**Universidad Tecnológica de La Habana “José Antonio Echeverría”**

**Facultad de Ingeniería Informática**



**Clasificación para kr vs kp y Agrupamiento para plants**

***Tarea final de Inteligencia Artifical***

**Autores:**

Javier García Hernández

Alejandro Roca Bernal

La Habana, Cuba

Noviembre,2023

**Resumen**

Este informe se centra en la aplicación de técnicas de minería de datos, específicamente clasificación y agrupamiento, en dos conjuntos de datos distintos. Se utiliza el algoritmo ID3 para la clasificación en el conjunto de datos de ajedrez kr-vs-kp y el algoritmo K-means para el agrupamiento en el conjunto de datos de plantas. A través de este estudio, se busca descubrir patrones y relaciones útiles en los datos que puedan proporcionar una visión significativa. Los resultados obtenidos de la minería de datos se presentan y se discuten en detalle en este informe.

Palabras claves: Minería de datos, Clasificación, Agrupamiento, Algoritmo ID3, Algoritmo K-means

**Abstract**

This report focuses on the application of data mining techniques, specifically classification and clustering, on two distinct datasets. The ID3 algorithm is used for classification on the kr-vs-kp chess dataset, and the K-means algorithm is used for clustering on the plant dataset. Through this study, we aim to uncover useful patterns and relationships in the data that can provide meaningful insight. The results obtained from data mining are presented and discussed in detail in this report.

Keywords: Data Mining, Classification, Clustering, ID3 Algorithm, K-means Algorithm

**Introducción**

En la era de la información, los datos se han convertido en una de las principales fuentes de conocimiento. Sin embargo, la cantidad masiva de datos disponibles puede ser abrumadora y, a menudo, es difícil de interpretar sin el uso de técnicas especializadas. Aquí es donde la minería de datos entra en juego[1-4].

La minería de datos es un proceso que utiliza una variedad de técnicas de análisis y algoritmos para descubrir patrones y relaciones en los datos que pueden ser de utilidad para la toma de decisiones[1]. Este informe, se centra en dos de las técnicas más comunes utilizadas en la minería de datos: la clasificación y el agrupamiento.

Se aplica el algoritmo ID3 para la clasificación en el conjunto de datos de ajedrez kr-vs-kp, y el algoritmo K-means para el agrupamiento en el conjunto de datos de plantas. A través de este informe, se explora en detalle estos algoritmos, se describen los conjuntos de datos, se presentan los resultados de la minería de datos y finalmente, se discuten conclusiones para estos.

**¿Qué es la clasificación y para qué se utiliza?**

La clasificación es una técnica de minería de datos que se utiliza para asignar una etiqueta o categoría a un conjunto de datos. Esta técnica se basa en el aprendizaje automático, que permite a los algoritmos identificar patrones en los datos y utilizarlos para predecir la categoría de nuevos datos.

La clasificación se utiliza en una amplia gama de aplicaciones, entre las que se incluyen:

1. **Predicción de comportamiento**: la clasificación se puede utilizar para predecir el comportamiento de los clientes, como si van a comprar un producto o si van a cancelar su suscripción.
2. **Identificación de fraude**: la clasificación se puede utilizar para identificar transacciones fraudulentas, como compras realizadas con tarjetas de crédito robadas.
3. **Clasificación de imágenes**: la clasificación se puede utilizar para clasificar imágenes en categorías, como perros, gatos o personas.

El proceso de clasificación en minería de datos se puede dividir en las siguientes etapas:

1. **Preparación de los datos**: esta etapa consiste en limpiar los datos y prepararlos para el proceso de clasificación.
2. **Selección del algoritmo**: esta etapa consiste en seleccionar el algoritmo de clasificación más adecuado para el conjunto de datos.
3. **Entrenamiento del modelo**: esta etapa consiste en entrenar el algoritmo de clasificación con un conjunto de datos de entrenamiento.
4. **Evaluación del modelo**: esta etapa consiste en evaluar el rendimiento del modelo con un conjunto de datos de evaluación.
5. **Implementación del modelo**: esta etapa consiste en implementar el modelo para su uso en producción.

**¿Qué es el agrupamiento y para qué se utiliza?**

El agrupamiento es una técnica que se utiliza para dividir un conjunto de datos en grupos de objetos similares. Esta técnica se basa en la identificación de patrones en los datos que permiten agrupar a los objetos que comparten características comunes.

El agrupamiento se utiliza en una amplia gama de aplicaciones, entre las que se incluyen:

1. **Segmentación de clientes**: el agrupamiento se puede utilizar para dividir a los clientes en grupos con características comunes, como edad, género, intereses o comportamiento de compra. Esto puede ayudar a las empresas a desarrollar estrategias de marketing más efectivas.
2. **Análisis de mercado**: el agrupamiento se puede utilizar para identificar grupos de productos o servicios con características comunes. Esto puede ayudar a las empresas a desarrollar nuevas estrategias de marketing o productos.
3. **Análisis de datos científicos**: el agrupamiento se puede utilizar para identificar grupos de datos con características comunes. Esto puede ayudar a los científicos a comprender mejor los datos y a desarrollar nuevas teorías.

El proceso de agrupamiento en minería de datos se puede dividir en las siguientes etapas:

1. **Preparación de los datos**: esta etapa consiste en limpiar los datos y prepararlos para el proceso de agrupamiento.
2. **Selección del algoritmo**: esta etapa consiste en seleccionar el algoritmo de agrupamiento más adecuado para el conjunto de datos.
3. **Ejecución del algoritmo**: esta etapa consiste en ejecutar el algoritmo de agrupamiento para dividir el conjunto de datos en grupos.
4. **Interpretación de los resultados**: esta etapa consiste en interpretar los resultados del agrupamiento para identificar los patrones en los datos.

**Descripción de los datos**

1. kr-vs-kp:

* Título: Final de partida de ajedrez: Rey+Torre contra Rey+Peón en a7 (generalmente abreviado KRKPA7). El peón en a7 significa que está a una casilla de distancia de reinar. Es el lado del Rey+Torre (blanco) el que debe moverse.
* Numero de instancias: 3196
* Numero de atributos: 36
* Resúmenes de atributos: Clases (2): -- Las blancas pueden ganar ("ganaron") y las blancas no pueden ganar ("ahora ganan")
* Atributos faltantes: ninguno.
* Class Distribution: In 1669 of the positions (52%), White can win. In 1527 of the positions (48%), White cannot win.

1. Plants:

* Características del conjunto de datos: multivariado
* Área temática: Biología
* Tareas asociadas: Agrupación
* Tipo de característica: Categórico
* Número de Instancias : 34781
* Número de Características: 70
* ¿Tiene valores faltantes?: Sí

**Algoritmo ID3**

El algoritmo ID3 es un algoritmo de aprendizaje supervisado que se utiliza para construir árboles de decisión. El algoritmo funciona dividiendo los datos de entrenamiento en subconjuntos más pequeños, cada uno de los cuales es más homogéneo que el conjunto de datos original. El algoritmo continúa dividiendo los datos hasta que cada subconjunto contenga solo un ejemplo o hasta que no se puedan encontrar más atributos que proporcionen una división significativa.

El algoritmo ID3 utiliza la ganancia de información para determinar qué atributo utilizar para dividir los datos. La ganancia de información es una medida de la cantidad de información que se puede obtener al saber el valor de un atributo. El algoritmo ID3 elige el atributo con la mayor ganancia de información para dividir los datos.

**Algoritmo Kmeans**

El algoritmo K-means es un algoritmo de aprendizaje no supervisado que se utiliza para agrupar datos en k grupos. El algoritmo funciona seleccionando k centroides iniciales al azar, asignando cada punto de datos al centroide más cercano y luego actualizando los centroides para que sean el punto medio de los puntos de datos asignados a ellos. El algoritmo continúa iterando este proceso hasta que los centroides no cambien o hasta que se alcance un número máximo de iteraciones.

**Resultados de la minería:**

**Referencias bibliográficas**

[1] J. Han, Kamber, M., Pei, J., *Data mining: concepts and techniques*, 3rd ed. Morgan Kaufmann, 2011.

[2] J. Hernández Orallo, *Introducción a la Minería de Datos*. 2004.

[3] P.-N. S. Tan, M.; Kumar, V., Introduction to Data Mining. Pearson Education India, 2016.

[4] I. H. Witten, Frank, E., Hall, M. A., Pal, C. J., Data mining: practical machine learning tools and techniques, 4th ed. Morgan Kaufmann, 2016.