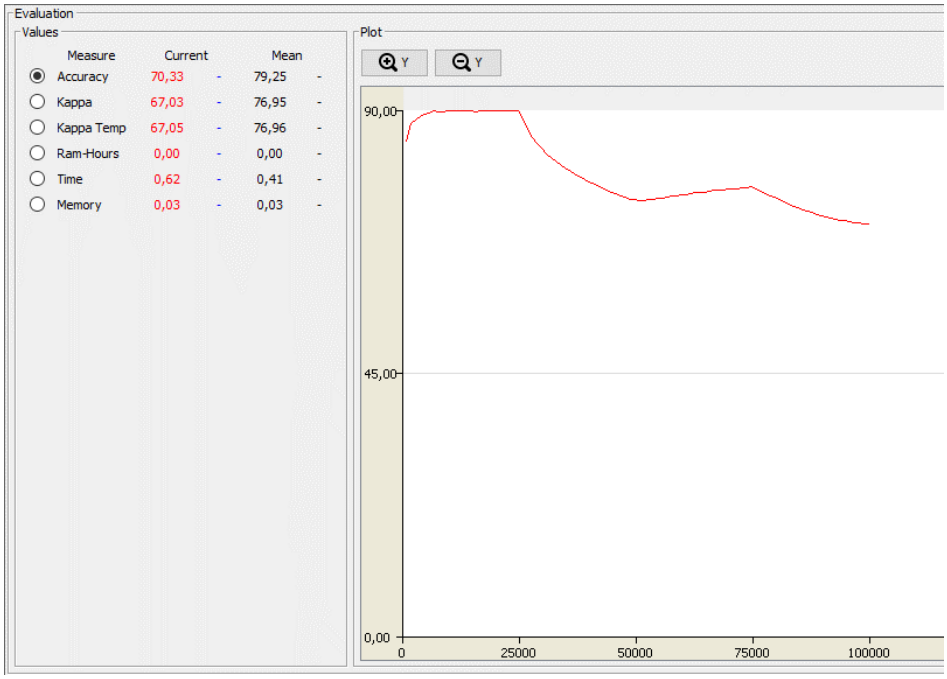

Ilustración 4: Gráfica de accuracy de Naive Bayes con datos sintéticos

2.2 ÁRBOLES DE HOEFFDING

A continuación, se procede a ejecutar el mismo proceso con los mismos parámetros, pero esta vez con el algoritmo de clasificación de Árboles de *Hoeffding*, los cuales son árboles de decisión con cotas de *Hoeffding*, es decir, cotas proporcionadas por la desigualdad de *Hoeffding* que son superiores a la probabilidad de que la suma de variables aleatorias se desvíe una cierta cantidad de su valor esperado. En esta ocasión, se ha conseguido un porcentaje de *accuracy* del 83.31%, que es un 12.98% mayor que el conseguido con el algoritmo de *Naive Bayes*. En la siguiente imagen, se puede ver la comparación de los resultados entre los dos modelos creados con los distintos algoritmos clasificadores.

Otro de los datos interesantes que se puede ver en la imagen es la media de los porcentajes de *accuracy* conseguidos en las evaluaciones de todos los bloques. Como se puede observar, con el algoritmo de Árboles de *Hoeffding*, se ha conseguido una media de accuracy del 84.07%, frente al 79.25% que se ha conseguido con el algoritmo de *Naive Bayes*.

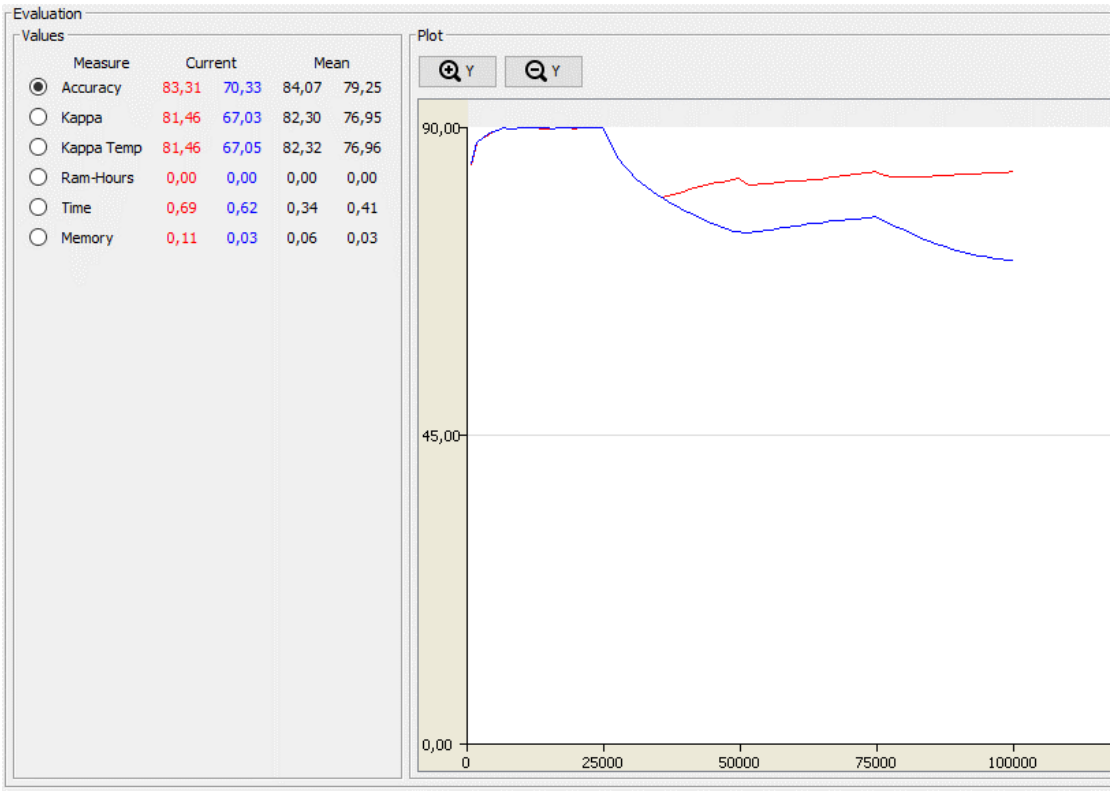

Ilustración 5: Comparación de accuracies Árbol de Hoeffding frente a Naive Bayes

2.2 K-NEAREST NEIGHBORS (KNN)

Lo siguiente que se ha hecho ha sido utilizar el algoritmo de los *k* vecinos más cercanos o **KNN** para realizar la tarea de clasificación. Este es un método de clasificación que se utiliza para calcular la distancia entre el ítem que se desea clasificar y el resto de los ítems del *dataset* de entrenamiento. Luego, selecciona los *k* elementos más cercanos, es decir, con menor distancia y, finalmente, realiza una votación de mayoría entre los *k* puntos. Por tanto, los de una clase o etiqueta que dominen decidirán la clasificación final.

Los resultados de clasificación con este algoritmo han sido de un 81.58% de *accuracy* que, en este caso, es inferior al conseguido con el algoritmo anterior, como se puede observar en la siguiente imagen. La media de clasificación sigue siendo menor que la del algoritmo de Árboles de *Hoeffding*, es decir, con este algoritmo se consigue una media del 80.92% frente a una media del 84.07%.

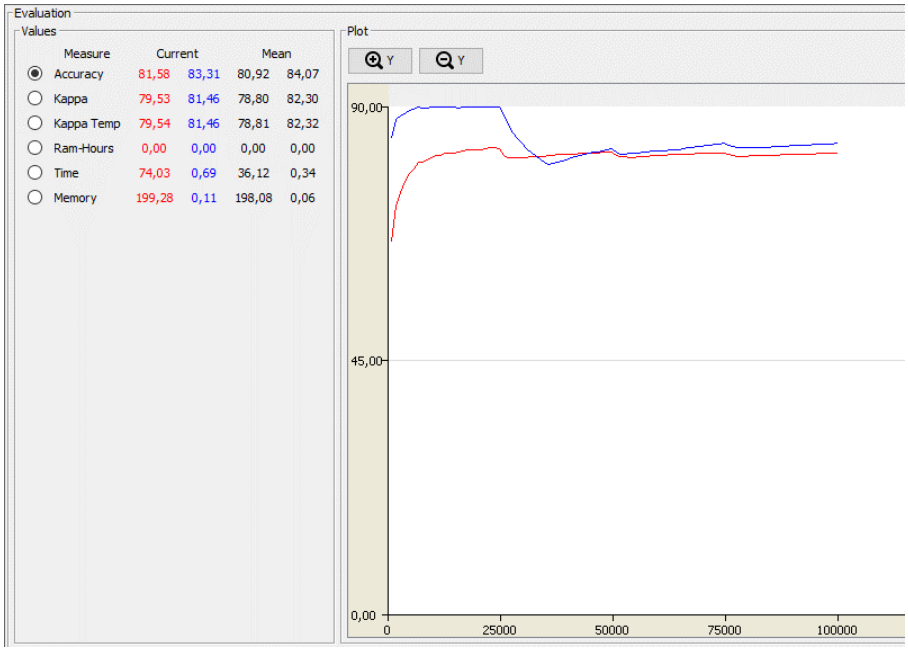

Ilustración 6: Comparación de accuracies de KNN frente a Árbol de Hoeffding

Si se hace una comparación del algoritmo KNN con el algoritmo de *Naive Bayes*, se puede ver que también se consiguen mejores resultados con el actual algoritmo de clasificación que con el clasificador de probabilidades. Se tiene un porcentaje de *accuracy* del 81.58% con el *KNN*, mientras que con el modelo de *Naive Bayes* se tiene un 70,33% de *accuracy* final.

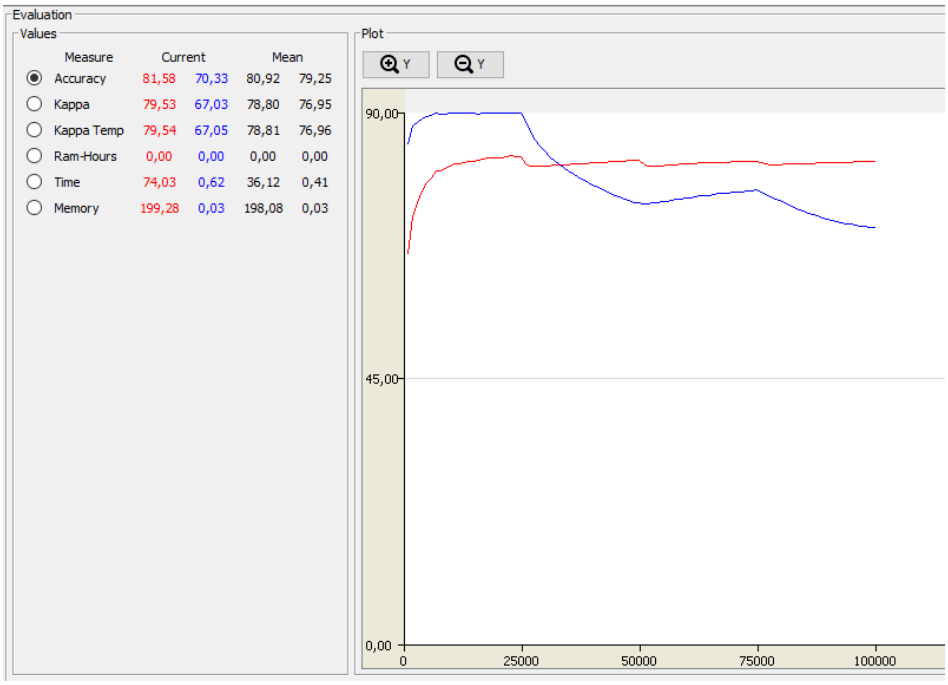

Ilustración 7: Comparación de accuracies de KNN frente a Naive Bayes

3. DATOS REALES

Después de haber realizado pruebas de la herramienta con los datos sintéticos, se procede a realizar un análisis con un conjunto de datos reales. Para este proyecto, se ha decidido utilizar el conjunto llamado ***Electricity***, que se ha descargado de la propia página de la herramienta MOA:

https://moa.cms.waikato.ac.nz/datasets/

Este *dataset* está constituido por datos que han sido recolectados del ***Australian New South Wales Electricity Market***. En este mercado los precios de la electricidad no son fijos y se ven afectados por la oferta y la demanda del mercado. Los precios son establecidos cada cinco minutos. La variable de clase de este conjunto de datos identifica el cambio del precio en relación con un promedio móvil de las últimas 24 horas y se puede obtener dos valores: *UP* o *DOWN*, que hacen referencia a si el precio ha subido o bajado en cuestión de un día. El formato del archivo es el siguiente:

```
3   @attribute date numeric
4   @attribute day {1,2,3,4,5,6,7}
5   @attribute period numeric
6   @attribute nswprice numeric
7   @attribute nswdemand numeric
8   @attribute vicprice numeric
9   @attribute vicdemand numeric
10  @attribute transfer numeric
11  @attribute class {UP,DOWN}
12
13  @data
14  0,2,0,0.056443,0.439155,0.003467,0.422915,0.414912,UP
15  0,2,0.021277,0.051699,0.415055,0.003467,0.422915,0.414912,UP
16  0,2,0.042553,0.051489,0.385004,0.003467,0.422915,0.414912,UP
17  0,2,0.06383,0.045485,0.314639,0.003467,0.422915,0.414912,UP
18  0,2,0.085106,0.042482,0.251116,0.003467,0.422915,0.414912,DOWN
```

Ilustración 8: Formato elecNormNew.arff

Por tanto, se trata nuevamente de un problema de clasificación, pues se desea clasificar si al final del día los precios de la electricidad subirán o bajarán. Para ello, se hará una comparación de los resultados de los modelos creados con los

distintos algoritmos de clasificación que se utilizaron en el apartado 2. Para esta tarea de clasificación, los parámetros que se van a utilizar son los siguientes:

El número de instancias será de 45.325, pues son las que constituyen el conjunto de datos. Por otro lado, el tamaño de los bloques que van a ir alimentando el modelo cada vez será de 100 datos por bloque. Finalmente, se utilizará el mismo evaluador que en el apartado anterior, es decir, el *IntervealedTestThenTrain*.

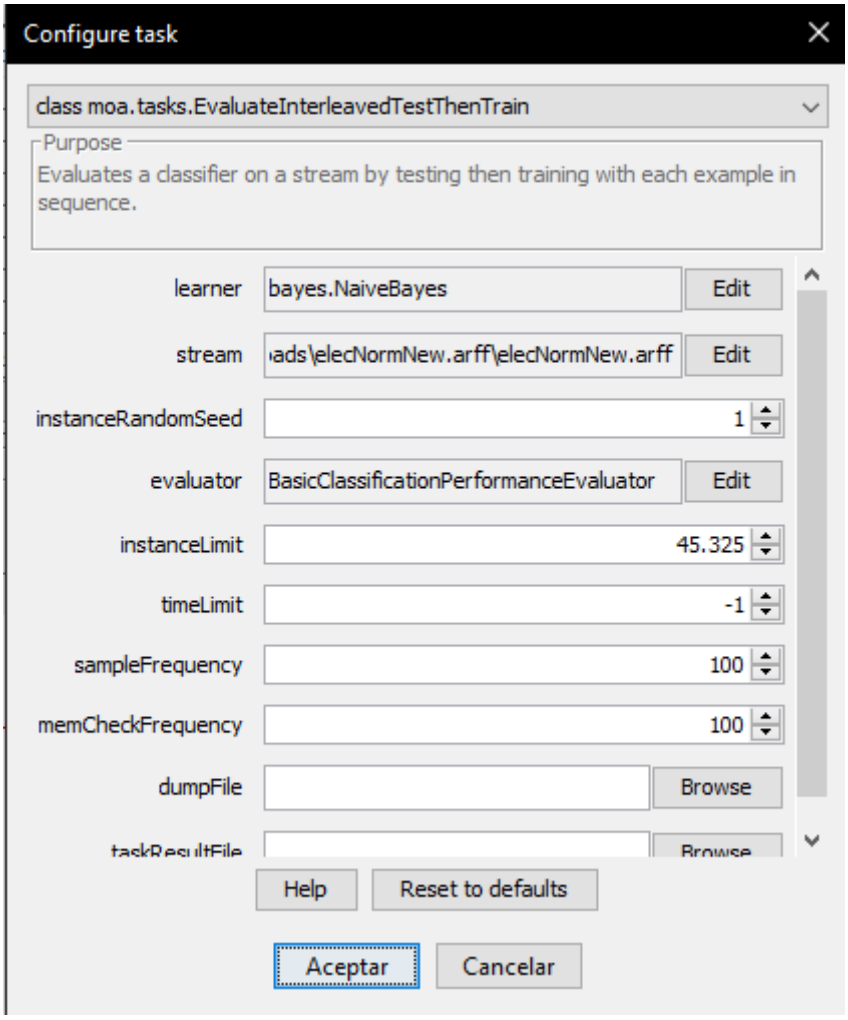

Ilustración 9: Parámetros de configuración de MOA con datos reales

3.1 NAIVE BAYES

El primer modelo del que se van a analizar los resultados es el creado con el algoritmo *Naive Bayes*. En este primer caso, se ha conseguido un porcentaje de *accuracy* del 73.36%, que se corresponde con un buen porcentaje de clasificación, aunque seguramente podría ser mejor.

Asimismo, como se puede observar en la siguiente imagen, la media de los porcentajes de *accuracy* conseguidos con el modelo en cada bloque de datos es del 76.18%.

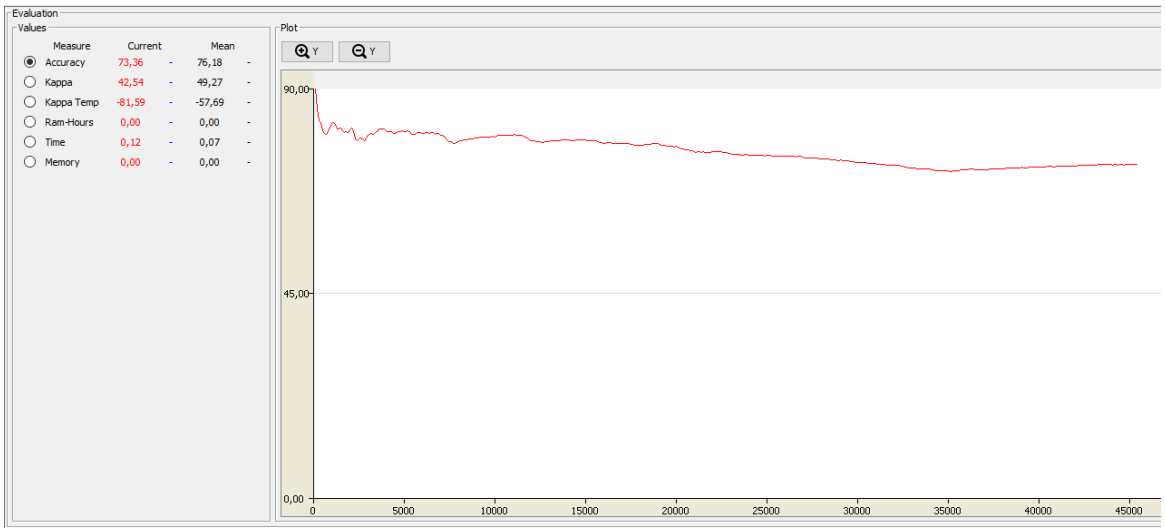

Ilustración 10: Gráfica de accuracy de Naive Bayes con datos reales

3.2 ÁRBOLES DE HOEFFDING

El siguiente algoritmo que se ha utilizado es el de Árboles de *Hoeffding*. Con dicho algoritmo, se consigue una *accuracy* del 79.20%, la cual es mayor que la conseguida con el algoritmo de *Naive Bayes*, con una diferencia de 5.84. Además, la media de los porcentajes de *accuracy* conseguida con este algoritmo es del 83.18%, que sigue siendo mayor que la media anterior resultante.

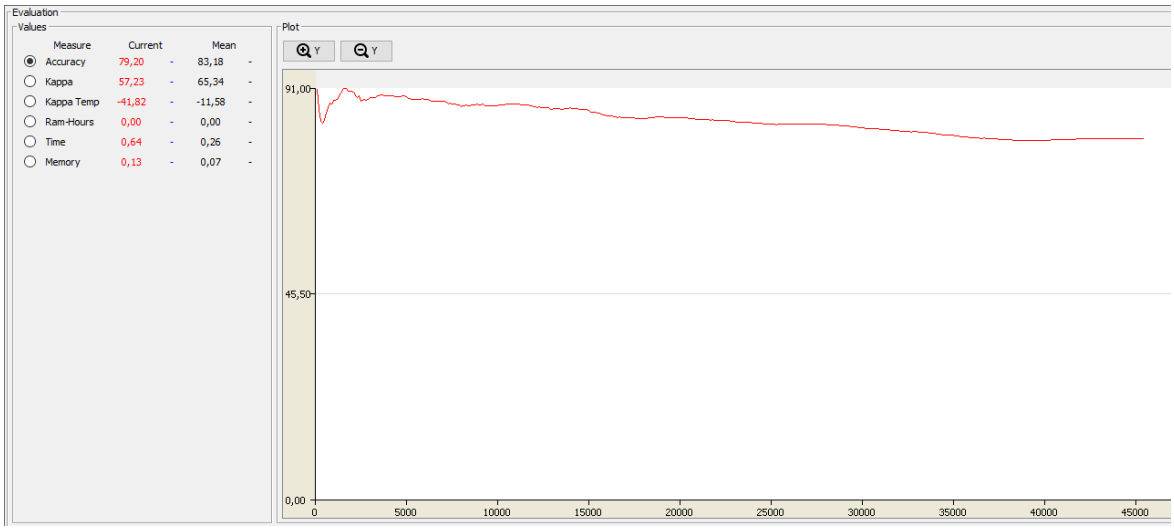

Ilustración 11: Gráfica de accuracy de Árbol de Hoeffding con datos reales

3.3 K-NEAREST NEIGHBORS (KNN)

Finalmente, se ha hecho uso del algoritmo de KNN para la obtención y análisis de los resultados de clasificación. Con este algoritmo se consigue un porcentaje de *accuracy* del 82.80%, que supera los dos porcentajes conseguidos con los algoritmos anteriores. Este algoritmo supera en un 3,6% al porcentaje de *accuracy* conseguido con el Árbol de *Hoeffding* y en un 9,44% al conseguido con *Naive Bayes*. Finalmente, con este algoritmo se consigue una media de los porcentajes del *accuracy* del 82.44% que, en este caso, resulta ser menor que la conseguida con el Árbol de *Hoeffding*, que fue del 83.18%.

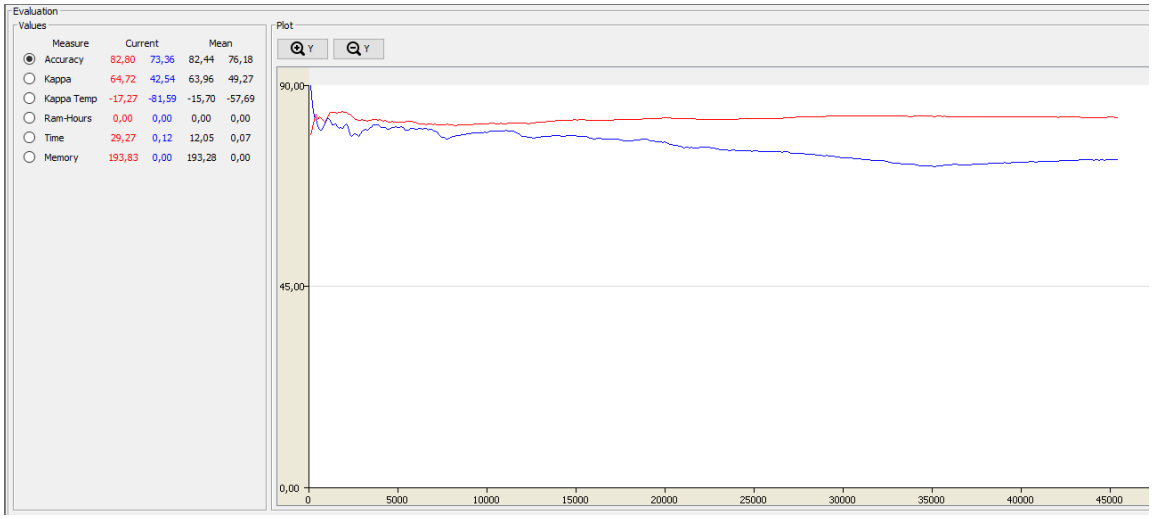

Ilustración 12: Comparación de accuracies de KNN frente a Árbol de Hoeffding

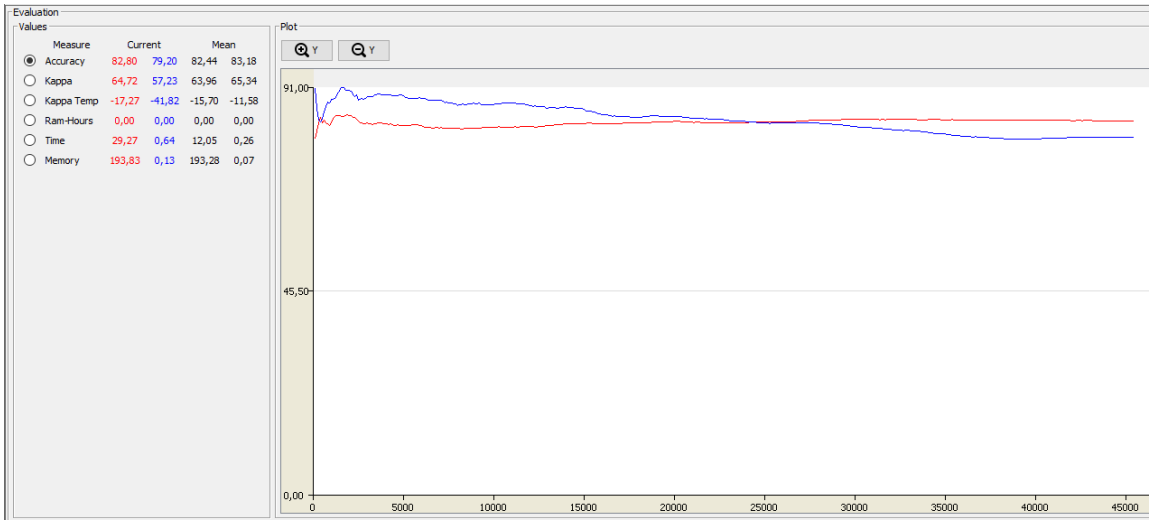

Ilustración 13: Comparación de accuracies de KNN frente a Naive Bayes

4.CONCLUSIONES

A lo largo de este proyecto se ha realizado un análisis del flujo de datos sobre la fluctuación del precio de la electricidad, con datos que han sido recogidos cada cinco minutos por el Mercado Eléctrico Australiano de Nueva Gales del Sur. Asimismo, se han utilizado tres algoritmos de clasificación para llevar a cabo la tarea de etiquetar si a lo largo de 24 horas el precio ha subido (*UP*) o ha bajado (*DOWN*). Dichos algoritmos de clasificación han sido: *Naive Bayes*, Árboles de *Hoeffding* y *K-Nearest Neighbors*. Los resultados de clasificación de los modelos, haciendo hincapié en la medida de la *accuracy* son los siguientes:

	% *accuracy* (último bloque)	**Media de % *accuracies***
Naive Bayes	73.36%	76.18%
Árboles de *Hoeffding*	79.20%	83.18%
K-Nearest Neighbors	82.20%	82.44%

Como se puede observar en la tabla, los mejores resultados conseguidos en el último bloque de datos pasado al modelo han sido con el algoritmo de clasificación KNN, con un 82.20% de *accuracy*, seguido por el algoritmo de Árboles de *Hoeffding*, con un porcentaje de *accuracy* del 79.20% y, finalmente, el algoritmo de *Naive Bayes*, con el que se ha conseguido un porcentaje del 73.36%.

Si bien se han conseguido mejores resultados de *accuracy* en el último bloque de datos, si se centra el foco en la media de los porcentajes de *accuracy* conseguidos por el modelo con los demás bloques de datos, se puede ver claramente que se consigue una media mayor, concretamente del 83.18%, con el algoritmo de Árboles de *Hoeffding*, frente al 82.44% de media conseguido por el KNN. Esto quiere decir que, aunque en el último bloque de datos se haya conseguido un mejor porcentaje de *accuracy*, no significa que el modelo haya clasificado mejor en los demás bloques de datos. Es decir, en general, el modelo creado con el algoritmo de KNN clasificó mejor que el modelo con el algoritmo de Árboles de *Hoeffding*.