UNIVERSIDAD AUTÓNOMA METROPOLITANA
UNIDAD IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

"USO DE MÉTRICAS GENERALIZADAS EN CLASIFICADORES"

Tesis que presenta:
Carlos Alberto Hernández Nava

Matrícula: 2183802157

Para obtener el grado de
Maestro en Ciencias y Tecnologías de la Información

Directores: Dr. Pedro Lara Velázquez
Dra. Hérica Sánchez Larios

Jurado:
Presidente: Dra. Hérica Sánchez Larios
Secretario: Dr. Eric Alfredo Rincón García
Vocal: Dr. Roman Anselmo Mora Gutiérrez

Iztapalapa, Ciudad de México, febrero 2021

Resumen

En inteligencia artificial existen diversos campos de aplicación, entre ellos se encuentran los clasificadores, que asignan elementos a una categoría, y son básicamente de dos tipos: supervisados y no supervisados. Este trabajo se centra en los no supervisados, específicamente en un algoritmo clasificador basado en el problema de coloración de gráficas suaves y uno basado en K-medias. El primero, dado un gráfico completo con ponderaciones en sus aristas, minimiza la suma de las penalizaciones entre los vértices con el mismo color. El algoritmo K-medias utilizado, es uno de los clasificadores más comunes, el cual realiza un agrupamiento con base en la distancia más corta entre cada elemento y alguno de los centroides que fueron seleccionados aleatoriamente, éste se actualiza de manera iterativa con los elementos asignados en ese grupo.

Las distancias euclidiana y euclidiana cuadrática son las más utilizadas en la mayoría de las investigaciones que usan sistemas clasificadores, pero no significa que sean con las que se obtienen los mejores resultados. En este trabajo se clasifican las instancias más comunes empleando la distancia Minkowski de orden superior, primeramente con valores enteros, a continuación con valores decimales, y una vez encontrados los valores más prometedores, se explora la región con cambios de una centésima. Finalmente, se realizaron experimentos con una combinación lineal de métricas, es decir un híbrido, todo lo anterior con el objetivo de observar el comportamiento de dos distintos clasificadores con diversas métricas y semimétricas (que satisfacen la definición de una métrica con excepción de la desigualdad del triángulo), y la precisión que alcanzan para cada base de datos.

Agradecimientos

En primer lugar, a mi madre Judith Laura Nava Vite, que me enseño todo de la vida, y me da su amor y apoyo, gracias por haberme impulsado en esta meta y por ser mi ejemplo a seguir. Todo se lo debo a ella.

A Victoria Amairani Romero Zuzuarregui, por su apoyo, paciencia, cariño y amor incondicional, gracias por ser parte de mi vida, Te amo.

A mis hermanas Mariana Esmeralda Mandujano Nava y Laura Itzel Mandujano Nava, gracias por su apoyo, consejos y el tiempo que pasamos juntos durante todo este proyecto.

A mi padre Abraham Mandujano Rodríguez, por su orientación y valiosos consejos sobre la vida a lo largo de los años.

Al Dr. Pedro Lara Velázquez por haber aceptado ser mi asesor, por su apoyo, sus enseñanzas y su dedicación a este proyecto de investigación.

A la Dra. Hérica Sánchez Larios por haber aceptado ser mi asesora, por sus valiosas aportaciones, así como las revisiones de este documento.

Al Posgrado en Ciencias y Tecnologías de la Información de la Universidad Autónoma Metropolitana y todo el personal que lo conforma por permitirme crecer profesionalmente.

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico brindado para poder cumplir con esta meta académica.

Por último, gracias a todos los familiares y amigos, que directa o indirectamente contribuyeron a lograr esta meta.

Índice

Introducción **9**

Objetivos **12**

Capítulo 1. Marco Teórico **13**
1.1 Métricas . . . 13
1.2 Clasificadores . . . 15
1.3 Procesos de agrupamiento . . . 16
1.4 Coloración de gráficas . . . 16
1.5 Coloración de gráficas suaves . . . 17
1.6 K-medias . . . 18
1.7 Diseño de experimentos . . . 20

Capítulo 2. Bases de datos de prueba **22**
2.1 Iris . . . 22
2.2 Wine . . . 22
2.3 Zoo . . . 23
2.4 Stone Flakes . . . 24
2.5 Titanic . . . 25
2.6 Bach . . . 26
2.7 CelebA . . . 26

Capítulo 3. Métricas y algoritmos utilizados **28**
3.1 Distancia Minkowski . . . 29
3.2 Distancia Euclidiana . . . 30
3.3 Distancia Manhattan . . . 31
3.4 Semimétricas . . . 32
3.5 Distancia Híbrida . . . 33
3.6 Preparación de las bases de datos . . . 34
3.7 Pseudocódigo de CGS . . . 36
3.8 Pseudocódigo de K-medias . . . 37

Capítulo 4. Análisis de resultados **38**
4.1 Resultados con $p > 1$. . . . 38
4.2 Resultados con $0 < p < 1$ . . . 43
4.3 Resultados con $0.88 \leq p \leq 1.02$ . . . 48
4.4 Resultados de la distancia híbrida . . . 55
4.5 Resultados finales . . . 63

Conclusiones **64**

Apéndice A. Diagramas de caja **65**
A.1. Diagramas de caja con valores de $p > 1$. . . . 65
A.2. Diagramas de caja con valores $0 < p < 1$. . . . 72

A.3. Diagramas de caja con valores $0.88 \leq p \leq 1.02$. 79
A.4. Diagramas de caja con distancia híbrida. 86

Referencias **93**

Índice de tablas

Tabla 1 Medidas de similaridad y disimilaridad para características cuantitativas. . . . 14
Tabla 2 Precisión de CGS en Iris con $p > 1$. . . . 38
Tabla 3 Precisión de K-medias en Iris con $p > 1$. . . . 38
Tabla 4 Precisión de CGS en Wine con $p > 1$. . . . 39
Tabla 5 Precisión de K-medias en Wine con $p > 1$. . . . 39
Tabla 6 Precisión de CGS en Zoo con $p > 1$. . . . 39
Tabla 7 Precisión de K-medias en Zoo con $p > 1$. . . . 40
Tabla 8 Precisión de CGS en Stone Flakes con $p > 1$. . . . 40
Tabla 9 Precisión de K-medias en Stone Flakes con $p > 1$. . . . 40
Tabla 10 Precisión de CGS en Titanic con $p > 1$. . . . 41
Tabla 11 Precisión de K-medias en Titanic con $p > 1$. . . . 41
Tabla 12 Precisión de CGS en Bach con $p > 1$. . . . 41
Tabla 13 Precisión de K-medias en Bach con $p > 1$. . . . 42
Tabla 14 Precisión de CGS en Celeb con $p > 1$. . . . 42
Tabla 15 Precisión de K-medias en Celeb con $p > 1$. . . . 42
Tabla 16 Precisión de CGS en Iris con $0 < p < 1$. . . . 43
Tabla 17 Precisión de K-medias en Iris con $0 < p < 1$. . . . 43
Tabla 18 Precisión de CGS en Wine con $0 < p < 1$. . . . 44
Tabla 19 Precisión de K-medias en Wine con $0 < p < 1$. . . . 44
Tabla 20 Precisión de CGS en Zoo con $0 < p < 1$. . . . 44
Tabla 21 Precisión de K-medias en Zoo con $0 < p < 1$. . . . 45
Tabla 22 Precisión de CGS en Stone Flakes con $0 < p < 1$. . . . 45
Tabla 23 Precisión de K-medias en Stone Flakes con $0 < p < 1$. . . . 45
Tabla 24 Precisión de CGS en Titanic con $0 < p < 1$. . . . 46
Tabla 25 Precisión de K-medias en Titanic con $0 < p < 1$. . . . 46
Tabla 26 Precisión de CGS en Bach con $0 < p < 1$. . . . 46
Tabla 27 Precisión de K-medias en Bach con $0 < p < 1$. . . . 47
Tabla 28 Precisión de CGS en Celeb con $0 < p < 1$. . . . 47
Tabla 29 Precisión de K-medias en Celeb con $0 < p < 1$. . . . 47
Tabla 30 Precisión de CGS en Iris con $0.88 \leq p \leq 1.02$. . . . 48
Tabla 31 Precisión de K-medias en Iris con $0.88 \leq p \leq 1.02$. . . . 49
Tabla 32 Precisión de CGS en Wine con $0.88 \leq p \leq 1.02$. . . . 49
Tabla 33 Precisión de K-medias en Wine con $0.88 \leq p \leq 1.02$. . . . 50
Tabla 34 Precisión de CGS en Zoo con $0.88 \leq p \leq 1.02$. . . . 50
Tabla 35 Precisión de K-medias en Zoo con $0.88 \leq p \leq 1.02$. . . . 51
Tabla 36 Precisión de CGS en Stone Flakes con $0.88 \leq p \leq 1.02$. . . . 51
Tabla 37 Precisión de K-medias en Stone Flakes con $0.88 \leq p \leq 1.02$. . . . 52
Tabla 38 Precisión de CGS en Titanic con $0.88 \leq p \leq 1.02$. . . . 52
Tabla 39 Precisión de K-medias en Titanic con $0.88 \leq p \leq 1.02$. . . . 53
Tabla 40 Precisión de CGS en Bach con $0.88 \leq p \leq 1.02$. . . . 53
Tabla 41 Precisión de K-medias en Bach con $0.88 \leq p \leq 1.02$. . . . 54
Tabla 42 Precisión de CGS en Celeb con $0.88 \leq p \leq 1.02$. . . . 54
Tabla 43 Precisión de K-medias en Celeb con $0.88 \leq p \leq 1.02$. . . . 55

Tabla 44 Precisión de CGS en Iris con distancia híbrida. 56
Tabla 45 Precisión de K-medias en Iris con distancia híbrida. 56
Tabla 46 Precisión de CGS en Wine con distancia híbrida. 57
Tabla 47 Precisión de K-medias en Wine con distancia híbrida. 57
Tabla 48 Precisión de CGS en Zoo con distancia híbrida. 58
Tabla 49 Precisión de K-medias en Zoo con distancia híbrida. 58
Tabla 50 Precisión de CGS en Stone Flakes con distancia híbrida. 59
Tabla 51 Precisión de K-medias en Stone Flakes con distancia híbrida. 59
Tabla 52 Precisión de CGS en Titanic con distancia híbrida. 60
Tabla 53 Precisión de K-medias en Titanic con distancia híbrida. 60
Tabla 54 Precisión de CGS en Bach con distancia híbrida. 61
Tabla 55 Precisión de K-medias en Bach con distancia híbrida. 61
Tabla 56 Precisión de CGS en Celeb con distancia híbrida. 62
Tabla 57 Precisión de K-medias en Celeb con distancia híbrida. 62
Tabla 58 Métricas con el mejor desempeño en CGS. 63
Tabla 59 Métricas con el mejor desempeño en K-medias. 63

Índice de figuras

Figura 1 Distancia euclidiana entre dos puntos. 30
Figura 2 Comparación entre distancia euclidiana y distancia Manhattan. 31
Figura 3 Círculos unitarios con varios valores de p utilizando distancia Minkowski. . 32
Figura 4 Precisión de CGS en Iris con $p > 1$. 65
Figura 5 Precisión de K-medias en Iris con $p > 1$. 65
Figura 6 Precisión de CGS en Wine con $p > 1$. 66
Figura 7 Precisión de K-medias en Wine con $p > 1$. 66
Figura 8 Precisión de CGS en Zoo con $p > 1$. 67
Figura 9 Precisión de K-medias en Zoo con $p > 1$. 67
Figura 10 Precisión de CGS en Stone Flakes con $p > 1$. 68
Figura 11 Precisión de K-medias en Stone Flakes con $p > 1$. 68
Figura 12 Precisión de CGS en Titanic con $p > 1$. 69
Figura 13 Precisión de K-medias en Titanic con $p > 1$. 69
Figura 14 Precisión de CGS en Bach con $p > 1$. 70
Figura 15 Precisión de K-medias en Bach con $p > 1$. 70
Figura 16 Precisión de CGS en Celeb con $p > 1$. 71
Figura 17 Precisión de K-medias en Celeb con $p > 1$. 71
Figura 18 Precisión de CGS en Iris con $0 < p < 1$. 72
Figura 19 Precisión de K-medias en Iris con $0 < p < 1$. 72
Figura 20 Precisión de CGS en Wine con $0 < p < 1$. 73
Figura 21 Precisión de K-medias en Wine con $0 < p < 1$. 73
Figura 22 Precisión de CGS en Zoo con $0 < p < 1$. 74
Figura 23 Precisión de K-medias en Zoo con $0 < p < 1$. 74
Figura 24 Precisión de CGS en Stone Flakes con $0 < p < 1$. 75
Figura 25 Precisión de K-medias en Stone Flakes con $0 < p < 1$. 75
Figura 26 Precisión de CGS en Titanic con $0 < p < 1$. 76
Figura 27 Precisión de K-medias en Titanic con $0 < p < 1$. 76
Figura 28 Precisión de CGS en Bach con $0 < p < 1$. 77
Figura 29 Precisión de K-medias en Bach con $0 < p < 1$. 77
Figura 30 Precisión de CGS en Celeb con $0 < p < 1$. 78
Figura 31 Precisión de K-medias en Celeb con $0 < p < 1$. 78
Figura 32 Precisión de CGS en Iris con $0.88 \leq p \leq 1.02$. 79
Figura 33 Precisión de K-medias en Iris con $0.88 \leq p \leq 1.02$. 79
Figura 34 Precisión de CGS en Wine con $0.88 \leq p \leq 1.02$. 80
Figura 35 Precisión de K-medias en Wine con $0.88 \leq p \leq 1.02$. 80
Figura 36 Precisión de CGS en Zoo con $0.88 \leq p \leq 1.02$. 81
Figura 37 Precisión de K-medias en Zoo con $0.88 \leq p \leq 1.02$. 81
Figura 38 Precisión de CGS en Stone Flakes con $0.88 \leq p \leq 1.02$. 82
Figura 39 Precisión de K-medias en Stone Flakes con $0.88 \leq p \leq 1.02$. 82
Figura 40 Precisión de CGS en Titanic con $0.88 \leq p \leq 1.02$. 83
Figura 41 Precisión de K-medias en Titanic con $0.88 \leq p \leq 1.02$. 83
Figura 42 Precisión de CGS en Bach con $0.88 \leq p \leq 1.02$. 84
Figura 43 Precisión de K-medias en Bach con $0.88 \leq p \leq 1.02$. 84

Figura 44 Precisión de CGS en Celeb con $0.88 \leq p \leq 1.02$. 85
Figura 45 Precisión de K-medias en Celeb con $0.88 \leq p \leq 1.02$. 85
Figura 46 Precisión de CGS en Iris con distancia híbrida. 86
Figura 47 Precisión de K-medias en Iris con distancia híbrida. 86
Figura 48 Precisión de CGS en Wine con distancia híbrida. 87
Figura 49 Precisión de K-medias en Wine con distancia híbrida. 87
Figura 50 Precisión de CGS en Zoo con distancia híbrida. 88
Figura 51 Precisión de K-medias en Zoo con distancia híbrida. 88
Figura 52 Precisión de CGS en Stone Flakes con distancia híbrida. 89
Figura 53 Precisión de K-medias en Stone Flakes con distancia híbrida. 89
Figura 54 Precisión de CGS en Titanic con distancia híbrida. 90
Figura 55 Precisión de K-medias en Titanic con distancia híbrida. 90
Figura 56 Precisión de CGS en Bach con distancia híbrida. 91
Figura 57 Precisión de K-medias en Bach con distancia híbrida. 91
Figura 58 Precisión de CGS en Celeb con distancia híbrida. 92
Figura 59 Precisión de K-medias en Celeb con distancia híbrida. 92

Introducción

La Inteligencia artificial abarca muchas áreas de aplicación, tan diversas como: comprensión del lenguaje, aprendizaje y sistemas adaptativos, resolución de problemas, percepción (visual), modelado, robótica y juegos [Pannu, 2015]; en esta última se prueban muchas de las aplicaciones nuevas en el mundo, tal como lo hizo Turing con su juego de ajedrez.

Dentro de las áreas de la inteligencia artificial se encuentra el reconocimiento de patrones, que tiene como objetivo catalogar diferentes objetos en un número predeterminado de clases. Los algoritmos diseñados para asignar cada objeto a una clase o categoría, son comúnmente llamados clasificadores. Cuando los clasificadores conocen la categoría real de cada objeto son denominados como algoritmos supervisados, en otro caso se les llama no supervisados.

El problema es, cómo realizar el análisis y gestión de la gran cantidad de información, con la que se está en contacto todos los días. Uno de los medios vitales para manejar estos datos es clasificarlos o agruparlos en un conjunto de categorías [Flores-Cruz et al., 2017]. Por lo anterior, resulta importante hablar de los procesos de agrupamiento o clustering, en los cuales se tiene un conjunto de elementos sin etiquetar, que deben ser agrupados en categorías dependiendo de su similitud. Así pues, las etiquetas se crean de forma posterior sin tener que alimentar el clasificador con información extra, y se establece que los elementos dentro de un grupo son parecidos entre ellos y diferentes a los de otros grupos.

Para comprender un nuevo objeto o fenómeno, las personas siempre intentan buscar las características o patrones que pueden describirlo, posteriormente lo comparan con otros objetos conocidos según la similitud o disimilitud generalizada como proximidad, de acuerdo a ciertas normas o reglas [Flores-Cruz et al., 2017]. Es decir, un grupo de objetos que están clasificados juntos deben tener características en común.

En este trabajo la categorización se realizará a través de dos estrategias. La primera es el problema de la coloración de gráficas suaves. Pero ¿qué es la coloración de gráficos?, se puede describir básicamente, como asignar un color a cada vértice, y al conjunto de vértices que reciben el mismo color se le llama clase de color, dicha asignación debe satisfacer algunas restricciones.

Por lo tanto, la coloración de gráficas suaves (CGS), se define como una generalización del problema de coloración, en donde dada una gráfica completa con ponderaciones en sus aristas, permite buscar una coloración que minimice la dureza, es decir, la suma de las penalizaciones entre los vértices con el mismo color [Lara-Velázquez et al., 2015].

La segunda estrategia con la que se clasificará en este trabajo, es el algoritmo K-medias propuesto por MacQueen en 1967 [MacQueen, 1967]; a pesar de su longevidad sigue siendo bastante útil. El algoritmo consiste en crear subconjuntos o grupos, que se determinan según la distancia más corta de cada elemento con alguno de los centroides que fueron iniciados de forma aleatoria. Una vez realizado lo anterior el centroide se actualiza de forma iterativa, con el cálculo del nuevo centroide como la media del grupo.

Para llevar a cabo la clasificación con ambos algoritmos, es necesario definir una distancia. En CGS, la distancia entre dos objetos, que se encuentran en la misma clase, está dada por la ponderación de la arista que los une. Por otro lado, en K-medias se calcula la distancia de cada objeto al centroide de su clase. De esta manera, si dos objetos están muy cerca respecto a una métrica dada, es muy probable que estén en la misma clase; si los objetos están lejos uno del otro, estarán en clases diferentes. Entre las métricas más utilizadas para K-medias se encuentran las siguientes: Jacquard, logarítmica, cuadrados independientes, coeficientes de correlación de Pearson y las métricas Lp.

En la literatura las métricas más comunes son las derivadas de la distancia Minkowski y su ecuación, que al variar el valor del exponente p y hacerlo igual a 1, se produce la distancia Manhattan, y si el valor es igual a 2 se obtiene la distancia euclidiana, pero con un cambio a la ecuación, eliminando la raíz cuadrada, da como resultado la distancia euclidiana cuadrática; para este trabajo se eliminó la raíz cuadrada para todos los valores de p que serán usados a lo largo del estudio.

La mayoría de las investigaciones con clasificadores se centran normalmente en la mejora del algoritmo, ya sea en eficiencia o rapidez y en cuanto a la métrica a utilizar regularmente se selecciona una, comúnmente la euclidiana, pero el interés de este estudio es más amplio, se enfoca en probar varias métricas que están bien definidas y deben cumplir con una serie de axiomas, así como semimétricas que satisfacen todo en la definición de una métrica, con excepción de la desigualdad del triángulo y así conocer el desempeño que obtienen.

El objetivo de este trabajo es identificar qué métrica obtiene el mejor rendimiento con los clasificadores utilizados, por medio de un estudio comparativo de las soluciones alcanzadas, generadas por métricas tradicionales, contra las métricas obtenidas a partir de combinaciones lineales positivas de métricas Lp $(0 < p \leq \infty)$. En general, se trata de probar con qué valor de p y con qué combinación de métricas Lp se obtienen mejores resultados.

En el Capítulo 1, se abordan todas las definiciones y contenidos necesarios para lograr comprender el tema de investigación, en el que se centra este trabajo; inicialmente se identifican las métricas y su conceptualización formal, así como ejemplos de las más comunes y las ecuaciones que las describen. Enseguida se especifican los clasificadores y los procesos de agrupamiento, para después hablar acerca de la coloración de gráficas y más específicamente de la coloración de gráficas suaves. Posteriormente se explica el funcionamiento del segundo clasificador utilizado en este trabajo, K-medias. Finalmente, el apartado de diseño de experimentos, se enfoca en el análisis de varianza (ANOVA), que es la prueba estadística que permitirá validar los resultados.

En el Capítulo 2, se hace una descripción a detalle de las bases de datos, tales como el número total de instancias en cada base, los atributos y lo que significan; además de los rangos de valores en los que oscilan cada uno de los atributos, también si existen valores faltantes y finalmente las clases que la constituyen y el número de elementos que las componen.

En el Capítulo 3, se detallan las métricas utilizadas en ambos clasificadores, comenzando con métricas enteras, por ejemplo, la altamente conocida distancia Manhattan, la distancia euclidiana y euclidiana cuadrática que, son las más comúnmente utilizadas en clasificadores. Después se men-

cionan valores decimales menores a 1, que se utilizan en la ecuación de la distancia Minkowski, lo que lleva a explorar de una manera aún más detallada introduciendo cambios en centésimas y finalmente probar un híbrido de la combinación de las distancias. En seguida, se integró la forma en que se trataron las bases de datos. Al final de este capítulo se muestra el funcionamiento de los dos clasificadores utilizados en este trabajo.

En el Capítulo 4, se hace un análisis de resultados obtenidos en cada métrica propuesta, a través del porcentaje de precisión alcanzado en la base de datos probada y para cada clasificador, lo anterior respaldado por pruebas ANOVA sobre los resultados generados de ambos clasificadores utilizados con cada tipo de métrica, para poder concluir indudablemente con qué métrica se obtienen los mejores resultados.

Finalmente, se presentan las conclusiones de este tema de investigación.

Objetivos

Objetivo general

- Implementar combinaciones de métricas Lp en la coloración de gráficas suaves y el algoritmo K-medias, para mejorar su desempeño en instancias de prueba.

Objetivos particulares

- Estudiar las métricas utilizadas en clasificadores.
- Desarrollar los algoritmos clasificadores.
- Aplicar diversas métricas en la coloración de gráficas suaves.
- Aplicar diversas métricas en el algoritmo K-medias.
- Emplear una combinación lineal de métricas en ambos clasificadores.

Capítulo 1. Marco Teórico

Un supuesto básico del análisis de conglomerados, es que inicialmente se sabe poco sobre los datos. Las reglas o características que definirán cómo se constituirá un grupo no se conocen a priori, y saber el grado de similitud entre los objetos, no revela qué causó que se formaran los grupos en primer lugar [Hausner, 2010].

1.1. Métricas

Si es posible calcular grados de similitud entre objetos, se puede afirmar que los objetos pertenecen a un espacio métrico, pero ¿qué es un espacio métrico?, consiste en un conjunto de objetos y una función llamada métrica, que toma dos objetos en el conjunto y produce un número real no negativo. Este número es la distancia entre los dos objetos, si dos objetos x y y son similares, por lo general la distancia entre ellos será pequeña, y si los objetos son muy diferentes, comúnmente será grande [Zarinbal, 2009], [Hausner, 2010].

La siguiente definición, es mucho más formal y clara, en cuanto a lo que se quiere establecer en relación a un espacio métrico.

Definición: Un espacio métrico es un par (X, d) donde X es un conjunto no vacío y d es una función real definida en $X \text{x} X$, llamada función distancia o métrica, y que satisface los siguientes tres axiomas [Guccione and Guccione, 2017]:

No negatividad, simetría y desigualdad triangular, respectivamente.

$i)$ $d(x, y) \geq 0$ y $d(x, y) = 0 \Leftrightarrow x = y$
$ii)$ $d(x, y) = d(y, x)$
$iii)$ $d(x, y) \leq d(x, z) + d(z, y)$

Como se mencionó, lo que produce la función que toma estos dos elementos es un número real no negativo, es decir la distancia entre ellos, lo que trae consigo la pregunta ¿qué es la distancia?, ésta es una descripción numérica de qué tan lejos están los objetos en un momento dado en el tiempo. En la física o en la discusión cotidiana, la distancia puede referirse a una longitud física, un período de tiempo o una estimación basada en otros criterios [Zarinbal, 2009].

Han habido muchos intentos de generalizar el ajuste de la métrica, modificando algunos de los axiomas de los espacios métricos. Por lo tanto, se han introducido otros tipos de espacios y resultados métricos, se han extendido a nuevas configuraciones tal como menciona Kiris [Kiris, 2017].

En el año 2000 Branciari, [Branciari, 2000] definió lo que se denomina espacio métrico generalizado. En éste, la desigualdad del triángulo de un espacio métrico ordinario se ha reemplazado por una nueva desigualdad, que involucra tres términos en lugar de dos. Ésta es conocida como la desigualdad cuadrilátera, tiene una suposición más débil que la triangular [Kiris, 2017].

Hubo una gran cantidad de literaturas que trabajaron no solo en un espacio métrico, sino también en otros espacios topológicos. Hace unos años, [Jleli and Samet, 2015] definieron una métrica generalizada, conocida como métrica JS. La ventaja de su idea es que muchos espacios topológicos están cubiertos por el métrico JS. Por esta razón, los resultados de los teoremas de punto fijo en espacios JS-métricos han sido estudiados recientemente [Thangthong and Khemphet, 2018].

Como se mencionó anteriormente, si se toman dos objetos x y y y resulta que son similares, su distancia será pequeña, y si los objetos son muy diferentes, por lo general será lo opuesto, lo anterior puede ser observado al utilizar medidas de similaridad o disimilaridad como las mostradas en la Tabla 1 [Xu and Wunsch, 2005]. En dos dimensiones, la distancia entre dos puntos se puede calcular utilizando el teorema de Pitágoras, el cual es ampliamente conocido y comprobable; en este teorema se generaliza a la distancia euclidiana utilizada en las dimensiones más altas [Zarinbal, 2009], [Hausner, 2010].

Tabla 1: Medidas de similaridad y disimilaridad para características cuantitativas.

Métrica	Ecuación
Distancia Minkowski	$D_{ij} = \left(\sum_{l=1}^{k} \lvert x_{il} - x_{jl}\rvert^{p}\right)^{\frac{1}{p}}$
Distancia Manhattan	$D_{ij} = \sum_{l=1}^{k} \lvert x_{il} - x_{jl}\rvert$
Distancia Euclidiana	$D_{ij} = \left(\sum_{l=1}^{k} (x_{il} - x_{jl})^{2}\right)^{\frac{1}{2}}$
Distancia Sup	$D_{ij} = \max_{1\leq l\leq d} \lvert x_{il} - x_{jl}\rvert$
Distancia de Mahalanobis	$D_{ij} = (x_i - x_j)^T * A^{-1}(x_i - x_j)$
Correlación de Pearson	$D_{ij} = \frac{1-r_{ij}}{2}$, donde $r_{ij} = \frac{\sum_{l=1}^{k}(x_{il}-\overline{x}_l)(x_{il}-\overline{x}_j)}{\sqrt{\sum_{l=1}^{k}(x_{il}-\overline{x}_l)(x_{il}-\overline{x}_j)}}$
Distancia de punto de simetría	$D_{ir} = \min_{j=1,\ldots,n \ \& \ j\neq i} \frac{\lVert(x_i-x_r)+(x_j-x_r)\rVert}{\lVert(x_i-x_r)\rVert+\lVert(x_j-x_r)\rVert}$
Similitud por coseno	$S_{nm} = \cos\alpha \frac{X_n^T X_m}{\lVert X_n\rVert\lVert X_m\rVert}$

1.2. Clasificadores

El reconocimiento de patrones, es la clasificación de grandes cantidades de objetos físicos o abstractos con el propósito de extraer información útil, que permita establecer agrupaciones de estos objetos. El flujo de clasificación consiste en la obtención de patrones, mediante un proceso de segmentación, extracción de características y reseña de cada objeto como una colección de descriptores [Flores-Cruz et al., 2017], [Pérez-Ortega et al., 2018].

Los sistemas clasificadores son un tipo especial de reconocimiento de patrones. Éstos colocan una etiqueta a un objeto de acuerdo a sus características, es decir, un sistema clasificador en si es una abstracción de la habilidad humana para reconocer clases, grupos o categorías de objetos observados, lo cual es una acción que se lleva a cabo diariamente, incluso a veces de forma inconsciente, por ejemplo, decidir los platillos que más nos gustan con base en los sabores que sabemos que nos agradan.

En este sistema, las etiquetas pueden ser determinadas previamente, y luego dado un objeto se debe decidir en qué clase particular tendría que ser clasificado, según las opciones existentes. Por otro lado cuando las etiquetas no están definidas con anterioridad, es necesario encontrar una forma de agrupar los elementos que, sean más parecidos entre ellos y a su vez diferentes a los otros grupos. Una vez hecho lo anterior, los clasificadores colocan una etiqueta a un objeto de acuerdo con sus características.

Para la tarea que implica el reconocimiento de patrones, se busca ordenar ciertos objetos dentro de una de las k categorías [Flores-Cruz et al., 2017].

Básicamente los sistemas clasificadores son de dos tipos:

- Supervisado: dado un conjunto de etiquetas previamente establecidas, la meta es encontrar una regla que, prediga correctamente la etiqueta que se asociará, con las entradas no vistas anteriormente.

- No supervisado: dado un conjunto de elementos, con los atributos que lo caracterizan y sin ninguna etiqueta, se agrupan dichos elementos en cúmulos naturales, creando así etiquetas para los grupos.

Para llevar a cabo la tarea señalada, se debe diseñar o utilizar un algoritmo que, a partir de un conjunto de objetos, sea capaz de asignar una etiqueta de un conjunto de clases ya preestablecidas; o bien que al tener un conjunto de objetos y sus atributos, se puedan crear las clases que los agrupen, a partir de características comunes entre ellos, empleando una métrica de similitud para los objetos dependiendo del tipo de algoritmo (supervisado o no supervisado).

Entre los clasificadores más conocidos se pueden mencionar, el clasificador bayesiano simple, clasificador parzen, discriminadores lineales, redes bayesianas, árboles de decisión, clasificador con PCA (Principal component analysis) y basados en redes neuronales.

1.3. Procesos de agrupamiento

El agrupamiento de datos es un área única y dinámica, que se ha convertido recientemente en un nicho de investigación muy activo para la minería de datos. Sirve también como el punto focal para estadísticas, aprendizaje automático, tecnología de bases de datos espaciales, recuperación de información, entre otros campos de aplicación [Flores-Cruz et al., 2017], [Dalatu et al., 2017].

La agrupación es un método de clasificación no supervisada debido a la ausencia de información etiquetada. Por lo tanto, es una forma de aprender por observación, en lugar de aprender con ejemplos. La agrupación se da en particiones de un conjunto de objetos, que están relacionados entre sí formando un grupo y que a su vez están escasamente relacionados con los objetos en otros cúmulos. Un grupo apropiado se logra cuando los elementos de éste están altamente interrelacionados y son diferentes de los elementos en otros grupos. [Dalatu et al., 2017].

El número de k grupos puede estar predeterminado o puede ser decidido por el algoritmo. Formalmente un algoritmo de agrupamiento produce un mapeo de la forma $C : \{1, ..., n\} \rightarrow \{1, ..., k\}$, donde n es el número de objetos analizados, es decir, a cada objeto se le asocia un color, etiqueta o grupo; además, C debe ser una función sobreyectiva, para garantizar que en cada conjunto k, hay al menos un objeto.

Algunos de los agrupadores de mayor importancia son: por medidas de similitud y distancia, jerárquicos, basados en error cuadrático y en teoría de grafos, así como en técnicas de búsqueda combinatoria, de lógica borrosa y agrupamiento de datos secuenciales. [Flores-Cruz et al., 2017].

1.4. Coloración de gráficas

La teoría de grafos surge por primera vez de la resolución de un conocido problema, llamando el problema de los siete puentes de Königsberg [Euler, 1741], este problema se plantea de manera informal de la siguiente manera:

- Dado el mapa de Königsberg, con el río Pregel dividiendo el plano en cuatro regiones distintas, que están unidas por medio de los siete puentes, ¿es posible encontrar un camino comenzando desde cualquiera de estas regiones, que pase por todos los puentes, recorriendo sólo una vez cada uno, y regresando al mismo punto de partida?

En la demostración, representó cada puente como una recta que une dos vértices, cada uno de los cuales corresponde con una región. Esta abstracción del problema dio pie a la noción de un grafo, es decir, líneas como aristas y puntos como vértices.

Posteriormente surge el problema de la coloración de grafos, que consiste en dividir el conjunto de vértices de una gráfica para satisfacer ciertas restricciones. Los problemas de coloración han sido ampliamente estudiados en matemáticas discretas y ciencias de la computación teórica [Kolay et al., 2019].

Dado un color X de una gráfica G, se dice que el conjunto de vértices que reciben el mismo color es una clase de color. En primer lugar, se debe decidir el número máximo de colores a utilizar, suponiendo que el grafo tiene n vértices, se necesitarán proporcionalmente n colores. En ese caso, se tienen un total de n^n soluciones posibles, número que crece de forma notable a medida que el valor de n aumenta [Seijas, 2017].

Lo anterior trae consigo uno de los problemas de coloración más conocidos, el problema del número cromático que busca la menor cantidad de colores necesarios para colorear sus vértices, sin que dos vértices adyacentes tengan el mismo color.

La coloración es un problema clásico de optimización con muchas aplicaciones prácticas, existe una amplia gama de métodos heurísticos para abordar el problema de la coloración de gráficos: desde algoritmos glotones rápidos hasta metaheurísticas que requieren un poco más de tiempo. Aunque el último produce mejores resultados en términos de minimizar el número de colores, el primero se utiliza ampliamente debido a su simplicidad y velocidad para producir buenos resultados [Galán, 2017].

1.5. Coloración de gráficas suaves

La coloración de gráficas suaves es un caso especial del problema de coloración de gráficas, por ser una generalización del problema de coloración robusta, se sabe que este problema es del tipo NP-duro y se necesita utilizar metaheurísticas en instancias mayores a 20 vértices [Lara-Velázquez et al., 2015].

En una gráfica completa ponderada, se busca una coloración que minimice la dureza; es decir, encontrar una coloración que minimice la suma de las penalizaciones entre los vértices que tienen el mismo color. Formalmente, la coloración de gráficas suaves es definida por distintos autores [Lara-Velázquez et al., 2015, Aparicio Reyes, 2017, Vásquez-Calderón et al., 2018, Urueta-Hinojosa et al., 2019], de la siguiente manera:

Sea $G = (V, E)$ una gráfica completa no dirigida, es decir, el conjunto de los vértices $V = \{1, 2, ..., n\}$ en G y todas sus aristas posibles (i, j), donde:

$$G = (V, E);\ |V| = n;\ |E| = n(n-1)/2. \tag{1}$$

Se define una penalización en cada arista (i, j), denotada por $p_{i,j}$ tal que:

$$p_{ij} \geq 0,\ \forall\ (i, j) \in E. \tag{2}$$

Una función de coloración en los vértices de $G = (V, E)$ se define como:

$$C^k : \{1, 2, ..., n\}\ \rightarrow\ \{1, 2, ..., k\}, \tag{3}$$

Donde k es el número total de colores $1 < k < n$ que identifica $C(i)$ como el color en el vértice i. Para una coloración C^k en una gráfica la función de dureza de la coloración C^k está dada por:

$$H(C^k) = \sum_{(i,j)\in E,\ C^k(i)=C^k(j)} p_{ij} \tag{4}$$

El objetivo del problema de CGS es encontrar la coloración C^k_{op} que minimiza la dureza $H(C^k_{op})$, para lograr dicho objetivo a continuación se describen dos propiedades de cada coloración.

1.5.1. Solidez de una coloración

Una vez realizada una coloración con k colores en una gráfica completa con n vértices, el número promedio de vértices pintados con el mismo color es $m = n/k$ y el número de aristas compartidas por m vértices son las combinaciones $C(m, 2) = m(m-1)/2$, por lo que el número promedio de penalizaciones incluidas en la función de dureza, es proporcional al número promedio de vértices pintados con el mismo color multiplicado por el número de colores.

La primera propiedad es la función de solidez de una coloración, que se define como la dureza de un gráfico dividido por el número medio de aristas que contribuyen a la dureza [Urueta-Hinojosa et al., 2019]:

$$S(C^k_{op}) = \frac{H(C^k_{op})}{km(m-1)/2} = \frac{2H(C^k_{op})}{k\frac{n}{k}(\frac{n}{k}-1)} = \frac{2H(C^k_{op})}{n(n-k)}. \tag{5}$$

1.5.2. Resiliencia de una coloración

La segunda propiedad es la resiliencia de una coloración C^k, que se define como el porcentaje que disminuye la solidez de una coloración con $k-1$ colores, con respecto a la realizada con k de ellos y se puede expresar como sigue [Urueta-Hinojosa et al., 2019]:

$$R(C^k_{op}) = \frac{S(C^{k-1}_{op}) - S(C^k_{op})}{S(C^k_{op})}. \tag{6}$$

Si resulta que al añadir un color adicional a la coloración de la gráfica, hay un aumento significativo en su resiliencia, eso significa que se ha encontrado un número adecuado de clases para clasificar.

1.6. K-medias

El término K-medias fue utilizado por primera vez por MacQueen [MacQueen, 1967], aunque la idea original se le atribuye a Steinhaus [Steinhaus, 1956], un poco más de 10 años antes. Mientras que el algoritmo estándar fue propuesto por Stuart Loyd en 1957, aunque no se publicó fuera

de los laboratorios Bell hasta 1982 [Lloyd, 1982].

El algoritmo de agrupación K-medias, consiste en un número de subconjuntos que se determinan según la iniciación de los centros en estos subconjuntos; en relación con la distancia más corta entre cada punto y su centro, normalmente llamados k centroides, cada dato es asignado a un subconjunto y una vez realizado esto, los centros de estos subconjuntos se van actualizando [Pham et al., 2019], es decir, el algoritmo consta de tres pasos, que se repiten hasta la convergencia, lo que dependiendo del autor se puede tomar como un paso 4.

- Paso 1: Se definen los puntos centrales (centroides).
- Paso 2: Los k grupos son generados al asociar cada punto a su centroide basado en la distancia mas corta.
- Paso 3: El centroide de cada k grupo se recalcula.
- Paso 4: Se repiten los pasos 2 y 3 hasta la convergencia.

Para lograr lo anterior, el algoritmo selecciona k centros al azar basándose en los objetos, por lo que cada uno representa una media o centroide inicial del grupo, luego, para los objetos restantes, asigna cada objeto al grupo con el que se relaciona mejor, establecido por la distancia entre el objeto y la media del grupo, después calcula una nueva media para cada grupo utilizando los objetos asignados al grupo en la iteración anterior. [Dalatu et al., 2017].

Como en este caso se trata de un algoritmo heurístico, no hay garantía de que converja al óptimo global, por lo que el resultado puede depender de los grupos iniciales; la elección de un valor k apropiado depende de las instancias consideradas, es una tarea bastante difícil, esta situación suele ser abordada por ensayo y error. Sin embargo, se han llevado a cabo varias investigaciones para determinar automáticamente el número k de grupos [Pérez-Ortega et al., 2018].

Se puede señalar que el algoritmo K-medias ha sido utilizado como un método de agrupamiento eficiente, sin embargo, como se ha mencionado el rendimiento de este algoritmo, es altamente dependiente de la selección de los centroides de agrupamiento iniciales o punto central, por lo tanto, el método para elegir los centros de agrupamiento iniciales es muy importante, tal como comentan Yang y Mexicano en sus respectivos trabajos [Yang et al., 2017], [Mexicano et al., 2016].

Las mejoras para este algoritmo se han aplicado en su fase de inicialización para mejorar la calidad del clúster, en la fase de clasificación, y finalmente en la fase de convergencia, cuya idea principal consiste en reducir el tiempo de ejecución sin una pérdida significativa de calidad, es decir, reducir el número de ciclos antes de alcanzar la convergencia [Mexicano et al., 2016].

1.7. Diseño de experimentos

En la publicación de Fisher, [Fisher, 1918] se utiliza por primera vez el término varianza, que se define como el cuadrado de la desviación estándar de una variable con respecto a su media, la cual tiene como valor mínimo el 0.

El análisis de varianza unidireccional (o de una vía) ANOVA, (del inglés Analysis of Variance), o prueba F nombrada así por Ronald Fisher, es una de las técnicas estadísticas más comunes en investigación, utilizada para determinar si existen diferencias estadísticamente significativas entre las medias de tres o más grupos [Blanca et al., 2017].

La prueba F supone que la variable se distribuye de manera normal e independiente con las mismas varianzas entre los grupos. Sin embargo, los datos en el mundo real a menudo no se distribuyen normalmente y las variaciones no siempre son iguales.

Uno de los primeros estudios sobre esta prueba fue llevado a cabo por Pearson, [Pearson, 1931] quien descubrió que la prueba F, era válida siempre que la desviación de la normalidad no fuera extrema y que el número de grados de libertad asignados a la variación residual, no fuera demasiado pequeño.

Desde 1931 hasta 1978 existieron algunos resultados contradictorios, los cuales pueden atribuirse al hecho de no haber utilizado un criterio estándar para evaluar la robustez. Según el criterio enunciado por Bradley, [Bradley, 1978], una prueba estadística se considera sólida si la tasa de error empírica Tipo I se encuentra entre .025 y .075 para un nivel alfa de .05, un error Tipo I o falso positivo, es el error que se comete cuando no se acepta la hipótesis nula siendo esta realmente verdadera [Blanca et al., 2017].

En resumen, la prueba F es robusta para desviaciones moderadas de la normalidad, cuando la desviación es de esta manera, las poblaciones tienen la misma forma de distribución y los tamaños de muestra son grandes e iguales.

Una vez dicho lo anterior, ANOVA permite rechazar la hipótesis nula H_0 [Peña et al., 2018]:

- Hipótesis nula, H_0: $\mu 1 = \mu 2 = \ldots = \mu k$.
- Hipótesis alternativa, H_1: $\mu i \neq \mu j$ para algún i, j donde $i \neq j$.

Donde $\mu i, i = 1, ..., k$ son las medias de las k poblaciones, asumiendo independencia, normalidad y la homocedasticidad.

Cada población se caracteriza por uno o más factores categóricos, luego ANOVA compara las medias de la variable en los diferentes niveles de factores. Mientras que ANOVA unidireccional solo considera un factor con dos o más o grupos [Peña et al., 2018]. Para aceptar la H_0, la prueba F se usa calculando las estadísticas F dadas por la ecuación (7):

$$F = \frac{\frac{\sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2}{K-1}}{\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\overline{Y}_{ij} - \overline{Y}_i)^2}{N-K}} \tag{7}$$

Donde Y_i indica la media de la muestra en el grupo i-ésimo, n_i es el número de muestras en el grupo i-ésimo, Y indica la media general de las muestras, Y_{ij} es la muestra j-ésima en el grupo i-ésimo y K denota el número de grupos. Se espera que el valor F sea aproximadamente 1, sin embargo, para rechazar H_0, se necesita un valor F alto.

Un valor F alto es difícil de interpretar por sí solo. Por lo que, el valor de referencia para rechazar H_0 se basa en el valor p, que es la probabilidad de observar un valor F que es al menos tan alto como el valor obtenido en el estudio, bajo el supuesto de que H_0 es verdadero, comúnmente, H_0 se rechaza si $p \leq 0.05$.

Capítulo 2. Bases de datos de prueba

Las bases de datos utilizadas en este trabajo se extrajeron del repositorio de aprendizaje automático de la Universidad de California en Irvine (UCI por sus siglas en inglés) [Aha et al., 1987] con excepción de Titanic: Machine Learning from Disaster que se extrajo del repositorio de Kaggle [Goldbloom and Hamner, 2010] y Large-scale CelebFaces Attributes (CelebA) [Ziwei Liu and Tang, 2015].

2.1. Iris

La base de datos Iris consta de 150 observaciones, 4 atributos sin valores faltantes y 3 clases que corresponden al tipo de planta, cada una constituida por 50 observaciones.

Los atributos de la base de datos son:

- Longitud del sépalo en cm.
- Ancho del sépalo en cm.
- Longitud del pétalo en cm.
- Ancho de pétalo en cm.

Los tipos de planta o su clase son:

- Iris setosa: 50 observaciones.
- Iris versicolor: 50 observaciones.
- Iris virginica: 50 observaciones.

2.2. Wine

Se utilizó la base de datos Wine que consta de 178 observaciones, 13 atributos sin valores faltantes, todos los datos son numéricos y 3 clases que corresponden a 3 tipos de vino.

Los atributos de la base de datos son:

- Grado de alcohol: con valores entre 11.03 y 14.83.
- Cantidad de ácido málico: presente en las frutas de sabor ácido como las uvas, con valores entre 0.74 y 5.8.
- Cenizas: producto de la incineración del extracto seco del vino, con valores que se encuentran entre 1.36 y 3.23.

- Alcalinidad de las cenizas: se trata de la suma de los cationes de amonio que se encuentran mezclados en los ácidos orgánicos del vino, con valores entre 10.6 y 30.
- Magnesio: se determina directamente en el vino por espectrofotometría de absorción atómica, teniendo valores entre 70 y 162.
- Fenoles totales: actúan como antioxidantes al combinarse con el oxígeno, con valores entre 0.98 y 3.88.
- Flavonoides: son compuestos polifenólicos naturales que se encuentran en las vides, sus valores se colocan entre 0.34 y 5.08.
- Fenoles no flavanoides: al igual que los anteriores son compuestos polifenólicos pero que se encuentran en una clasificación diferente al ser no flavonoides, con valores entre 0.13 y 0.66.
- Proantocianidinas: son los flavonoides cuantitativamente más importantes en la uva y el vino, donde inciden de manera importante, otorgando propiedades beneficiosas para la salud humana, sus valores oscilan entre 0.41 y 3.58.
- Intensidad del color: hace referencia al grado en que la luz lo puede atravesar o al grado de opacidad del vino, sus valores se colocan entre 1.28 y 13.
- Matiz: dentro del matiz se puede hablar por ejemplo de tonalidades, ribetes o reflejos, sus valores se encuentran entre 0.48 y 1.71.
- OD280/OD315: concentración de esas proteínas, con valores entre 1.27 y 4.
- Prolina: es un aminoácido proteinogénico, con valores distribuidos entre 278 y 1680.

Los tipos de vino son:

- Vino de mesa: 59 observaciones.
- Vino de crianza: 71 observaciones.
- Vino de reserva: 48 observaciones.

2.3. Zoo

Se utilizó la base de datos Zoo, que consta de 101 observaciones, de las cuales se utilizaron 84, 17 atributos sin valores faltantes y 4 clases que corresponden al tipo de animal.

Los atributos de la base de datos son binarios, excepto en el atributo de número de patas:

- Nombre del animal (este atributo fue suprimido).
- Cabello.

- Plumas.
- Huevos.
- Leche.
- Volador.
- Acuático.
- Depredador.
- Dentado.
- Vertebrado.
- Respira.
- Venenoso.
- Aletas.
- Número de patas.
- Cola.
- Doméstico
- Tamaño parecido a un gato.

Las clases son de tipo numérico (valores enteros en el rango [1,4]:

- Tipo 1: 41 observaciones con animales como son oso, elefante, y antílope.
- Tipo 2: 20 observaciones con aves como pato, cisne y cuervo.
- Tipo 3: 13 observaciones con animales marinos como el caballito de mar y las mantarrayas.
- Tipo 4: 10 observaciones que incluyen animales como cangrejos y langostas.

2.4. Stone Flakes

Se utilizó la base de datos Stone Flakes la cual se encuentra constituida por 79 observaciones, de las cuales se utilizaron 70 que cumplían con los criterios de limpieza de los datos, 8 atributos y 4 clases propuestas por arqueólogos.

Los atributos de la base de datos son:

- Índice de longitud y anchura: ancho del instrumento con valores de 1.02 hasta 1.69 cm.

- Índice de espesor: grosor del instrumento entre 16.5 y 43.7 cm.
- Índice de ancho-profundidad: sus valores oscilan entre 1.66 y 4.9 cm.
- Ángulo de descamación: ángulo de golpeo entre el objeto y la superficie en la que se talló, desde los 105 hasta los 131 grados.
- Plataforma primaria: frecuencias relativas que se encuentran entre 0 y 67.2.
- Plataforma facetada: frecuencias relativas cuyos valores se encuentran entre 0 y 67.2.
- Superficie dorsal totalmente trabajada: área con mayor desgaste con valores entre 5 y 94.1 cm^2.
- Proporción de superficie dorsal trabajada: proporción de desgaste respecto a la superficie total con valores que varían entre 30 y 98 %.

Las 4 clases propuestas por arqueólogos son:

- Paleolítico inferior: se considera que comenzó hace unos 2,5 millones de años y duró hasta hace unos 125-127,000 años.
- Técnica de Levallois: es una técnica de fabricación de instrumentos que data del comienzo del paleolítico medio, es decir de hace 120,000 años.
- Paleolítico medio: se considera que abarca desde hace 120,000 años y duró hasta hace 40,000 años.
- Homo sapiens: los restos más antiguos atribuidos a Homo sapiens se encuentran en Marruecos, con 315 000 años de antigüedad y hasta la actualidad (esta clase fue suprimida ya que se sobrepone a las 3 anteriores).

2.5. Titanic

La base de datos Titanic consta de 887 observaciones, con 6 atributos y 2 clases que corresponden a la supervivencia del pasajero.

Los atributos de la base de datos son:

- Clase del boleto.
- Sexo.
- Edad.
- Número de hermanos / cónyuges a bordo.
- Número de padres / hijos a bordo.

- Tarifa que pagó el pasajero.

Las 2 clases son:

- Superviviente: si.
- Superviviente: no.

2.6. Bach

Se utilizó la base de datos Bach Choral Harmony que se encuentra constituida por 5665 observaciones de las cuales se utilizaron 4554, que cumplían con los criterios de limpieza de los datos, 16 atributos y 17 clases que corresponden a la tonalidad del fragmento de la cantata.

Los atributos de la base de datos son:

- ID coral: correspondiente a los nombres de archivo de Bach Central (este atributo fue suprimido).
- Número de evento: índice del evento dentro del coral.
- Atributo 3-14, clases de tono: si o no (binario), dependiendo de si un tono dado está presente.
- Bajo: clase de tono de la nota de bajo.
- Medidor: enteros del 1 al 5, los números más bajos denotan eventos menos acentuados y los números más altos denotan eventos más acentuados.

Las clases son de tipo numérico (valores enteros en el rango [1,17]):

- Etiqueta del acorde: tonalidad del fragmento de la cantata.

2.7. CelebA

Se utilizó la base de datos CelebA, constituida por 202,599 imágenes de celebridades, localización de puntos destacados del rostro y 40 atributos binarios para cada imagen. Se extrajo una muestra de 4,000 observaciones de la base original, y se utilizó la localización de puntos destacados del rostro (ojo izquierdo, ojo derecho, boca y nariz) que, previamente fueron alineadas y recortadas de acuerdo con las dos ubicaciones de los ojos, con dos clases que corresponden con el sexo de la celebridad.

Los atributos de la base de datos son:

- Ojos: corresponde a la longitud del segmento entre los ojos, cuyos valores se encuentran entre 7 y 56.03.

- Ojo izquierdo - nariz: es la longitud del segmento entre el ojo izquierdo y el centro de la nariz, con valores distribuidos entre 6.08 y 52.47.
- Ojo derecho - nariz: es la longitud del segmento entre el ojo derecho y el centro de la nariz, sus valores se colocan entre 6 y 50.91.
- Nariz - boca izquierda: corresponde a la longitud del segmento entre la nariz y donde termina la boca en el lado izquierdo del rostro, sus valores se colocan entre 6 y 50.91.
- Nariz - boca derecha: corresponde a la longitud del segmento entre la nariz y donde termina la boca en el lado derecho del rostro, sus valores oscilan entre 7.07 y 61.58.
- Boca: corresponde a la longitud del segmento entre el lugar donde termina la boca en el lado izquierdo y el lado derecho del rostro, cuyos valores se encuentran entre 5.38 y 54.
- Ojo izquierdo - boca izquierda: es la longitud del segmento entre el ojo izquierdo y donde termina la boca en el lado izquierdo del rostro, con valores distribuidos entre 30.80 y 62.
- Ojo derecho - boca derecha: corresponde a la longitud del segmento entre el ojo derecho y donde termina la boca en el lado derecho del rostro, sus valores oscilan entre 30.67 y 65.19.
- Ojo izquierdo - boca derecha: es la longitud del segmento entre el ojo izquierdo y donde termina la boca en el lado derecho del rostro, sus valores se colocan entre 41.34 y 68.96.
- Ojo derecho - boca izquierda: corresponde a la longitud del segmento entre el ojo derecho y donde termina la boca en el lado izquierdo del rostro, cuyos valores se encuentran entre 40.60 y 68.87.

Las 2 clases son:

- Hombre: si.
- Hombre: no.

Capítulo 3. Métricas y algoritmos utilizados

Al utilizar coloración de gráficas suaves se requiere de una matriz de distancias, la cual representa la gráfica completa de los datos. El valor de cada elemento de la matriz es más pequeño cuanto más cercanos se encuentran dos elementos, e inversamente más grande cuanto más lejanos se encuentran dichos elementos.

Por otro lado, en K-medias al llevar a cabo el proceso de asignación, se requiere la distancia de cada elemento a todos los centroides, para realizar la asignación de dichos elementos al más cercano, así que tanto para coloración de gráficas suaves como para K-medias se requiere de calcular distancias, para este proceso se proponen primeramente dos tipos de métricas procedentes de la distancia de Minkowski, las cuales se mencionan a continuación:

- Métrica con valores enteros: al utilizar la distancia de Minkowski con valores enteros, se garantiza que es una métrica ya que cumple con 3 axiomas, de no negatividad, simetría y desigualdad del triángulo, para este tipo se utilizaron los valores de $p = 1$, 2, 3 y 4.

- Métrica con valores menores a 1: se refiere a una semimétrica, donde se garantizan todos los axiomas a excepción de la desigualdad del triángulo, en este caso se utilizaron los valores de $p = 0.1$, 0.3, 0.5, 0.7 y 0.9.

De lo observado en los resultados de los dos tipos de métricas anteriores, y apoyado en lo que menciona Love [Love and Dowling, 1985], que el valor de p = 0.9 es el indicado para realizar la clasificación, se decidió explorar más a detalle lo que sucede alrededor de los valores 0.9 y 1 (Manhattan).

- Valores en centésimas: para poder explorar alrededor de los valores mencionados, se tomaron algunos, un poco más pequeños que 0.9 y un poco más grandes a 1, para ser más preciso, se utilizaron los valores de $p = 0.88$, 0.89, 0.90, 0.91, ... , 1.0, 1.01 y 1.02.

- Distancia híbrida: para que la observación de estos valores sea aún más exhaustiva, se realizaron pruebas con una combinación de métricas, teniendo como base la distancia Manhattan y combinándola con el valor de p = 0.9, la forma en que se realizó esta exploración será abordada más adelante.

Todas las métricas comentadas aquí, serán explicadas más a detalle en las páginas siguientes para tener un panorama completo, para posteriormente realizar el análisis de los resultados de todas las pruebas con las diferentes métricas y en ambos clasificadores.

3.1. Distancia Minkowski

La distancia de Minkowski, es de gran importancia para este estudio, pero antes de especificar cómo se define, debemos asumir que si tenemos dos puntos $X = (x_1, x_2)$ y $Y = (y_1, y_2)$, entonces $d(X, Y)$ es la función distancia entre los puntos X y Y, y tienen algunas características que se satisfacen:

- $d(X, Y) \geq 0 \quad \forall\, X, Y$ No negatividad,
- $d(X, Y) = 0 \Leftrightarrow X = Y \quad \forall\, X, Y$ Propiedad idéntica,
- $d(X, Y) = d(X, Y) \quad \forall\, X, Y$ Simetría,
- $d(X, Y) \leq d(X, Z) + d(Z, Y) \quad \forall\, X, Y$ Desigualdad triangular.

Lo anterior se puede describir de la siguiente forma: la no negatividad establece que la distancia entre X y Y es mayor o igual a cero para cualquier X y Y que se tomen; la siguiente expresa que la distancia entre los dos puntos es igual a 0 solamente en caso de que esos dos puntos sean iguales; la de simetría indica que la distancia entre los puntos X y Y es igual a la distancia entre Y y X; finalmente la desigualdad triangular denota que la distancia entre X y Y es menor o igual a la distancia al incluir un tercer punto Z, y que se considere la distancia de X a Z más la distancia de Z a Y.

Dicho lo anterior, la distancia Minkowski de orden p entre dos puntos $X = (x_1, x_2, ..., x_n)$ y $Y = (y_1, y_2, ..., y_n)$ se escribe como $d_{k,p}(X, Y)$ [Zarinbal, 2009], que se expresa como:

$$d_{k,p}(X, Y) = \left(\sum_{i=1}^{n} k_i |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{8}$$

Si $k_1 = k_2 = ... = k_n = k_p$ entonces se tiene la ecuación $d_{k,p}$ ponderada:

$$d_{k,p}(X, Y) = K \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{9}$$

Si $k_1 = k_2 = ... = k_n = 1$ entonces se tiene la ecuación:

$$d_{k,p}(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{10}$$

donde $(x_1, x_2, ..., x_n)$ y $(y_1, y_2, ...y_n)$ son dos vectores en el espacio euclidiano de dimensión n, y p es un número real positivo. Es una generalización de todas las métricas de uso común en un espacio euclidiano.

3.2. Distancia Euclidiana

La distancia euclidiana es a menudo conocida como la distancia ordinaria, ya que sigue la idea de una línea recta entre dos puntos en el espacio euclidiano, en la literatura antigua se le conocía como métrica pitagórica.

El ejemplo más básico, es la distancia euclidiana entre dos puntos $X = (x_1, x_2, ..., x_n)$ y $Y = (y_1, y_2, ...y_n)$, la cual viene dada por la ecuación:

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2} \tag{11}$$

La Ecuación (11) también se puede escribir como sigue:

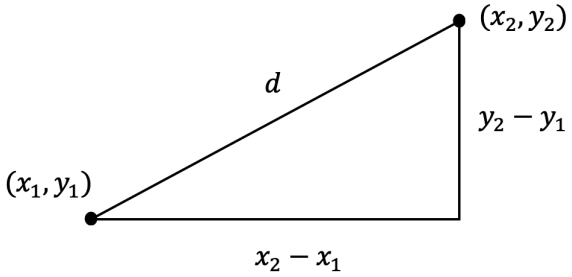
$$d(X, Y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{12}$$

La Ecuación (12) es exactamente lo que se obtiene al sustituir el valor $p = 2$ en la Ecuación (10), dando como resultado:

$$d(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{\frac{1}{2}} \tag{13}$$

Para lograr observar lo mencionado en las ecuaciones anteriores, es muy intuitivo y sencillo de representar gráficamente en dos dimensiones la distancia entre dos puntos, utilizando la distancia euclidiana tal como se muestra en la Figura 1.

Figura 1: Distancia euclidiana entre dos puntos.



El cuadrado de la distancia euclidiana se conoce como la distancia euclidiana cuadrada o cuadrática, es de gran importancia para estimar parámetros en modelos estadísticos, por ejemplo, es utilizada en el método de mínimos cuadrados para el análisis de regresión, dicha distancia se puede expresar de la siguiente manera:

$$d^2(X,Y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2 \tag{14}$$

Lo anterior se puede expresar como se muestra a continuación:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|^2 \tag{15}$$

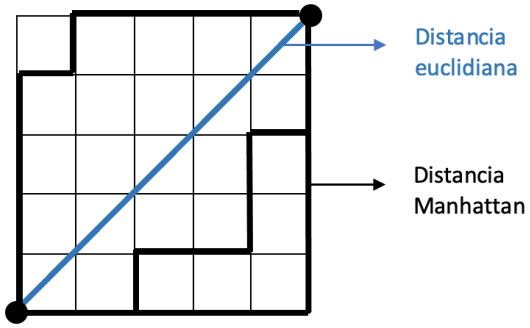
Tanto la distancia euclidiana como la distancia euclidiana cuadrática, son de las más utilizadas en problemas de clasificación, incluso es la elección casi predeterminada al utilizar el algoritmo K-medias.

3.3. Distancia Manhattan

La distancia Manhattan (en inglés llamada taxicab), también conocida como rectilínea o norma l_1, obtiene su nombre del diseño en cuadrícula de las calles en la isla de Manhattan, reflejando la distancia que un auto tendría que recorrer al circular por las calles de dicha ciudad para llegar de X a Y.

Para poder observar de una manera más clara lo que esto significa, y a su vez, poder contrastarlo con la métrica de la sección anterior, en la Figura 2 se aprecia el camino a seguir utilizando la distancia euclidiana, y algunas rutas posibles utilizando la distancia Manhattan.

Figura 2: Comparación entre distancia euclidiana y distancia Manhattan.



La distancia entre dos puntos X y Y con n dimensiones se puede expresar como sigue:

$$d(X,Y) = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n| \tag{16}$$

Como se acaba de mencionar se le conoce también como norma l_1, esto es debido a que si sustituimos el valor de $p = 1$ en (10) obtendremos entonces la ecuación de la distancia Manhattan:

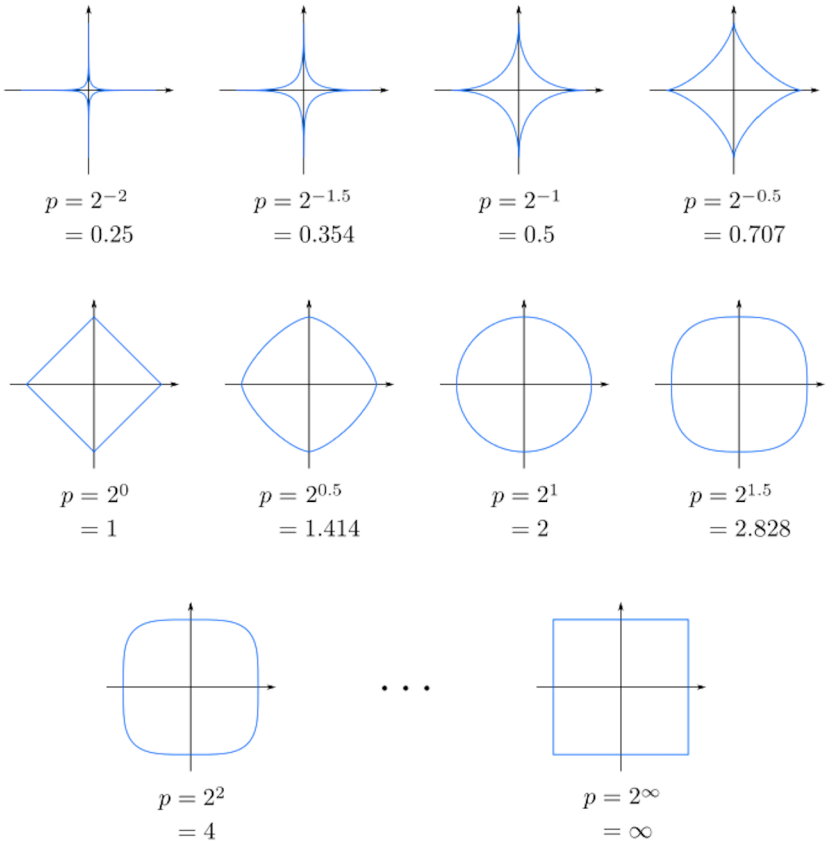
$$d(X, Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{17}$$

3.4. Semimétricas

Ahora que se revisó la ecuación de la distancia Minkowski, donde comúnmente p toma valores enteros mayores o igual a 1, como puede verse si $p = 1$ se obtiene la distancia Manhattan, pero si $p = 2$ se consigue la distancia euclidiana; si se utilizan valores menores a 1, como: $p = 0.1$, 0.3, 0.5, 0.7 y 0.9, es decir, los valores escogidos para este estudio, lo que produce es la definición de una semimétrica, ya que esta distancia cumple los primeros 3 axiomas pero no necesariamente con la desigualdad triangular.

Puede que sea más sencillo de observar lo que se obtiene al utilizar estos valores de p menores a 1, si se introduce la definición de un círculo: un círculo es un conjunto de puntos con una distancia fija llamada radio, hacia otro punto llamado centro, tal como se observa en la Figura 3.

Figura 3: Círculos unitarios con varios valores de p utilizando distancia Minkowski.



Love [Love and Dowling, 1985] realizó un estudio para lograr el mejor ajuste de los parámetros de (8), es decir, tanto el parámetro k como el parámetro p, el cual exactamente se hace variar en este estudio, en sus resultados se observa que el ajuste para el valor de p, que obtenía mejores resultados era menor a 1, en la mayoría de los casos rondando el valor de 0.9, lo que motivó la exploración en este trabajo de utilizar valores menores a 1; a pesar de que en este estudio no se utiliza el parámetro k que se refiere a un peso o ponderación.

Una vez que se realizó un análisis sobre los resultados con los valores de p menores a 1, (que se abordará más adelante), surgió la duda sobre el comportamiento que se obtiene al utilizar valores alrededor de las dos métricas que se sabe tienen un buen desempeño, es decir, de p = 1 y p = 0.9.

Los nuevos valores a probar alrededor de p = 1 y p = 0.9, comprenden el intervalo de (0.88, 1.02), es decir, se varió el valor de p desde dos centésimas por debajo del valor de p = 0.9, así como todos los valores intermedios entre este valor y el valor de p = 1, además de dos centésimas por encima del valor de p = 1, generando como resultado la prueba con 15 valores diferentes de p, para observar más a detalle el comportamiento obtenido por los clasificadores utilizando estos valores.

Se puede especificar que, en la ecuación que se utilizó para probar diferentes valores de p se eliminó la raíz cuadrada, ya que es una función creciente para $p > 0$, dando como resultado la siguiente ecuación:

$$d(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right) \tag{18}$$

3.5. Distancia Híbrida

Para hacer aún más exhaustivo el estudio sobre el comportamiento que tienen los clasificadores con los dos valores del parámetro p, que se comentaron en la sección anterior, se propone la utilización de una combinación lineal de distancias, que identificaremos a partir de este punto como distancia híbrida.

La distancia híbrida podría implementarse de diferentes maneras, por ejemplo, Dalatu [Dalatu et al., 2017] propone utilizar la distancia Manhattan y la distancia Chebyshev, que por definición es el promedio de ambas distancias aplicadas a cada par de puntos.

La distancia híbrida también puede definirse de otra manera, Yang [Yang et al., 2017] y Ulucay [Uluçay et al., 2019], introducen la letra α en la distancia híbrida, esta letra se establece con un peso y cuyo valor determina la proporción de la distancia Manhattan, así como de la segunda distancia en la distancia híbrida propuesta.

La distancia híbrida fue definida como: la combinación de la distancia Manhattan obtenida a partir de la distancia Minkowski al utilizar el valor de p = 1:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{19}$$

así como la distancia obtenida de utilizar el valor de p = 0.9:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i|^{0.9} \tag{20}$$

y al combinarlas se obtiene finalmente la ecuación a utilizar para la distancia híbrida:

$$d_H(X,Y) = \sum_{i=1}^{n} \Big((\alpha)(|x_i - y_i|) + (1-\alpha)(|x_i - y_i|)^{0.9} \Big) \tag{21}$$

El parámetro α en (21) inicia con un valor de 0, e irá aumentando con cambios de 0.1 hasta 1, ahora que se tiene la ecuación es más claro lo que se obtiene al variar el valor de α. Si α toma el valor de 0, lo que se obtiene es exactamente la distancia al utilizar únicamente el valor de p = 0.9, mientras que cuando α toma el valor de 1, el resultante es la distancia que se obtendría al utilizar únicamente el valor de p = 1 o la distancia Manhattan.

Con valores intermedios, por ejemplo, α = 0.2 implica que la proporción de la distancia Manhattan es de 0.2, por lo tanto, la proporción de la distancia con p = 0.9 es de 1 - α; cuando el parámetro α toma el valor de 0.5, significa que, la distancia obtenida para cada par de puntos, es la mitad en proporción de distancia Manhattan y la mitad de p = 0.9.

Se decidió utilizar los valores intermedios entre 0 y 1 para el parámetro α, a fin de poder observar si la distancia híbrida, mejora el desempeño obtenido por ambos clasificadores, al cambiar la proporción de las dos distancias elegidas.

3.6. Preparación de las bases de datos

Durante este estudio, para utilizar ambos clasificadores se llevó a cabo un proceso de preparación de las siete bases de datos, el proceso cuenta con las siguientes etapas:

- Limpieza de datos faltantes.

  En las bases de datos a utilizar en este estudio, en algunos casos existían elementos con datos faltantes, ya que se van a utilizar dos clasificadores distintos, y el comportamiento derivado de la falta de estos datos puede ser diferente para cada uno de ellos, se decidió excluir

dichos elementos que muestran datos faltantes; para hacer más homogéneas las pruebas y, que sea solamente el cambio de la métrica lo que indique si existe un cambio o no en el comportamiento de los clasificadores.

- Ponderación de columnas alfanuméricas.

  En las bases de datos existen columnas con valores alfanuméricos, es decir, no todos los valores que se encuentran en ellas son solamente números. Se pueden localizar (en estas bases de datos), categorías tales como: ”alto, medio, bajo”, por ejemplo, en Zoo se encuentran ”grande y pequeño”, así como en la base de datos de Titanic la supervivencia de un pasajero se debe sustituir por un valor numérico tal como 0 para "muerto" y 1 para "vivo", está sustitución se realizó en todos los casos donde existían valores que no eran exclusivamente numéricos.

- Normalización de los datos.

  Entre los atributos de cada elemento, se puede distinguir que, algunas columnas tienen escalas mayores a las de otros, como consecuencia los datos no son uniformes. En este estudio se considera que todos los atributos son igual de importantes para la clasificación, por lo tanto, se deben someter todas las bases de datos a un proceso de normalización, y así evitar la influencia de los atributos con mayor escala.

  Existen varias técnicas para la normalización de datos, entre ellas están min-max, z-score y por escala decimal, habitualmente la primera es la más utilizada; pero es de libre elección para cada investigador la técnica a utilizar.

  En este estudio se decidió utilizar la primera técnica, este proceso de normalización consiste en tomar el valor mínimo y máximo de cada atributo, una vez obtenidos estos límites superior e inferior, los demás valores se sustituyen con la siguiente fórmula, garantizando que los valores del atributo tomarán valores entre 0 y 1:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{22}$$

3.7. Pseudocódigo de CGS

El primer clasificador que se utilizó en este estudio es la implementación del problema de coloración de gráficas suaves (CGS) descrito en el Algoritmo 1; el número de colores se busca mediante la solidez y la resiliencia, pero dado que se quiere observar el rendimiento con distintas métricas, el número de colores a utilizar se establece desde un principio, determinando así el número correcto de colores o clases para clasificar a cada base de datos.

Algoritmo 1: Pseudocódigo del clasificador con CGS.

```
Se establece el número de colores a utilizar
Calcular la matriz de distancias
para cada elemento en base de datos hacer
    Calcula la distancia entre el elemento y todos los demás
fin
Se genera la primer coloración
para cada elemento en base de datos hacer
    Asigna un color a elemento
fin
Calcular la dureza que viene dada por la suma de las penalizaciones
para cada elemento i en elementos hacer
    para cada elemento j en elementos hacer
        si color de elemento i == color de elemento j entonces
            Se agrega el valor de la penalización a la dureza
        fin
    fin
fin
mientras iteración < parámetro de paro hacer
    Se selecciona aleatoriamente uno de los elementos
    Se le asigna un color
    si color nuevo == color anterior entonces
        Asigna otro color
    fin
    Se calcula la nueva dureza
    si nueva dureza < dureza anterior entonces
        Se establece la nueva coloración como solución actual
        en otro caso
            Se conserva la coloración anterior
        fin
    fin
fin
```

3.8. Pseudocódigo de K-medias

El segundo clasificador que se utilizó en este estudio es la implementación del algoritmo K-medias descrito en el Algoritmo 2, que como se mencionó anteriormente, consta de 4 pasos que se repiten hasta la convergencia. A pesar de que la elección de un número correcto de k centroides es de vital importancia para este algoritmo, no es el punto crucial de este estudio, por lo que el valor de k que se establece para cada base de datos, es el valor correcto de centroides. Lo que se desea observar es el rendimiento del clasificador con distintas métricas.

Algoritmo 2: Pseudocódigo del clasificador con K-medias.

```
Establecer el valor de k
para i =1 hasta k hacer
    Genera un centroide de forma aleatoria
fin
mientras diferencia > 0 hacer
    para cada elemento en base de datos hacer
        para cada centroide en centroides hacer
            Calcula la distancia de elemento con centroide
        fin
        Asigna el elemento al grupo del centroide más cercano
    fin
    Se deben recalcular la posición de los centroides
    para cada grupo hacer
        Se obtiene la media aritmética de las posiciones de los elementos asignados al
         grupo
        Asigna lo obtenido como nuevo centroide
    fin
    Se debe verificar la condición de paro
    para cada centroide hacer
        Se calcula la diferencia entre el nuevo centroide y el anterior correspondiente
        si diferencia == 0 entonces
            Detener el algoritmo
        fin
    fin
fin
```

Capítulo 4. Análisis de resultados

4.1. Resultados con $p > 1$.

Para efectos de este estudio, se calcularon las matrices de distancia de cada una de las bases de datos, en las cuales se probó el algoritmo de CGS utilizando la ecuación (18) con los valores de p, mencionados en la sección anterior, es decir, $p = 1$, 2, 3 y 4, mientras que para el algoritmo K-medias se utilizaron al momento de calcular la distancia de cada observación, a los k centroides.

Cabe mencionar, que se ejecutó el algoritmo 30 veces para cada valor de p, es decir, con las diferentes métricas se puede observar la tendencia que tienen los resultados al ir aumentando el valor del exponente.

En la sección Apéndice, pueden encontrarse de manera gráfica mediante diagramas de caja, los resultados que se muestran en tablas a continuación.

4.1.1. Resultados de Iris con $p > 1$

Tabla 2: Precisión de CGS en Iris con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	0.624s	**92.06 %**
p = 2	3	0.672s	89.33 %
p = 3	3	0.654s	86.66 %
p = 4	3	0.700s	86.00 %

En la Tabla 2 se muestran los porcentajes de precisión con CGS en Iris, el mejor se obtuvo utilizando el valor de $p = 1$ (Manhattan), con 92.06 % de precisión, y respaldado por la prueba ANOVA indica que efectivamente ese valor de p da resultados significativamente mejores a los demás.

Tabla 3: Precisión de K-medias en Iris con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	0.017s	**86.39 %**
p = 2	3	0.017s	84.28 %
p = 3	3	0.013s	83.73 %
p = 4	3	0.015s	82.15 %

En la Tabla 3 se observan los porcentajes de precisión con K-medias en Iris, el mejor se obtuvo al utilizar el valor de $p = 1$ (Manhattan), con 86.39 % de precisión, pero mediante la prueba ANOVA muestra que todas las medias de utilizar esos valores en K-medias, son estadísticamente iguales.

4.1.2. Resultados de Wine con $p > 1$

Tabla 4: Precisión de CGS en Wine con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	1.070s	**93.82 %**
p = 2	3	1.047s	93.25 %
p = 3	3	1.061s	93.32 %
p = 4	3	1.047s	92.22 %

En la Tabla 4 se muestran los porcentajes de precisión con CGS en Wine, el mejor se obtuvo utilizando el valor de $p = 1$, con 93.82 % de precisión, mientras que la prueba ANOVA, indica que este valor de p obtiene resultados significativamente mejores.

Tabla 5: Precisión de K-medias en Wine con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	0.040s	**94.79 %**
p = 2	3	0.052s	92.26 %
p = 3	3	0.042s	89.98 %
p = 4	3	0.044s	89.21 %

En la Tabla 5 se observan los porcentajes de precisión con K-medias en Wine, el mejor se obtuvo utilizando el valor de $p = 1$, con 94.79 % de precisión, y una vez realizada la prueba ANOVA muestra que tanto $p = 1$ y 2 obtienen medias de sus resultados iguales pero mejores que con 3 y 4.

4.1.3. Resultados de Zoo con $p > 1$

Tabla 6: Precisión de CGS en Zoo con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	4	0.330s	**78.61 %**
p = 2	4	0.259s	76.82 %
p = 3	4	0.258s	73.65 %
p = 4	4	0.285s	76.86 %

En la Tabla 6 se muestran los porcentajes de precisión con CGS en Zoo, el mejor se obtuvo al utilizar el valor de $p = 1$, con 78.61 % de precisión, pero lo que la prueba ANOVA dice es que con cualquiera de los valores de p se obtienen medias iguales.

Tabla 7: Precisión de K-medias en Zoo con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	4	0.021s	**90.79 %**
p = 2	4	0.017s	83.53 %
p = 3	4	0.021s	81.82 %
p = 4	4	0.020s	80.19 %

En la Tabla 7 se observan los porcentajes de precisión con K-medias en Zoo, el mejor se obtuvo utilizando el valor de $p = 1$, con 90.79 % de precisión, al realizar la prueba ANOVA muestra que con este valor de p se obtienen resultados significativamente mejores que con los demás.

4.1.4. Resultados de Stone Flakes con $p > 1$

Tabla 8: Precisión de CGS en Stone Flakes con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	0.177s	78.57 %
p = 2	3	0.195s	81.76 %
p = 3	3	0.180s	**82.85 %**
p = 4	3	0.191s	80.76 %

En la Tabla 8 se muestran los porcentajes de precisión con CGS en Stone Flakes, el mejor se obtuvo al usar el valor de $p = 3$, con 82.85 % de precisión, la prueba ANOVA indica que el resultado de este valor es significativamente mejor a los otros valores de p utilizados.

Tabla 9: Precisión de K-medias en Stone Flakes con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	3	0.009s	**78.47 %**
p = 2	3	0.010s	78.04 %
p = 3	3	0.011s	78.33 %
p = 4	3	0.009s	77.42 %

En la Tabla 9 se observan los porcentajes de precisión con K-medias en Stone Flakes, el mejor se obtuvo al utilizar el valor de $p = 1$, con 78.47 % de precisión, pero es un tanto obvio por lo parecido de los resultados, que la prueba ANOVA indica que para todos los valores de p se obtienen medias iguales.

4.1.5. Resultados de Titanic con $p > 1$

Tabla 10: Precisión de CGS en Titanic con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	2	15.15s	**78.57 %**
p = 2	2	13.88s	77.25 %
p = 3	2	13.73s	77.19 %
p = 4	2	13.70s	**78.57 %**

En la Tabla 10 se observan los porcentajes de precisión con CGS en Titanic, el mejor se obtuvo con dos valores, utilizando el valor de $p = 1$ y el valor $p = 4$, con 78.57 % de precisión para ambos, pero en este caso, la prueba ANOVA muestra que los 4 valores utilizados dan medias iguales.

Tabla 11: Precisión de K-medias en Titanic con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	2	0.051s	**77.13 %**
p = 2	2	0.043s	76.00 %
p = 3	2	0.038s	73.82 %
p = 4	2	0.041s	73.07 %

En la Tabla 11 se muestran los porcentajes de precisión con K-medias en Titanic, el mejor se obtuvo al usar el valor de $p = 1$, con 77.13 % de precisión, al igual que en el experimento anterior la prueba ANOVA indica que todos los valores de p dan como resultado medias iguales.

4.1.6. Resultados de Bach con $p > 1$

Tabla 12: Precisión de CGS en Bach con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	17	355.9s	**70.03 %**
p = 2	17	346.3s	68.82 %
p = 3	17	344.8s	68.81 %
p = 4	17	348.6s	69.61 %

En la Tabla 12 se observan los porcentajes de precisión con CGS en Bach, el mejor se obtuvo utilizando el valor de $p = 1$, con 70.03 % de precisión, pero una vez realizada la prueba ANOVA correspondiente, muestra que todos los valores de p utilizados dan como resultado medias iguales.

Tabla 13: Precisión de K-medias en Bach con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	17	9.194s	**67.18 %**
p = 2	17	7.371s	66.62 %
p = 3	17	8.104s	63.66 %
p = 4	17	9.832s	61.35 %

En la Tabla 13 se muestran los porcentajes de precisión con K-medias en Bach, el mejor se obtuvo utilizando el valor de $p = 1$, con 67.18 % de precisión, al utilizar la prueba ANOVA indica que, tanto el valor de $p = 1$ y $p = 2$ obtienen medias iguales y mejores que las de $p = 3$ y 4.

4.1.7. Resultados de Celeb con $p > 1$

Tabla 14: Precisión de CGS en Celeb con $p > 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 1	2	179.3s	**77.85 %**
p = 2	2	165.6s	76.11 %
p = 3	2	160.1s	71.17 %
p = 4	2	159.8s	67.07 %

En la Tabla 14 se observan los porcentajes de precisión con CGS en Celeb, el mejor se obtuvo al utilizar el valor de $p = 1$, con 77.85 % de precisión, una vez realizada la prueba ANOVA correspondiente muestra que este valor es significativamente mejor a los otros valores de p utilizados.

Tabla 15: Precisión de K-medias en Celeb con $p > 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 1	2	1.010s	**79.01 %**
p = 2	2	1.458s	76.97 %
p = 3	2	1.649s	75.53 %
p = 4	2	1.727s	73.28 %

En la Tabla 15 se muestran los porcentajes de precisión con K-medias en Celeb, el mejor se obtuvo al usar el valor de $p = 1$, con 79.01 % de precisión, la prueba ANOVA indica que este valor de p obtiene resultados significativamente mejores.

4.2. Resultados con $0 < p < 1$

Por otro lado, al observar la tendencia cuando el valor p se aleja de 1, surge la duda sobre lo que sucede al utilizar valores entre 0 y 1, por lo que se prosiguió a ejecutar de nuevo los algoritmos con nuevos valores de p, en está ocasión con $p = 0.1, 0.3, 0.5, 0.7$ y 0.9.

4.2.1. Resultados de Iris con $0 < p < 1$

Tabla 16: Precisión de CGS en Iris con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.777s	86.06 %
p = 0.3	3	0.757s	94.35 %
p = 0.5	3	0.650s	**94.55 %**
p = 0.7	3	0.661s	94.37 %
p = 0.9	3	0.645s	92.37 %

En la Tabla 16 se observan los porcentajes de precisión con CGS en Iris, el mejor se obtuvo utilizando el valor de $p = 0.5$, con 94.55 % de precisión, al realizar la prueba ANOVA se muestra que las medias de los valores $p = 0.3$, 0.5 y 0.7 son iguales y mejores que las obtenidas con 0.1 y 0.9.

Tabla 17: Precisión de K-medias en Iris con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.017s	84.00 %
p = 0.3	3	0.017s	85.55 %
p = 0.5	3	0.015s	**87.97 %**
p = 0.7	3	0.013s	87.73 %
p = 0.9	3	0.013s	85.35 %

En la Tabla 17 se muestran los porcentajes de precisión con K-medias en Iris, el mejor se obtuvo al utilizar el valor $p = 0.5$, con 87.97 % de precisión, la prueba ANOVA señala que los valores de $p = 0.5$ y 0.7 obtienen medias iguales y mejores que con los otros 3 valores.

4.2.2. Resultados de Wine con $0 < p < 1$

Tabla 18: Precisión de CGS en Wine con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.777s	92.00 %
p = 0.3	3	0.757s	92.39 %
p = 0.5	3	0.650s	93.46 %
p = 0.7	3	0.661s	93.82 %
p = 0.9	3	0.645s	**94.38 %**

En la Tabla 18 se observan los porcentajes de precisión con CGS en Wine, el mejor se obtuvo utilizando el valor $p = 0.9$, con 94.38 % de precisión, la prueba ANOVA muestra que con los valores de $p = 0.5$, 0.7 y 0.9 se obtienen medias iguales, pero mejores a las de 0.1 y 0.3.

Tabla 19: Precisión de K-medias en Wine con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.017s	90.65 %
p = 0.3	3	0.017s	93.03 %
p = 0.5	3	0.015s	94.55 %
p = 0.7	3	0.013s	94.94 %
p = 0.9	3	0.013s	**95.24 %**

En la Tabla 19 se muestran los porcentajes de precisión con K-medias en Wine, el mejor se obtuvo al usar el valor $p = 0.9$, con 95.24 % de precisión, y respaldado por la prueba de ANOVA señala que este valor efectivamente da los mejores resultados y que son diferentes a los obtenidos de los otros valores p.

4.2.3. Resultados de Zoo con $0 < p < 1$

Tabla 20: Precisión de CGS en Zoo con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	4	0.299s	**80.27 %**
p = 0.3	4	0.258s	79.48 %
p = 0.5	4	0.239s	77.34 %
p = 0.7	4	0.333s	78.73 %
p = 0.9	4	0.236s	77.65 %

En la Tabla 20 se muestran los porcentajes de precisión con CGS en Zoo, el mejor se obtuvo utilizando el valor de $p = 0.1$, con 80.27 % de precisión, pero lo que indica la prueba ANOVA es que con cualquiera de estos valores de p se obtienen medias iguales.

Tabla 21: Precisión de K-medias en Zoo con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	4	0.177s	69.62 %
p = 0.3	4	0.175s	79.68 %
p = 0.5	4	0.083s	89.96 %
p = 0.7	4	0.048s	88.73 %
p = 0.9	4	0.039s	**90.23 %**

En la Tabla 21 se observan los porcentajes de precisión con K-medias en Zoo, el mejor se obtuvo al utilizar el valor de $p = 0.9$, con 90.23 % de precisión, la prueba ANOVA muestra que con los valores de $p = 0.5$, 0.7 y 0.9 se obtienen medias iguales y mejores a las obtenidas con 0.1 y 0.3.

4.2.4. Resultados de Stone Flakes con $0 < p < 1$

Tabla 22: Precisión de CGS en Stone Flakes con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.185s	67.42 %
p = 0.3	3	0.189s	73.80 %
p = 0.5	3	0.167s	75.38 %
p = 0.7	3	0.173s	76.23 %
p = 0.9	3	0.173s	**77.66 %**

En la Tabla 22 se muestran los porcentajes de precisión con CGS en Stone Flakes, el mejor se obtuvo utilizando el valor de $p = 0.9$, con 77.66 % de precisión, lo que indica la prueba ANOVA es que efectivamente este valor da el mejor resultado y es diferente a los demás valores de p utilizados.

Tabla 23: Precisión de K-medias en Stone Flakes con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	3	0.012s	78.71 %
p = 0.3	3	0.009s	78.09 %
p = 0.5	3	0.010s	78.66 %
p = 0.7	3	0.008s	78.19 %
p = 0.9	3	0.008s	**78.76 %**

En la Tabla 23 se muestran los porcentajes de precisión con K-medias en Stone flakes, el mejor se obtuvo utilizando el valor de $p = 0.9$, con 78.76 % de precisión, como era de esperarse al observar los resultados, la prueba ANOVA arroja que todos los valores dan medias iguales.

4.2.5. Resultados de Titanic con $0 < p < 1$

Tabla 24: Precisión de CGS en Titanic con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	2	13.69s	73.31 %
p = 0.3	2	13.57s	74.07 %
p = 0.5	2	13.60s	75.01 %
p = 0.7	2	13.48s	75.18 %
p = 0.9	2	15.74s	**75.46 %**

En la Tabla 24 se observan los porcentajes de precisión con CGS en Titanic, el mejor se obtuvo al utilizar el valor de $p = 0.9$, con 75.46 % de precisión, al realizar la prueba de ANOVA muestra que con todos los valores de p utilizados se obtienen medias iguales.

Tabla 25: Precisión de K-medias en Titanic con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	2	0.105s	70.26 %
p = 0.3	2	0.085s	72.38 %
p = 0.5	2	0.064s	75.10 %
p = 0.7	2	0.069s	75.54 %
p = 0.9	2	0.047s	**77.01 %**

En la Tabla 25 se muestran los porcentajes de precisión con K-medias en Titanic, el mejor se obtuvo utilizando el valor de $p = 0.9$, con 77.01 % de precisión, la prueba de ANOVA indica que los valores de $p = 0.5, 0.7, 0.9$ dan medias iguales y mejores que las obtenidas con 0.1 y 0.3.

4.2.6. Resultados de Bach con $0 < p < 1$

Tabla 26: Precisión de CGS en Bach con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	17	343.1s	67.74 %
p = 0.3	17	344.4s	68.39 %
p = 0.5	17	343.7s	69.09 %
p = 0.7	17	342.5s	69.36 %
p = 0.9	17	340.8s	**69.63 %**

En la Tabla 26 se observan los porcentajes de precisión con CGS en Bach, el mejor se obtuvo al usar el valor de $p = 0.9$, con 69.63 % de precisión, pero al realizar la prueba de ANOVA muestra que con todos los valores de p se obtienen medias iguales.

Tabla 27: Precisión de K-medias en Bach con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	17	29.93s	23.02 %
p = 0.3	17	30.10s	25.07 %
p = 0.5	17	29.66s	50.92 %
p = 0.7	17	29.58s	67.93 %
p = 0.9	17	20.33s	**68.60 %**

En la Tabla 27 se muestran los porcentajes de precisión con K-medias en Bach, el mejor se obtuvo utilizando el valor de $p = 0.9$, con 68.60 % de precisión, al realizar la prueba de ANOVA indica que los valores de $p = 0.7$, 0.9 dan medias iguales y mejores que las obtenidas con 0.1, 0.3 y 0.5.

4.2.7. Resultados de Celeb con $0 < p < 1$

Tabla 28: Precisión de CGS en Celeb con $0 < p < 1$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	2	175.5s	73.76 %
p = 0.3	2	159.4s	76.19 %
p = 0.5	2	165.7s	77.37 %
p = 0.7	2	180.3s	77.62 %
p = 0.9	2	180.2s	**77.87 %**

En la Tabla 28 se observan los porcentajes de precisión con CGS en Celeb, el mejor se obtuvo al utilizar el valor de $p = 0.9$, con 77.87 % de precisión, y al realizar la prueba de ANOVA señala que efectivamente este valor da el mejor resultado y es significativamente diferente a los demás valores de p utilizados.

Tabla 29: Precisión de K-medias en Celeb con $0 < p < 1$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.1	2	1.889s	75.41 %
p = 0.3	2	1.316s	76.57 %
p = 0.5	2	1.084s	77.64 %
p = 0.7	2	1.155s	78.30 %
p = 0.9	2	1.195s	**78.71 %**

En la Tabla 29 se muestran los porcentajes de precisión con K-medias en Celeb, el mejor se obtuvo utilizando el valor de $p = 0.9$, con 78.71 % de precisión, al realizar la prueba de ANOVA indica que el valor de $p = 0.9$ efectivamente da resultados significativamente mejores a los demás.

4.3. Resultados con $0.88 \leq p \leq 1.02$

De los resultados obtenidos en los experimentos anteriores, se detectó que al utilizar la distancia Manhattan se obtienen los mejores resultados; mientras que al utilizar valores menores a 1, se identificó el valor de $p = 0.9$ como el que mejores resultados otorga, por ende es de interés lo que sucede con valores muy cercanos a los dos mencionados anteriormente, y también en todos los valores entre uno y otro con un cambio de una centésima, es decir se probarán los valores de $p =$ 0.88, 0.89, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1, 1.01 y 1.02.

4.3.1. Resultados de Iris con $0.88 \leq p \leq 1.02$

Tabla 30: Precisión de CGS en Iris con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	0.668s	91.60 %
p = 0.89	3	0.651s	92.37 %
p = 0.90	3	0.645s	92.37 %
p = 0.91	3	0.628s	92.22 %
p = 0.92	3	0.609s	92.06 %
p = 0.93	3	0.646s	91.13 %
p = 0.94	3	0.653s	92.06 %
p = 0.95	3	0.622s	91.91 %
p = 0.96	3	0.627s	91.28 %
p = 0.97	3	0.664s	91.60 %
p = 0.98	3	0.665s	91.44 %
p = 0.99	3	0.620s	91.28 %
p = 1.00	3	0.624s	92.06 %
p = 1.01	3	0.638s	91.28 %
p = 1.02	3	0.645s	91.13 %

En la Tabla 30 se muestran los porcentajes de precisión con CGS en Iris, por medio de una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

Tabla 31: Precisión de K-medias en Iris con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	0.012s	86.55 %
p = 0.89	3	0.012s	86.62 %
p = 0.90	3	0.013s	85.35 %
p = 0.91	3	0.013s	86.42 %
p = 0.92	3	0.013s	85.44 %
p = 0.93	3	0.014s	82.02 %
p = 0.94	3	0.014s	85.39 %
p = 0.95	3	0.013s	86.44 %
p = 0.96	3	0.014s	82.28 %
p = 0.97	3	0.014s	84.24 %
p = 0.98	3	0.013s	83.22 %
p = 0.99	3	0.013s	83.46 %
p = 1.00	3	0.017s	86.39 %
p = 1.01	3	0.012s	85.44 %
p = 1.02	3	0.011s	82.17 %

En la Tabla 31 se observan los porcentajes de precisión con K-medias en Iris, al utilizar una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de los valores de p utilizados son iguales.

4.3.2. Resultados de Wine con $0.88 \leq p \leq 1.02$

Tabla 32: Precisión de CGS en Wine con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	1.031s	94.38 %
p = 0.89	3	1.049s	94.38 %
p = 0.90	3	1.061s	94.38 %
p = 0.91	3	1.052s	94.38 %
p = 0.92	3	1.031s	93.82 %
p = 0.93	3	1.026s	93.82 %
p = 0.94	3	1.042s	93.82 %
p = 0.95	3	1.079s	93.82 %
p = 0.96	3	1.029s	93.82 %
p = 0.97	3	1.053s	93.82 %
p = 0.98	3	1.037s	93.82 %
p = 0.99	3	1.078s	93.82 %
p = 1.00	3	1.070s	93.82 %
p = 1.01	3	1.025s	93.82 %
p = 1.02	3	1.057s	93.82 %

En la Tabla 32 se muestran los porcentajes de precisión con CGS en Wine, y respaldado por una prueba ANOVA se puede afirmar que el intervalo de $p = 0.88$ a 0.91, obtiene medias iguales y mejores a las de $p = 0.92$ a 1.02.

Tabla 33: Precisión de K-medias en Wine con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	0.044s	93.98 %
p = 0.89	3	0.040s	95.28 %
p = 0.90	3	0.037s	95.24 %
p = 0.91	3	0.046s	94.00 %
p = 0.92	3	0.044s	95.29 %
p = 0.93	3	0.050s	95.18 %
p = 0.94	3	0.050s	94.15 %
p = 0.95	3	0.051s	95.35 %
p = 0.96	3	0.051s	95.35 %
p = 0.97	3	0.048s	95.22 %
p = 0.98	3	0.039s	95.74 %
p = 0.99	3	0.043s	94.51 %
p = 1.00	3	0.040s	94.79 %
p = 1.01	3	0.037s	95.86 %
p = 1.02	3	0.041s	94.92 %

En la Tabla 33 se observan los porcentajes de precisión con K-medias en Wine, al realizar una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

4.3.3. Resultados de Zoo con $0.88 \leq p \leq 1.02$

Tabla 34: Precisión de CGS en Zoo con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	4	0.230s	78.65 %
p = 0.89	4	0.236s	76.78 %
p = 0.90	4	0.236s	77.65 %
p = 0.91	4	0.238s	79.64 %
p = 0.92	4	0.260s	79.12 %
p = 0.93	4	0.239s	81.54 %
p = 0.94	4	0.237s	80.71 %
p = 0.95	4	0.246s	80.55 %
p = 0.96	4	0.239s	77.65 %
p = 0.97	4	0.235s	77.65 %
p = 0.98	4	0.235s	77.61 %
p = 0.99	4	0.239s	78.09 %
p = 1.00	4	0.330s	78.61 %
p = 1.01	4	0.239s	78.57 %
p = 1.02	4	0.262s	78.49 %

En la Tabla 34 se muestran los porcentajes de precisión con CGS en Zoo, al realizar la prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

Tabla 35: Precisión de K-medias en Zoo con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	4	0.025s	92.10 %
p = 0.89	4	0.030s	90.03 %
p = 0.90	4	0.039s	90.23 %
p = 0.91	4	0.020s	88.92 %
p = 0.92	4	0.020s	90.31 %
p = 0.93	4	0.020s	93.21 %
p = 0.94	4	0.022s	85.87 %
p = 0.95	4	0.021s	92.06 %
p = 0.96	4	0.029s	89.24 %
p = 0.97	4	0.019s	86.58 %
p = 0.98	4	0.026s	88.25 %
p = 0.99	4	0.018s	88.84 %
p = 1.00	4	0.021s	90.79 %
p = 1.01	4	0.018s	89.99 %
p = 1.02	4	0.021s	89.36 %

En la Tabla 35 se observan los porcentajes de precisión con K-medias en Zoo, al utilizar una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

4.3.4. Resultados de Stone Flakes con $0.88 \leq p \leq 1.02$

Tabla 36: Precisión de CGS en Stone Flakes con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	0.175s	77.52 %
p = 0.89	3	0.173s	77.38 %
p = 0.90	3	0.173s	77.66 %
p = 0.91	3	0.178s	77.99 %
p = 0.92	3	0.167s	77.85 %
p = 0.93	3	0.165s	78.23 %
p = 0.94	3	0.168s	77.90 %
p = 0.95	3	0.184s	78.57 %
p = 0.96	3	0.197s	78.57 %
p = 0.97	3	0.192s	78.57 %
p = 0.98	3	0.173s	78.57 %
p = 0.99	3	0.174s	78.57 %
p = 1.00	3	0.177s	78.57 %
p = 1.01	3	0.171s	78.57 %
p = 1.02	3	0.171s	78.57 %

En la Tabla 36 se muestran los porcentajes de precisión con CGS en Stone Flakes, y utilizando ANOVA se afirma que el intervalo de $p = 0.95$ a 1.02 obtiene medias iguales y mejores a las del intervalo de $p = 0.88$ a 0.94.

Tabla 37: Precisión de K-medias en Stone Flakes con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	3	0.009s	78.71 %
p = 0.89	3	0.008s	76.76 %
p = 0.90	3	0.008s	78.76 %
p = 0.91	3	0.009s	77.04 %
p = 0.92	3	0.009s	77.28 %
p = 0.93	3	0.009s	78.23 %
p = 0.94	3	0.011s	78.57 %
p = 0.95	3	0.008s	78.09 %
p = 0.96	3	0.011s	77.47 %
p = 0.97	3	0.009s	77.90 %
p = 0.98	3	0.009s	77.85 %
p = 0.99	3	0.010s	78.00 %
p = 1.00	3	0.009s	78.47 %
p = 1.01	3	0.009s	79.23 %
p = 1.02	3	0.009s	78.71 %

En la Tabla 37 se observan los porcentajes de precisión con K-medias en Stone Flakes, y por medio de una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

4.3.5. Resultados de Titanic con $0.88 \leq p \leq 1.02$

Tabla 38: Precisión de CGS en Titanic con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	2	13.67s	75.37 %
p = 0.89	2	13.68s	73.91 %
p = 0.90	2	15.74s	75.46 %
p = 0.91	2	13.53s	74.36 %
p = 0.92	2	13.69s	75.09 %
p = 0.93	2	14.73s	73.69 %
p = 0.94	2	14.52s	75.92 %
p = 0.95	2	14.70s	74.44 %
p = 0.96	2	15.11s	74.52 %
p = 0.97	2	15.22s	74.52 %
p = 0.98	2	15.09s	75.27 %
p = 0.99	2	15.09s	73.76 %
p = 1.00	2	15.15s	78.57 %
p = 1.01	2	15.05s	74.52 %
p = 1.02	2	15.15s	73.14 %

En la Tabla 38 se muestran los porcentajes de precisión con CGS en Titanic, y al realizar la prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

Tabla 39: Precisión de K-medias en Titanic con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	2	0.060s	77.37 %
p = 0.89	2	0.058s	77.37 %
p = 0.90	2	0.047s	77.01 %
p = 0.91	2	0.050s	77.13 %
p = 0.92	2	0.053s	77.73 %
p = 0.93	2	0.054s	76.89 %
p = 0.94	2	0.049s	77.61 %
p = 0.95	2	0.056s	76.89 %
p = 0.96	2	0.052s	77.13 %
p = 0.97	2	0.053s	77.25 %
p = 0.98	2	0.047s	77.49 %
p = 0.99	2	0.043s	77.25 %
p = 1.00	2	0.051s	77.13 %
p = 1.01	2	0.054s	77.01 %
p = 1.02	2	0.061s	77.25 %

En la Tabla 39 se observan los porcentajes de precisión con K-medias en Titanic, al realizar la prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

4.3.6. Resultados de Bach con $0.88 \leq p \leq 1.02$

Tabla 40: Precisión de CGS en Bach con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	17	343.3s	69.22 %
p = 0.89	17	344.2s	68.72 %
p = 0.90	17	340.8s	69.63 %
p = 0.91	17	344.1s	68.13 %
p = 0.92	17	345.6s	68.91 %
p = 0.93	17	345.5s	69.49 %
p = 0.94	17	346.9s	67.95 %
p = 0.95	17	345.0s	69.27 %
p = 0.96	17	344.9s	69.66 %
p = 0.97	17	340.4s	69.28 %
p = 0.98	17	341.8s	69.91 %
p = 0.99	17	341.6s	69.39 %
p = 1.00	17	355.9s	70.03 %
p = 1.01	17	341.5s	69.22 %
p = 1.02	17	344.5s	68.76 %

En la Tabla 40 se muestran los porcentajes de precisión con CGS en Bach, y por medio de una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p son iguales.

Tabla 41: Precisión de K-medias en Bach con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	17	19.28s	66.84 %
p = 0.89	17	18.81s	66.82 %
p = 0.90	17	20.33s	68.60 %
p = 0.91	17	20.36s	67.89 %
p = 0.92	17	18.75s	67.42 %
p = 0.93	17	20.48s	66.72 %
p = 0.94	17	16.67s	66.88 %
p = 0.95	17	14.42s	67.71 %
p = 0.96	17	13.96s	67.12 %
p = 0.97	17	16.17s	67.14 %
p = 0.98	17	16.15s	66.64 %
p = 0.99	17	12.46s	68.07 %
p = 1.00	17	09.19s	67.18 %
p = 1.01	17	13.56s	68.01 %
p = 1.02	17	12.10s	67.28 %

En la Tabla 41 se observan los porcentajes de precisión con K-medias en Bach, por medio de una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados en este experimento son iguales.

4.3.7. Resultados de Celeb con $0.88 \leq p \leq 1.02$

Tabla 42: Precisión de CGS en Celeb con $0.88 \leq p \leq 1.02$.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	2	171.0s	77.85 %
p = 0.89	2	177.7s	77.85 %
p = 0.90	2	180.2s	77.87 %
p = 0.91	2	171.6s	77.86 %
p = 0.92	2	171.2s	77.84 %
p = 0.93	2	186.7s	77.84 %
p = 0.94	2	155.4s	77.84 %
p = 0.95	2	168.7s	77.84 %
p = 0.96	2	156.5s	77.86 %
p = 0.97	2	178.5s	77.85 %
p = 0.98	2	174.9s	77.85 %
p = 0.99	2	173.8s	77.85 %
p = 1.00	2	179.3s	77.85 %
p = 1.01	2	174.5s	77.78 %
p = 1.02	2	170.6s	77.75 %

En la Tabla 42 se muestran los porcentajes de precisión con CGS en Celeb, y utilizando ANOVA indica que los valores de $p = 0.90$, 0.91 y 0.96 obtienen medias iguales y mejores a las de los demás valores de p utilizados.

Tabla 43: Precisión de K-medias en Celeb con $0.88 \leq p \leq 1.02$.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
p = 0.88	2	1.264s	77.95 %
p = 0.89	2	1.226s	77.94 %
p = 0.90	2	1.195s	78.71 %
p = 0.91	2	1.120s	78.72 %
p = 0.92	2	1.238s	78.75 %
p = 0.93	2	1.122s	78.80 %
p = 0.94	2	1.366s	78.83 %
p = 0.95	2	1.283s	78.00 %
p = 0.96	2	1.154s	78.17 %
p = 0.97	2	1.287s	78.97 %
p = 0.98	2	1.133s	78.98 %
p = 0.99	2	1.162s	79.00 %
p = 1.00	2	1.010s	79.01 %
p = 1.01	2	1.091s	79.03 %
p = 1.02	2	1.087s	78.34 %

En la Tabla 43 se observan los porcentajes de precisión con K-medias en Celeb, por medio de una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todos los valores de p utilizados son iguales.

4.4. Resultados de la distancia híbrida

Como se mencionó en el capítulo 3, se utilizó una distancia híbrida, que consiste en la combinación de el valor $p = 1$ y el valor $p = 0.9$, en la cual se introduce un parámetro α con valores entre 0 y 1, lo que resultará en observar el comportamiento desde utilizar solamente $p = 0.9$ o $p = 1$, y en el medio las combinaciones dando más peso a una u otra.

4.4.1. Resultados de Iris con la distancia híbrida

Tabla 44: Precisión de CGS en Iris con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	0.645s	92.37 %
$\alpha = 0.1$	3	0.700s	92.06 %
$\alpha = 0.2$	3	0.776s	91.75 %
$\alpha = 0.3$	3	0.683s	92.22 %
$\alpha = 0.4$	3	0.664s	91.44 %
$\alpha = 0.5$	3	0.669s	91.28 %
$\alpha = 0.6$	3	0.637s	91.75 %
$\alpha = 0.7$	3	0.676s	91.13 %
$\alpha = 0.8$	3	0.687s	91.28 %
$\alpha = 0.9$	3	0.702s	91.91 %
$\alpha = 1$	3	0.624s	92.06 %

En la Tabla 44 se observan los porcentajes de precisión con CGS y la distancia híbrida en Iris, al realizar una prueba ANOVA se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

Tabla 45: Precisión de K-medias en Iris con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	0.013s	85.35 %
$\alpha = 0.1$	3	0.017s	87.59 %
$\alpha = 0.2$	3	0.020s	86.44 %
$\alpha = 0.3$	3	0.016s	85.39 %
$\alpha = 0.4$	3	0.017s	86.53 %
$\alpha = 0.5$	3	0.018s	83.31 %
$\alpha = 0.6$	3	0.021s	87.51 %
$\alpha = 0.7$	3	0.021s	85.42 %
$\alpha = 0.8$	3	0.017s	87.62 %
$\alpha = 0.9$	3	0.021s	83.31 %
$\alpha = 1$	3	0.017s	86.39 %

En la Tabla 45 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Iris, se puede afirmar por medio de una prueba ANOVA que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

4.4.2. Resultados de Wine con la distancia híbrida

Tabla 46: Precisión de CGS en Wine con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	1.061s	94.38 %
$\alpha = 0.1$	3	1.106s	94.38 %
$\alpha = 0.2$	3	1.216s	93.82 %
$\alpha = 0.3$	3	1.086s	93.82 %
$\alpha = 0.4$	3	1.109s	93.82 %
$\alpha = 0.5$	3	1.079s	93.82 %
$\alpha = 0.6$	3	1.103s	93.82 %
$\alpha = 0.7$	3	1.093s	93.82 %
$\alpha = 0.8$	3	1.155s	93.82 %
$\alpha = 0.9$	3	1.208s	93.82 %
$\alpha = 1$	3	1.070s	93.82 %

En la Tabla 46 se observan los porcentajes de precisión con CGS y la distancia híbrida en Wine, respaldado por una prueba ANOVA se puede afirmar que con los valores de $\alpha = 0$ y 0.1, se obtienen medias iguales, y mejores a las obtenidas con los valores de $\alpha = 0.2$ hasta 1.

Tabla 47: Precisión de K-medias en Wine con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	0.037s	95.24 %
$\alpha = 0.1$	3	0.060s	95.31 %
$\alpha = 0.2$	3	0.057s	93.95 %
$\alpha = 0.3$	3	0.057s	95.26 %
$\alpha = 0.4$	3	0.056s	94.13 %
$\alpha = 0.5$	3	0.056s	95.33 %
$\alpha = 0.6$	3	0.061s	95.33 %
$\alpha = 0.7$	3	0.056s	95.31 %
$\alpha = 0.8$	3	0.057s	95.59 %
$\alpha = 0.9$	3	0.048s	95.80 %
$\alpha = 1$	3	0.040s	94.79 %

En la Tabla 47 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Wine, podemos afirmar por medio de una prueba ANOVA, que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

4.4.3. Resultados de Zoo con la distancia híbrida

Tabla 48: Precisión de CGS en Zoo con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	4	0.236s	77.65 %
$\alpha = 0.1$	4	0.246s	78.21 %
$\alpha = 0.2$	4	0.265s	77.77 %
$\alpha = 0.3$	4	0.243s	76.43 %
$\alpha = 0.4$	4	0.322s	76.86 %
$\alpha = 0.5$	4	0.257s	75.27 %
$\alpha = 0.6$	4	0.251s	77.57 %
$\alpha = 0.7$	4	0.262s	77.22 %
$\alpha = 0.8$	4	0.266s	77.85 %
$\alpha = 0.9$	4	0.252s	77.34 %
$\alpha = 1$	4	0.330s	78.61 %

En la Tabla 48 se observan los porcentajes de precisión con CGS y la distancia híbrida en Zoo, se puede afirmar por medio de una prueba ANOVA que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

Tabla 49: Precisión de K-medias en Zoo con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	4	0.039s	90.23 %
$\alpha = 0.1$	4	0.040s	90.00 %
$\alpha = 0.2$	4	0.029s	88.01 %
$\alpha = 0.3$	4	0.048s	91.03 %
$\alpha = 0.4$	4	0.041s	89.16 %
$\alpha = 0.5$	4	0.041s	88.09 %
$\alpha = 0.6$	4	0.038s	88.65 %
$\alpha = 0.7$	4	0.035s	88.69 %
$\alpha = 0.8$	4	0.052s	88.80 %
$\alpha = 0.9$	4	0.038s	89.64 %
$\alpha = 1$	4	0.021s	90.79 %

En la Tabla 49 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Zoo, se puede afirmar por medio de una prueba ANOVA, que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales para este experimento.

4.4.4. Resultados de Stone con la distancia híbrida

Tabla 50: Precisión de CGS en Stone Flakes con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	0.173s	77.66 %
$\alpha = 0.1$	3	0.174s	77.66 %
$\alpha = 0.2$	3	0.196s	77.95 %
$\alpha = 0.3$	3	0.181s	77.80 %
$\alpha = 0.4$	3	0.189s	78.09 %
$\alpha = 0.5$	3	0.217s	78.57 %
$\alpha = 0.6$	3	0.182s	78.57 %
$\alpha = 0.7$	3	0.204s	78.57 %
$\alpha = 0.8$	3	0.194s	78.57 %
$\alpha = 0.9$	3	0.186s	78.57 %
$\alpha = 1$	3	0.177s	78.57 %

En la Tabla 50 se observan los porcentajes de precisión con CGS y la distancia híbrida en Stone Flakes, al realizar la prueba ANOVA se puede afirmar que, con los valores $\alpha = 0.5$ a 1, se obtienen medias iguales, y que son mejores que las obtenidas con los valores de $\alpha = 0$ a 0.4.

Tabla 51: Precisión de K-medias en Stone Flakes con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	3	0.008s	78.76 %
$\alpha = 0.1$	3	0.013s	79.33 %
$\alpha = 0.2$	3	0.017s	79.19 %
$\alpha = 0.3$	3	0.015s	77.57 %
$\alpha = 0.4$	3	0.014s	77.95 %
$\alpha = 0.5$	3	0.016s	76.19 %
$\alpha = 0.6$	3	0.015s	77.23 %
$\alpha = 0.7$	3	0.014s	79.09 %
$\alpha = 0.8$	3	0.013s	79.76 %
$\alpha = 0.9$	3	0.014s	77.80 %
$\alpha = 1$	3	0.009s	78.47 %

En la Tabla 51 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Stone Flakes, se puede afirmar por medio de una prueba ANOVA que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

4.4.5. Resultados de Titanic con la distancia híbrida

Tabla 52: Precisión de CGS en Titanic con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	2	15.74s	75.46 %
$\alpha = 0.1$	2	15.38s	75.83 %
$\alpha = 0.2$	2	15.56s	76.94 %
$\alpha = 0.3$	2	15.55s	76.67 %
$\alpha = 0.4$	2	15.00s	76.30 %
$\alpha = 0.5$	2	15.03s	75.55 %
$\alpha = 0.6$	2	15.11s	76.39 %
$\alpha = 0.7$	2	15.83s	76.02 %
$\alpha = 0.8$	2	14.44s	75.27 %
$\alpha = 0.9$	2	15.96s	76.03 %
$\alpha = 1$	2	15.15s	78.57 %

En la Tabla 52 se observan los porcentajes de precisión con CGS y la distancia híbrida en Titanic, al realizar una prueba ANOVA, se puede afirmar que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales en este experimento.

Tabla 53: Precisión de K-medias en Titanic con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	2	0.047s	77.01 %
$\alpha = 0.1$	2	0.068s	77.49 %
$\alpha = 0.2$	2	0.079s	77.13 %
$\alpha = 0.3$	2	0.068s	77.49 %
$\alpha = 0.4$	2	0.063s	77.37 %
$\alpha = 0.5$	2	0.081s	76.53 %
$\alpha = 0.6$	2	0.065s	77.37 %
$\alpha = 0.7$	2	0.067s	77.01 %
$\alpha = 0.8$	2	0.063s	77.49 %
$\alpha = 0.9$	2	0.064s	77.49 %
$\alpha = 1$	2	0.051s	77.13 %

En la Tabla 53 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Titanic, se puede afirmar a través de una prueba ANOVA que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

4.4.6. Resultados de Bach con la distancia híbrida

Tabla 54: Precisión de CGS en Bach con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	17	340.8s	69.63 %
$\alpha = 0.1$	17	440.9s	68.82 %
$\alpha = 0.2$	17	445.1s	69.06 %
$\alpha = 0.3$	17	463.1s	68.81 %
$\alpha = 0.4$	17	460.1s	68.85 %
$\alpha = 0.5$	17	458.9s	69.17 %
$\alpha = 0.6$	17	456.8s	69.67 %
$\alpha = 0.7$	17	456.7s	69.57 %
$\alpha = 0.8$	17	459.0s	69.13 %
$\alpha = 0.9$	17	455.1s	69.54 %
$\alpha = 1$	17	355.9s	70.03 %

En la Tabla 54 se observan los porcentajes de precisión con CGS y la distancia híbrida en Bach, al realizar una prueba ANOVA se puede afirmar que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

Tabla 55: Precisión de K-medias en Bach con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	17	20.33s	68.60 %
$\alpha = 0.1$	17	27.15s	68.61 %
$\alpha = 0.2$	17	28.01s	67.49 %
$\alpha = 0.3$	17	30.53s	66.52 %
$\alpha = 0.4$	17	28.51s	66.27 %
$\alpha = 0.5$	17	26.60s	66.71 %
$\alpha = 0.6$	17	19.15s	67.62 %
$\alpha = 0.7$	17	16.58s	67.61 %
$\alpha = 0.8$	17	21.65s	67.41 %
$\alpha = 0.9$	17	18.39s	68.11 %
$\alpha = 1$	17	09.19s	67.18 %

En la Tabla 55 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Bach, se puede afirmar por medio de una prueba ANOVA que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales en este experimento.

4.4.7. Resultados de Celeb con la distancia híbrida

Tabla 56: Precisión de CGS en Celeb con distancia híbrida.

Métrica	Colores	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	2	180.2s	77.87 %
$\alpha = 0.1$	2	233.0s	77.86 %
$\alpha = 0.2$	2	218.4s	77.86 %
$\alpha = 0.3$	2	260.1s	77.84 %
$\alpha = 0.4$	2	225.3s	77.84 %
$\alpha = 0.5$	2	232.2s	77.85 %
$\alpha = 0.6$	2	211.7s	77.87 %
$\alpha = 0.7$	2	232.7s	77.84 %
$\alpha = 0.8$	2	216.1s	77.85 %
$\alpha = 0.9$	2	234.9s	77.85 %
$\alpha = 1$	2	179.3s	77.85 %

En la Tabla 56 se observan los porcentajes de precisión con CGS y la distancia híbrida en Celeb, al realizar la prueba ANOVA se puede afirmar que con los valores $\alpha = 0$, 0.1, 0.2 y 0.6, se obtienen medias iguales, y que son mejores que las obtenidas con los demás valores de α.

Tabla 57: Precisión de K-medias en Celeb con distancia híbrida.

Métrica	Centroides	Tiempo (segundos)	Porcentaje de precisión
$\alpha = 0$	2	1.195s	78.71 %
$\alpha = 0.1$	2	1.661s	78.72 %
$\alpha = 0.2$	2	1.626s	77.90 %
$\alpha = 0.3$	2	1.856s	78.78 %
$\alpha = 0.4$	2	1.614s	78.80 %
$\alpha = 0.5$	2	1.705s	78.80 %
$\alpha = 0.6$	2	1.654s	78.89 %
$\alpha = 0.7$	2	1.925s	78.98 %
$\alpha = 0.8$	2	1.796s	78.98 %
$\alpha = 0.9$	2	1.638s	79.00 %
$\alpha = 1$	2	1.010s	79.01 %

En la Tabla 57 se muestran los porcentajes de precisión con K-medias y la distancia híbrida en Celeb, al realizar una prueba ANOVA se puede afirmar que se acepta la hipótesis nula, es decir, que las medias de todas las combinaciones son iguales.

4.5. Resultados finales

Finalmente, se presentan dos tablas con los resultados de las pruebas anteriores, es decir, las métricas que obtienen los mejores resultados con cada clasificador, así como los porcentajes de precisión.

Tabla 58: Métricas con el mejor desempeño en CGS.

	Entera	Décimas	Centésimas	Híbrida
Iris	p = 0.9, 1 (92.37)	p = 0.3, 0.5, 0.7 (94.55)	p= 0.88, ... , 1.02 (92.37)	α = 0, 0.1, ... , 1 (92.37)
Wine	p = 0.9 (94.38)	p = 0.5, 0.7, 0.9 (94.38)	p= 0.88, ... , 0.91 (94.38)	α = 0, 0.1 (94.38)
Zoo	p = 0.9, 1, 2, 3, 4 (78.61)	p = 0.1, ... ,0.9 (80.27)	p= 0.88, ... , 1.02 (81.54)	α = 0, 0.1, ... , 1 (78.61)
Stone	p = 3 (82.85)	p = 0.9 (77.66)	p= 0.95, ... , 1.02 (78.57)	α = 0.5, ... , 1 (78.57)
Titanic	p = 1, 2, 3, 4 (78.57)	p = 0.1, ... , 0.9 (75.46)	p= 0.88, ... , 1.02 (78.57)	α = 0, 0.1, ... , 1 (78.57)
Bach	p = 0.9, 1, 2, 3, 4 (70.03)	p = 0.5, 0.7, 0.9 (69.63)	p= 0.88, ... , 1.02 (70.03)	α = 0, 0.1, ... , 1 (70.03)
Celeb	p = 0.9, 1 (77.87)	p = 0.9 (77.87)	p= 0.90, 0.91, 0.96 (77.87)	α = 0, 0.1, 0.2, 0.6 (77.87)

Tabla 59: Métricas con el mejor desempeño en K-medias.

	Entera	Décimas	Centésimas	Híbrida
Iris	p = 0.9, 1, 2, 3, 4 (86.39)	p = 0.5, 0.7 (87,97)	p= 0.88, ... ,1.02 (86.62)	α = 0, 0.1, ... , 1 (87.62)
Wine	p = 0.9, 1 (95.24)	p = 0.9 (95.24)	p= 0.88, ... , 1.02 (95.86)	α = 0, 0.1, ... , 1 (95.80)
Zoo	p = 0.9, 1 (90.79)	p = 0.5, 0.7, 0.9 (90.23)	p= 0.88, ... ,1.02 (93.21)	α = 0, 0.1, ... , 1 (91.03)
Stone	p = 0.9, 1, 2, 3, 4 (78.76)	p = 0.1, ... , 0.9 (78.76)	p= 0.88, ... ,1.02 (79.23)	α = 0, 0.1, ... , 1 (79.76)
Titanic	p = 0.9, 1, 2, 3, 4 (77.13)	p = 0.5, 0.7, 0.9 (77.01)	p= 0.88, ... ,1.02 (77.73)	α = 0, 0.1, ... , 1 (77.49)
Bach	p = 0.9, 1, 2 (68.60)	p = 0.7, 0.9 (68.60)	p= 0.88, ... ,1.02 (68.60)	α = 0, 0.1, ... , 1 (68.61)
Celeb	p = 0.9, 1 (79.01)	p = 0.9 (78.71)	p= 0.88, ... ,1.02 (79.03)	α = 0, 0.1, ... , 1 (79.01)

Conclusiones

En este trabajo se utilizaron dos algoritmos agrupadores no supervisados, es decir, que no necesitan ser entrenados, al contrario de un clasificador supervisado, que requiere un entrenamiento donde se le enseña como debe categorizar.

Es posible observar la forma en que se modifica la precisión, que es el parámetro que permite ver el comportamiento de las diferentes métricas, ejecutando cada algoritmo con las diferentes bases de datos de prueba.

Con los algoritmos utilizados, la distancia Manhattan que satisface completamente la definición de métrica, obtiene los mejores resultados en la mayoría de las bases de datos de prueba, a pesar de que la distancia euclidiana es la que más se usa. Se puede afirmar que estos algoritmos tienen un buen desempeño bajo condiciones de limpieza de ruido y normalización de datos.

Al observar que el porcentaje de precisión incrementa significativamente con el valor de p = 1, se prosiguió a utilizar valores de p entre 0 y 1 que corresponden a una semimétrica, ya que se satisfacen todos los axiomas con excepción de la desigualdad del triángulo, así que, de los resultados obtenidos con estos valores, se puede afirmar que en la mayoría de los casos el mejor valor utilizado es $p = 0.9$.

Una vez encontrados estos dos valores, surge la pregunta sobre si, el comportamiento mejorará con un valor intermedio o incluso con algún valor un poco por encima o por debajo de ambos, por lo que se prosiguió a utilizar nuevos valores de p entre 0.88 y 1.02, y se pudo observar que se mantenía la precisión a lo largo de todo el intervalo.

Como último paso se hizo una combinación lineal de las distancias elegidas, teniendo como base la distancia Manhattan y combinándola con el valor de $p = 0.9$, variando por medio de un parámetro α la proporción de cada una de ellas en la distancia híbrida, se observó que se mantiene el porcentaje de precisión desde utilizar solamente $p = 0.9$ hasta $p = 1$, así como en todas sus combinaciones con α (desde 0 hasta 1).

Finalmente se reafirma, que entre las métricas más utilizadas, la distancia Manhattan obtiene los mejores resultados con ambos clasificadores en la mayoría de las bases de datos, en cuanto a las demás pruebas se identifica que $p = 0.9$ obtiene valores tan buenos como los de Manhattan. Tal como se esperaba al utilizar valores entre estos dos, se mantenía la precisión, así como sucedió con la distancia híbrida, siendo interesante que la combinación de una métrica con una semimétrica conserva el porcentaje de precisión.

Considerando lo anterior, la mejor opción para utilizar en un clasificador es la distancia Manhattan, ya que es bien conocida, así como sencilla de implementar, además de que se garantiza que es una métrica puesto que cumple con los axiomas, además gracias a los experimentos y pruebas realizados se sabe que obtiene excelentes resultados.

Apéndice A

A.1. Diagramas de caja con valores de $p > 1$.

Se muestran los diagramas de caja con la precisión obtenida en cada base de datos, con ambos clasificadores al utilizar valores de $p > 1$, es decir, con $p = 1, 2, 3$ y 4.

A.1.1. Diagramas de caja de Iris con valores de $p > 1$.

Figura 4: Precisión de CGS en Iris con $p > 1$.

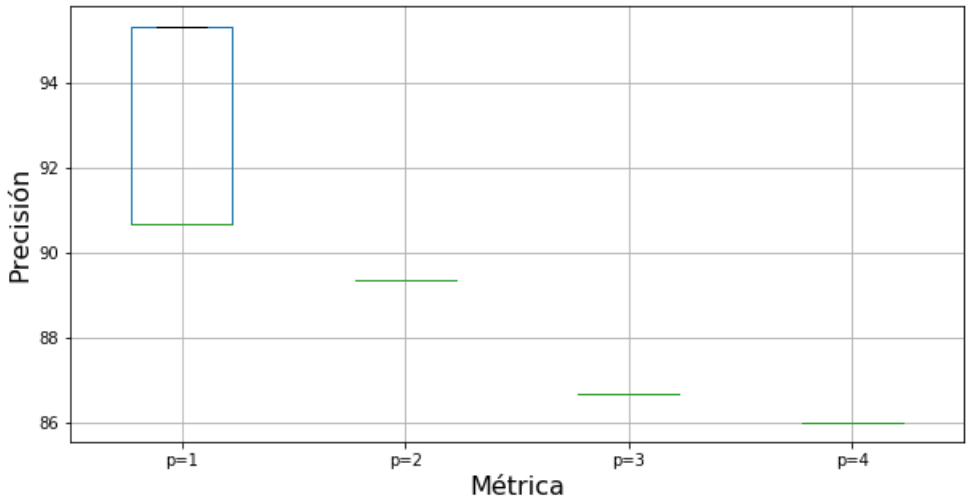
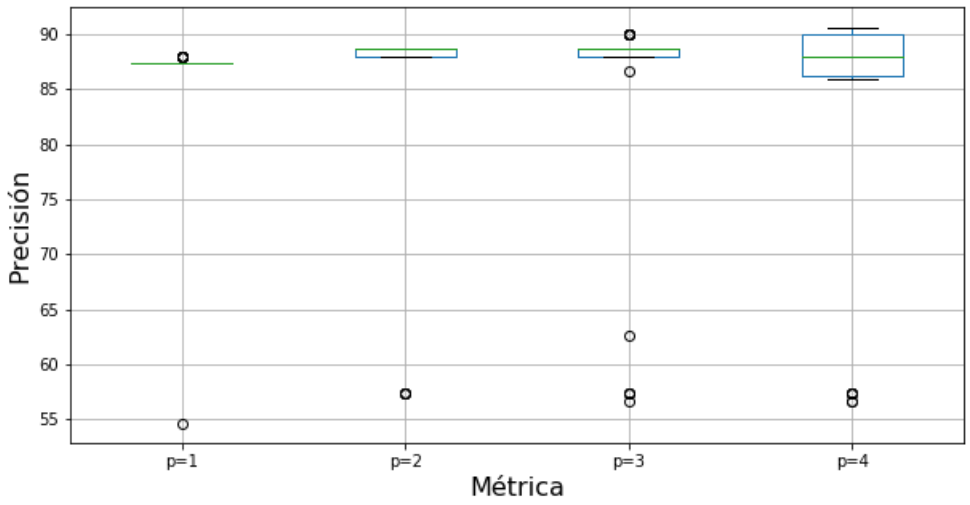


Figura 5: Precisión de K-medias en Iris con $p > 1$.



A.1.2. Diagramas de caja de Wine con valores de $p > 1$.

Figura 6: Precisión de CGS en Wine con $p > 1$.

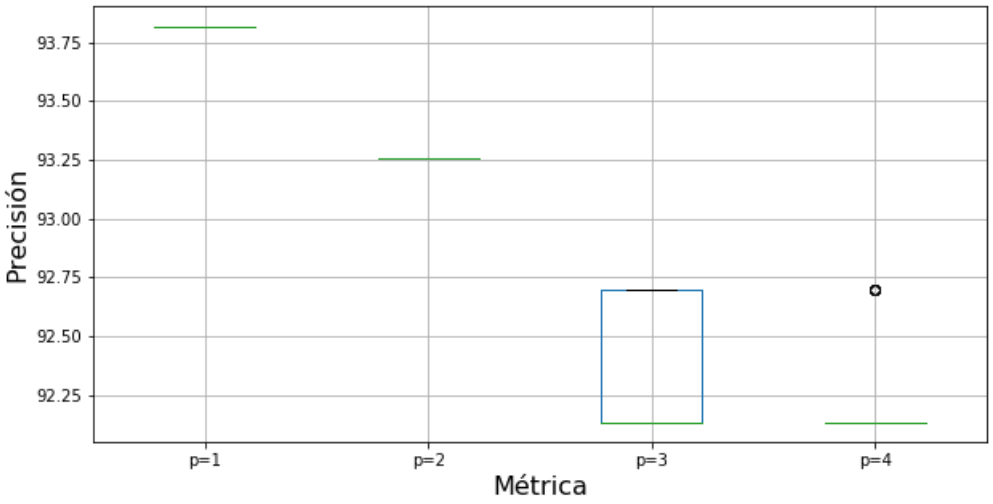
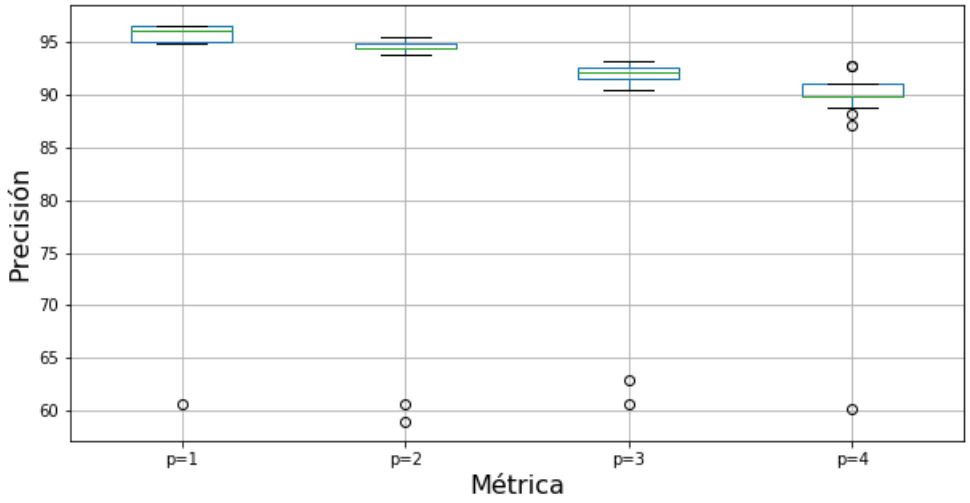


Figura 7: Precisión de K-medias en Wine con $p > 1$.



A.1.3. Diagramas de caja de Zoo con valores de $p > 1$.

Figura 8: Precisión de CGS en Zoo con $p > 1$.

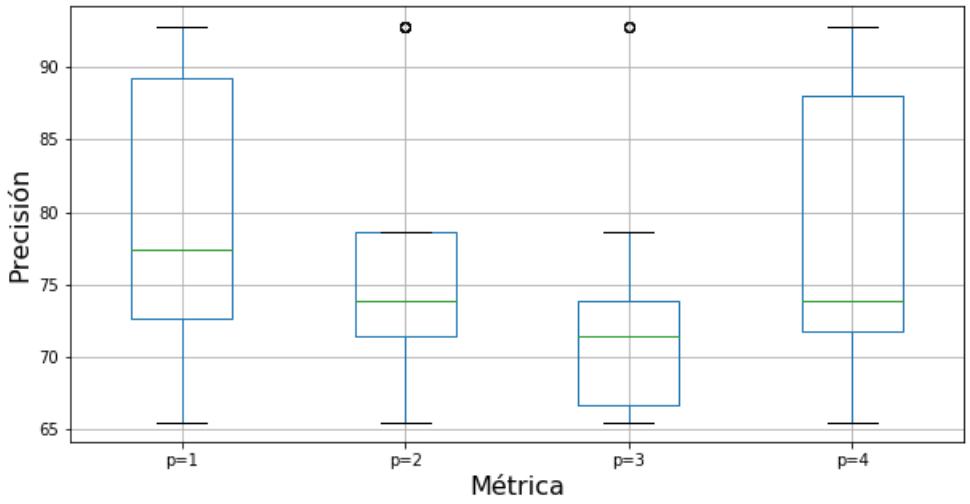
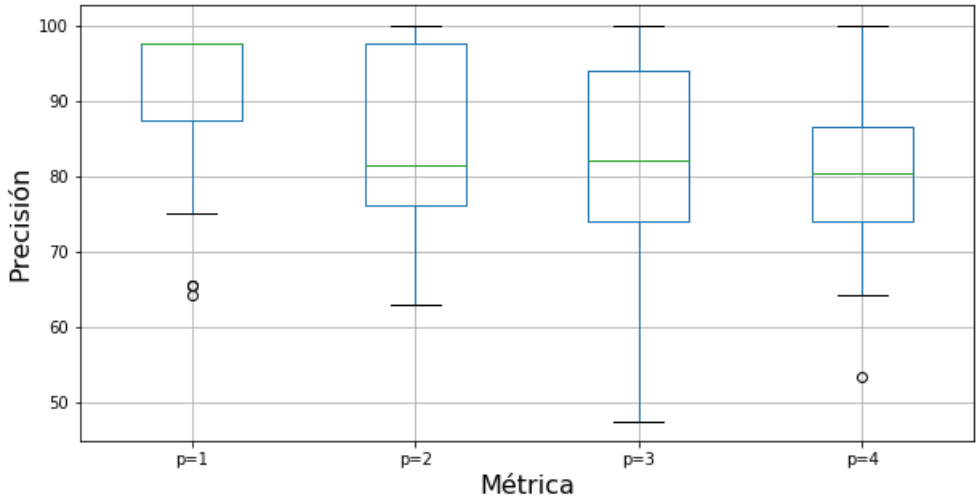


Figura 9: Precisión de K-medias en Zoo con $p > 1$.



A.1.4. Diagramas de caja de Stone Flakes con valores de $p > 1$.

Figura 10: Precisión de CGS en Stone Flakes con $p > 1$.

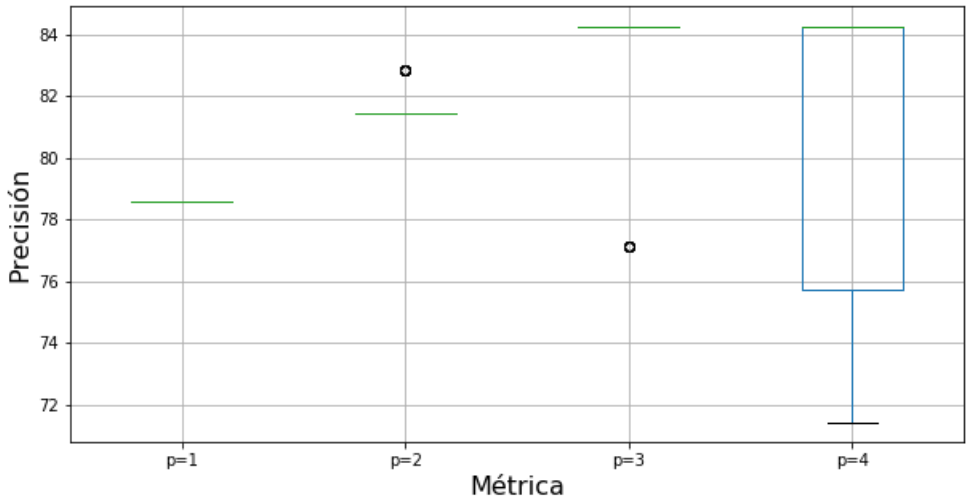
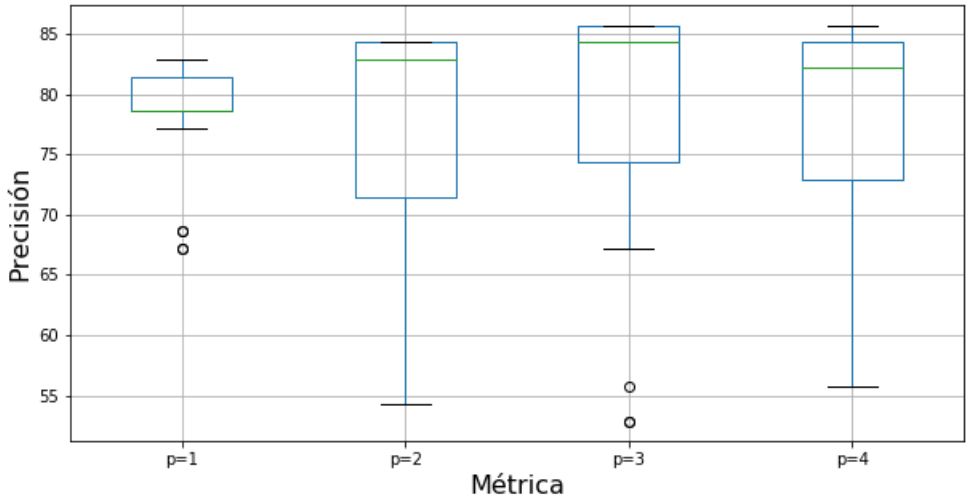


Figura 11: Precisión de K-medias en Stone Flakes con $p > 1$.



A.1.5. Diagramas de caja de Titanic con valores de $p > 1$.

Figura 12: Precisión de CGS en Titanic con $p > 1$.

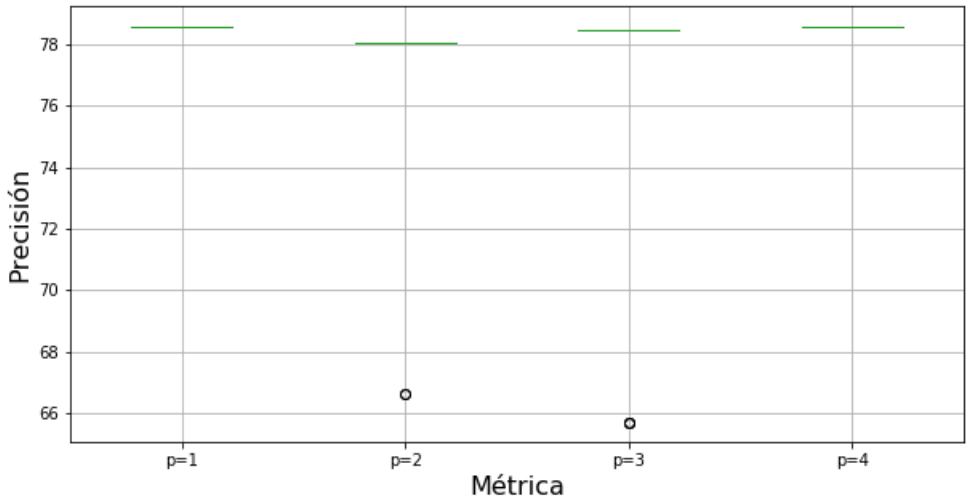
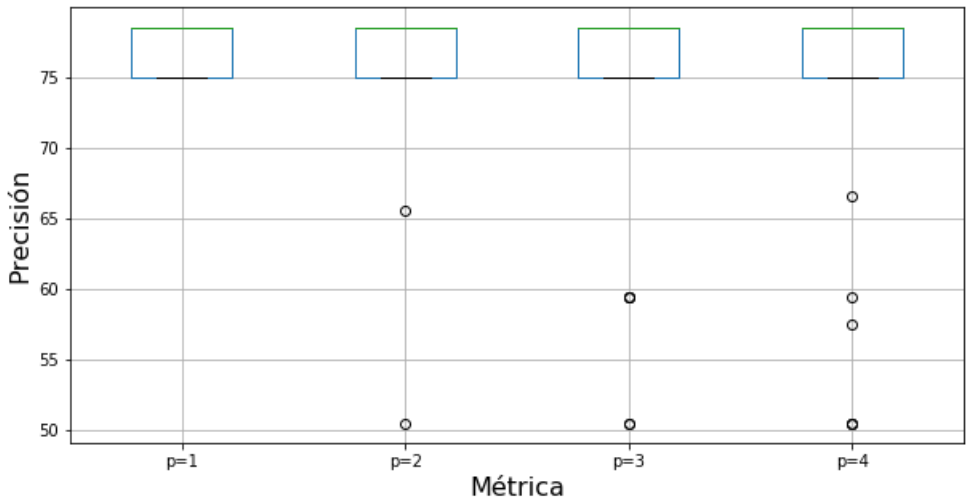


Figura 13: Precisión de K-medias en Titanic con $p > 1$.



A.1.6. Diagramas de caja de Bach con valores de $p > 1$.

Figura 14: Precisión de CGS en Bach con $p > 1$.

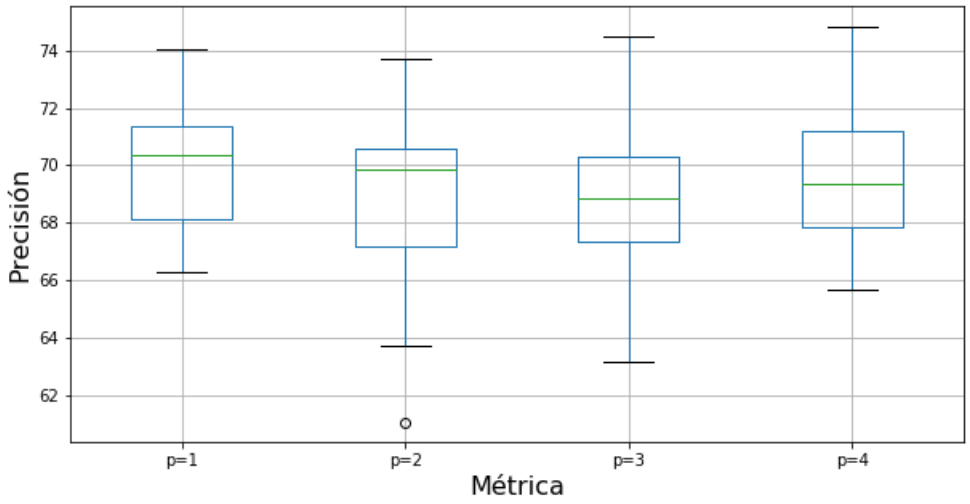
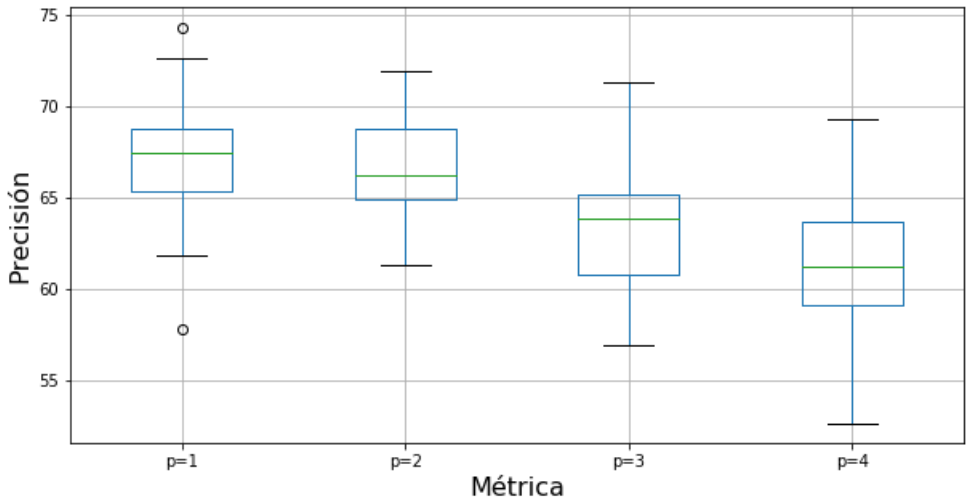


Figura 15: Precisión de K-medias en Bach con $p > 1$.



A.1.7. Diagramas de caja de Celeb con valores de $p > 1$.

Figura 16: Precisión de CGS en Celeb con $p > 1$.

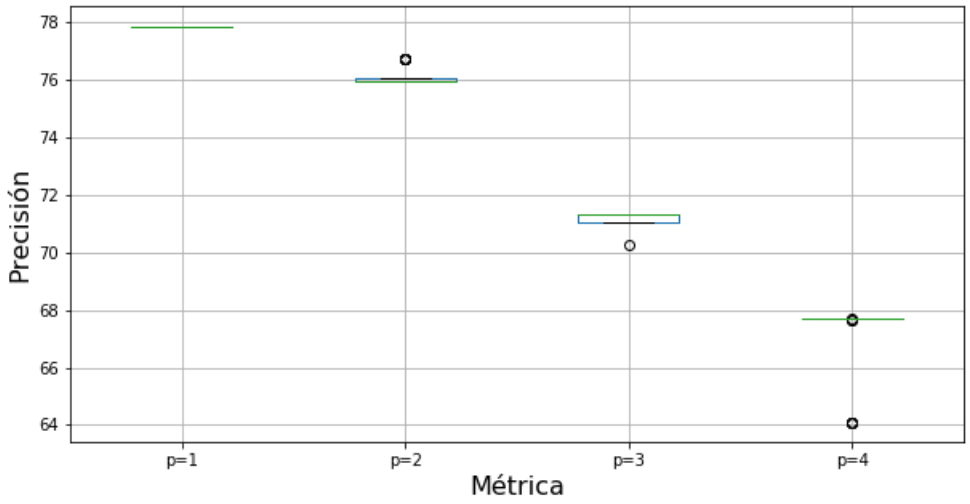
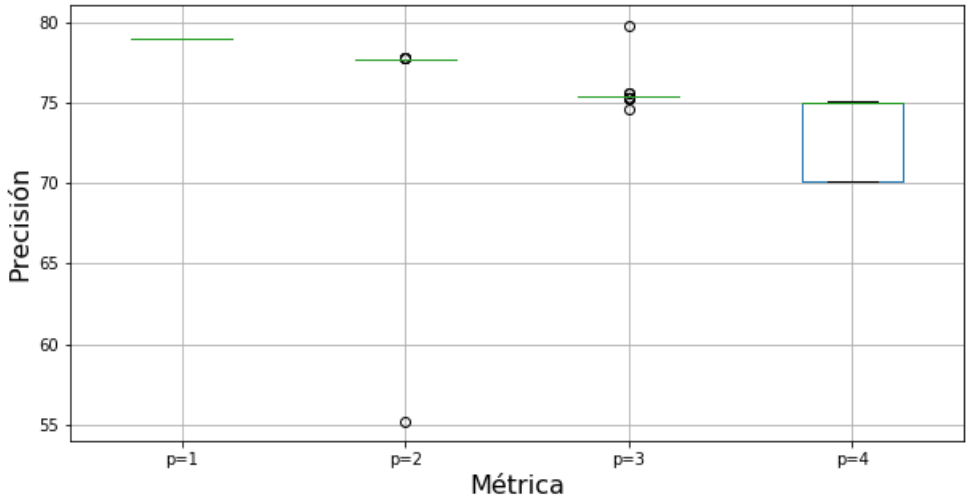


Figura 17: Precisión de K-medias en Celeb con $p > 1$.



A.2. Diagramas de caja con valores $0 < p < 1$.

Se muestran los diagramas de caja con la precisión obtenida en cada base de datos, con ambos clasificadores al utilizar valores $0 < p < 1$, es decir, con p = 0.1, 0.3, 0.5, 0.7 y 0.9.

A.2.1. Diagramas de caja de Iris con valores $0 < p < 1$.

Figura 18: Precisión de CGS en Iris con $0 < p < 1$.

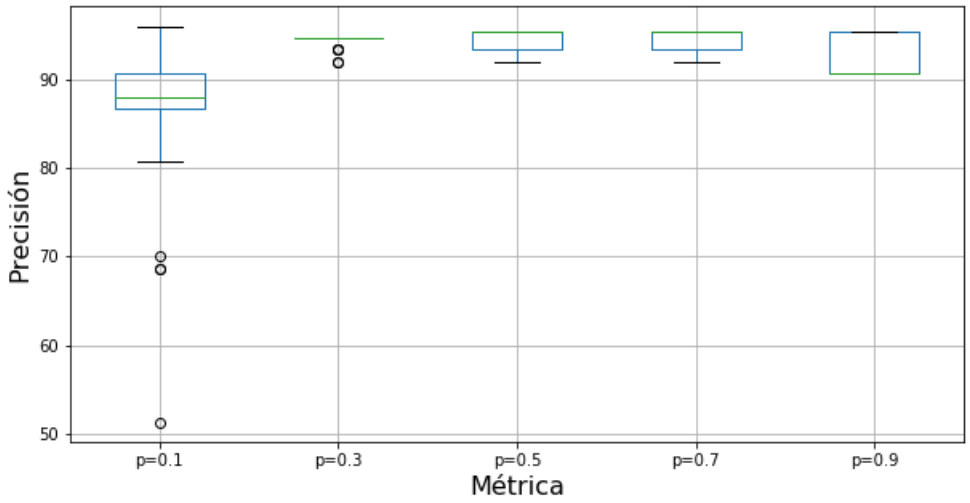
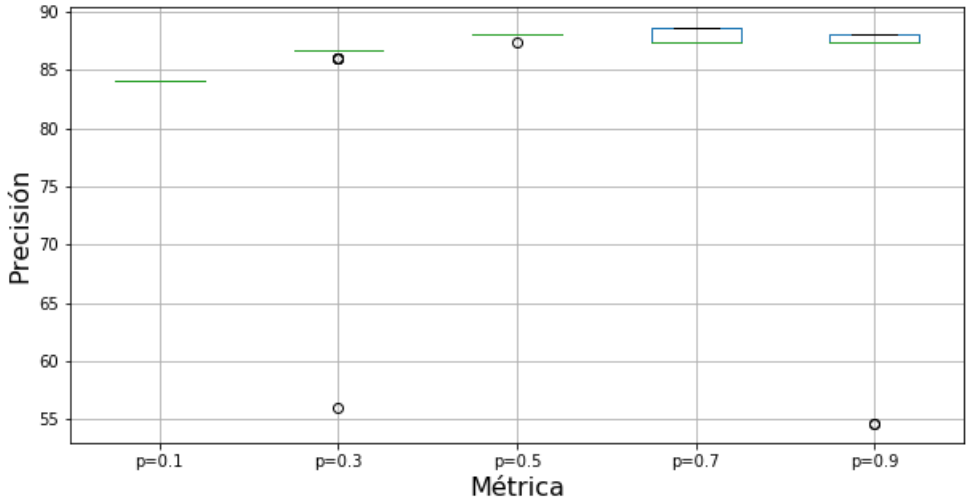


Figura 19: Precisión de K-medias en Iris con $0 < p < 1$.



A.2.2. Diagramas de caja de Wine con valores $0 < p < 1$.

Figura 20: Precisión de CGS en Wine con $0 < p < 1$.

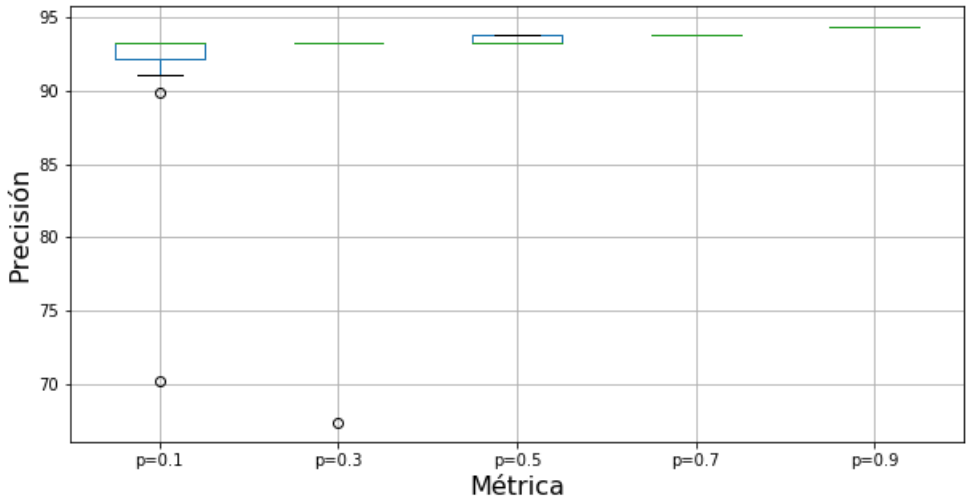
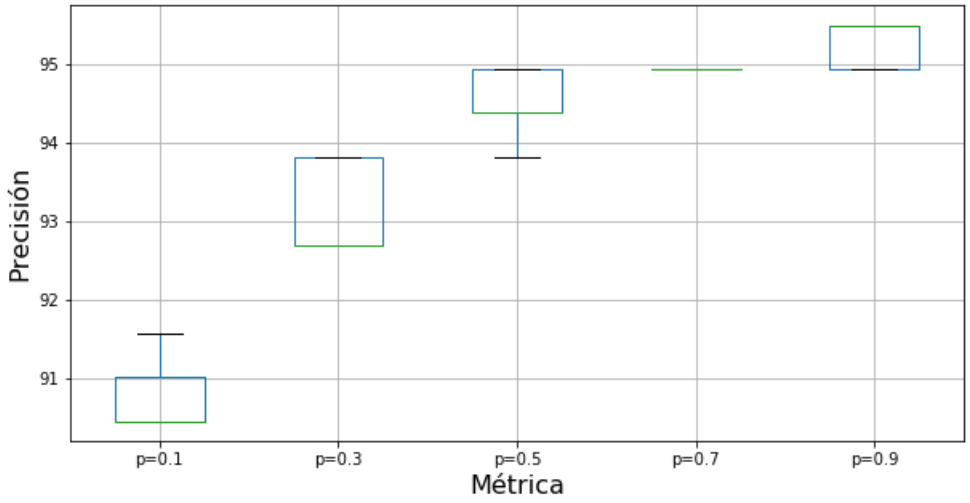


Figura 21: Precisión de K-medias en Wine con $0 < p < 1$.



A.2.3. Diagramas de caja de Zoo con valores $0 < p < 1$.

Figura 22: Precisión de CGS en Zoo con $0 < p < 1$.

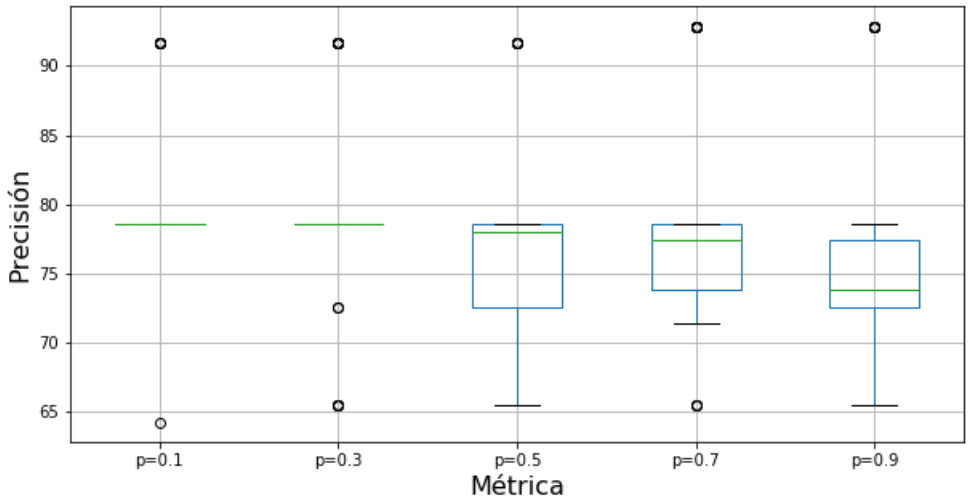
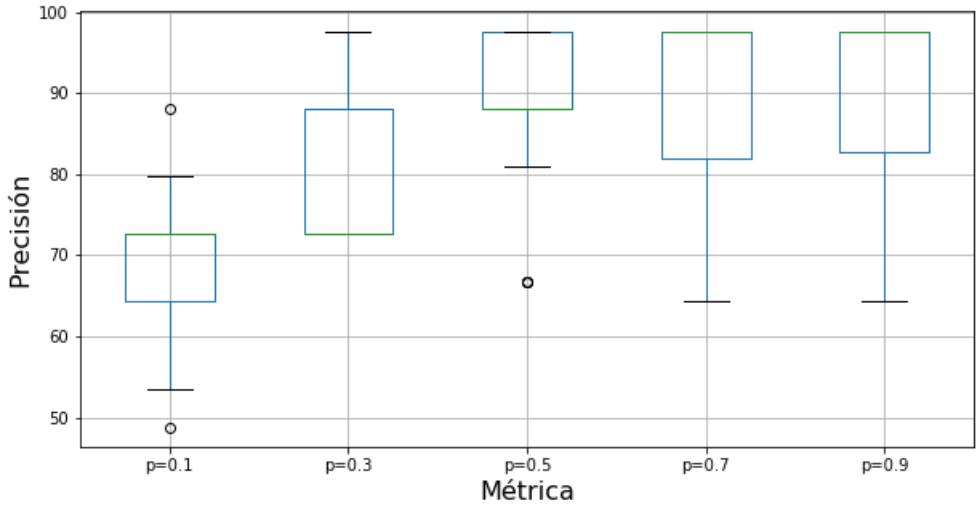


Figura 23: Precisión de K-medias en Zoo con $0 < p < 1$.



A.2.4. Diagramas de caja de Stone Flakes con valores $0 < p < 1$.

Figura 24: Precisión de CGS en Stone Flakes con $0 < p < 1$.

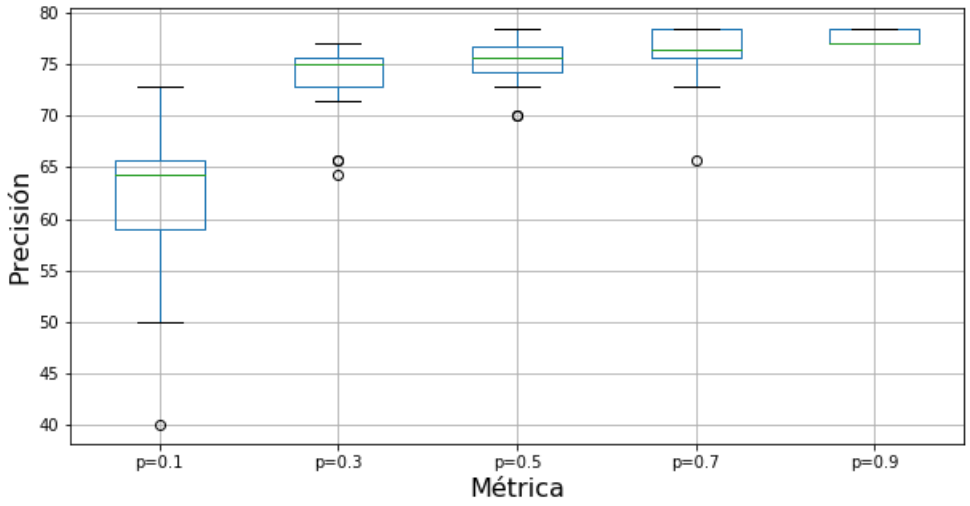
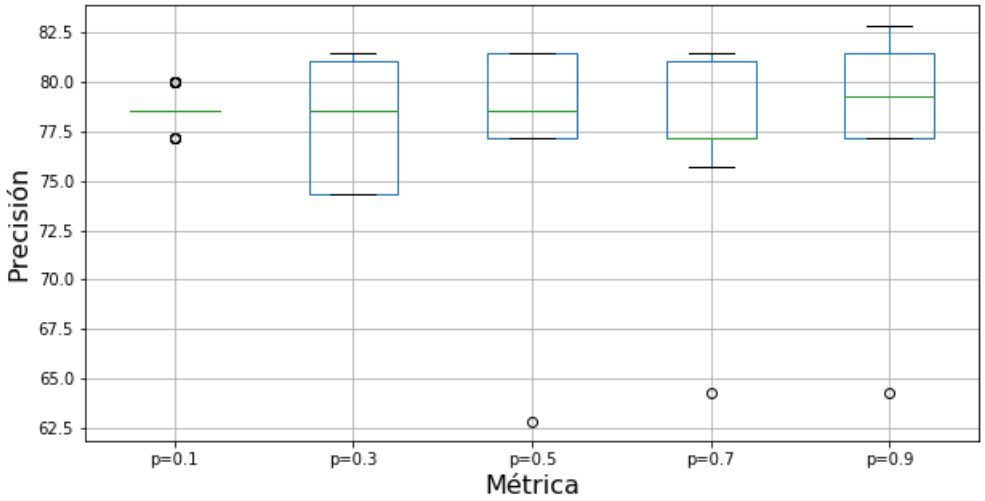


Figura 25: Precisión de K-medias en Stone Flakes con $0 < p < 1$.



A.2.5. Diagramas de caja de Titanic con valores $0 < p < 1$.

Figura 26: Precisión de CGS en Titanic con $0 < p < 1$.

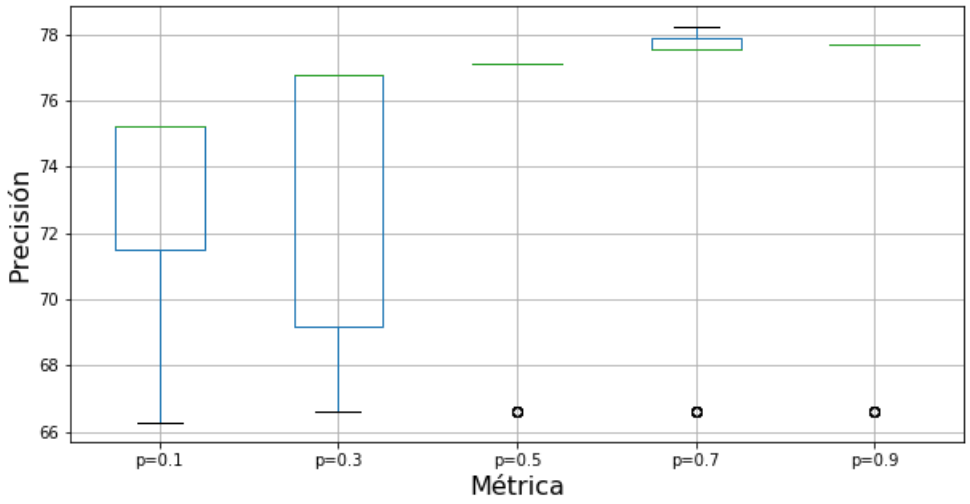
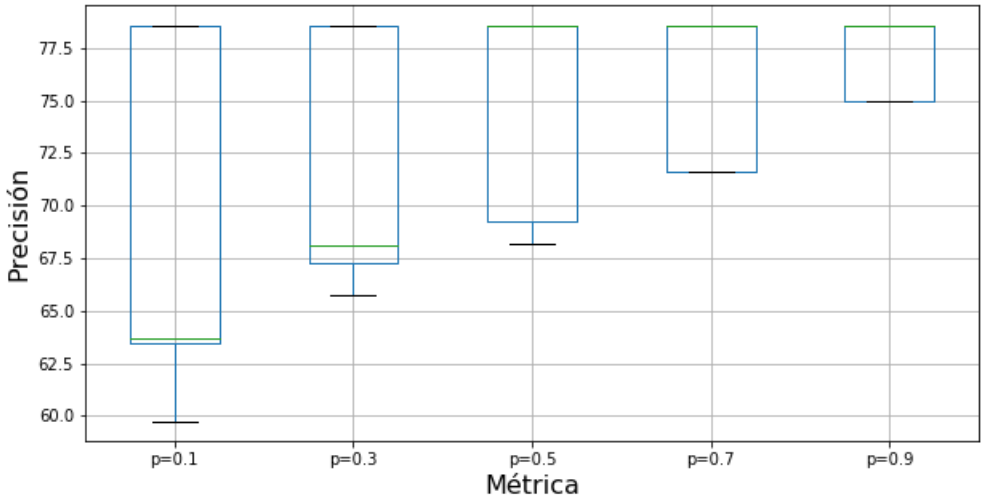


Figura 27: Precisión de K-medias en Titanic con $0 < p < 1$.



A.2.6. Diagramas de caja de Bach con valores $0 < p < 1$.

Figura 28: Precisión de CGS en Bach con $0 < p < 1$.

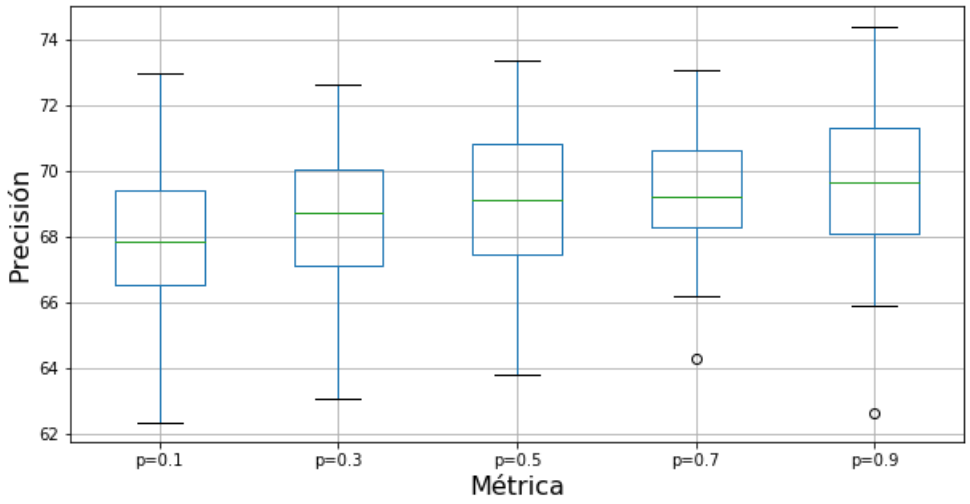
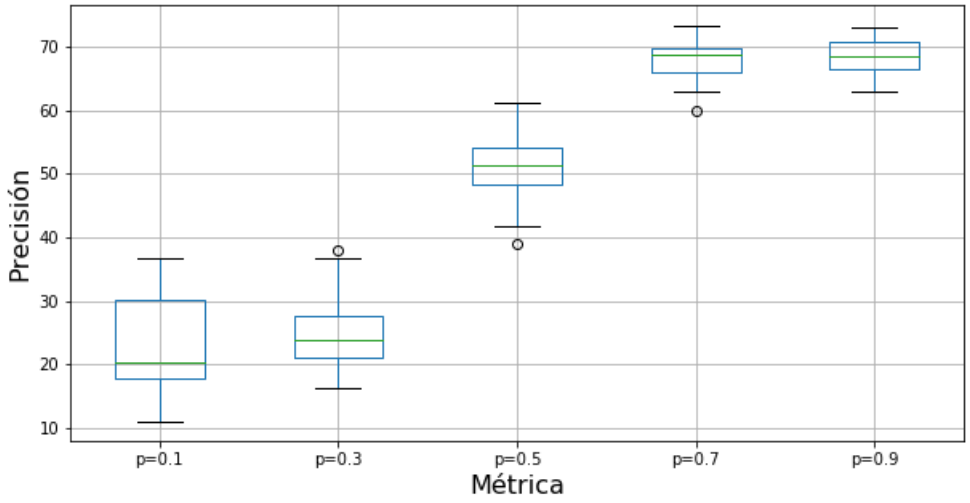


Figura 29: Precisión de K-medias en Bach con $0 < p < 1$.



A.2.7. Diagramas de caja de Celeb con valores $0 < p < 1$.

Figura 30: Precisión de CGS en Celeb con $0 < p < 1$.

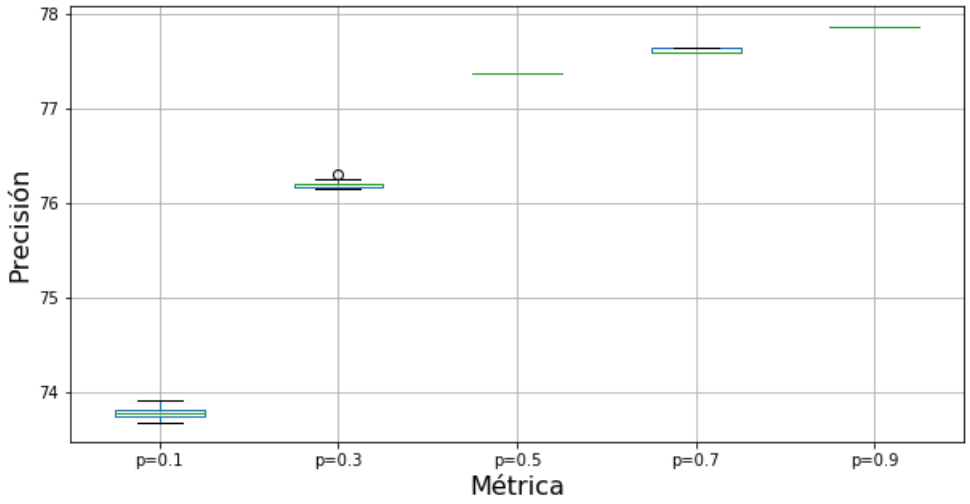
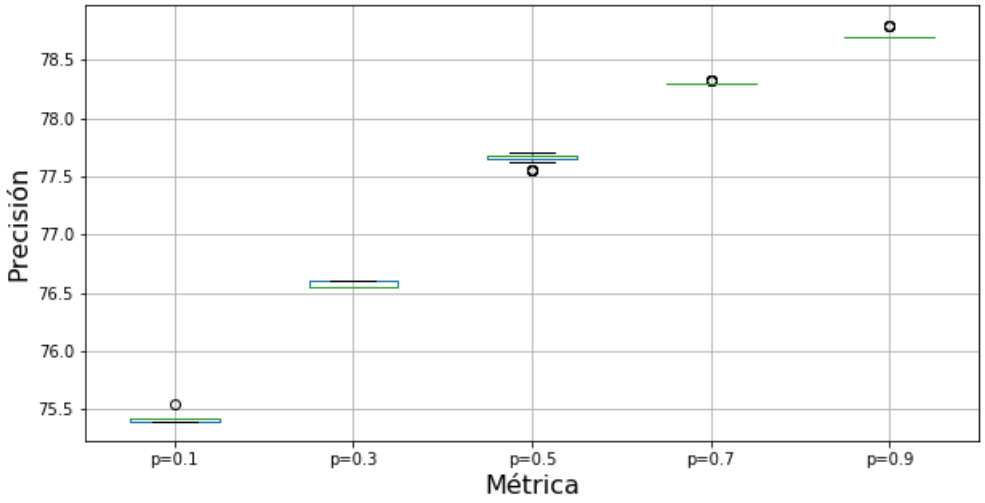


Figura 31: Precisión de K-medias en Celeb con $0 < p < 1$.



A.3. Diagramas de caja con valores $0.88 \leq p \leq 1.02$.

Se muestran los diagramas de caja con la precisión obtenida en cada base de datos, con ambos clasificadores al utilizar valores $0.88 \leq p \leq 1.02$.

A.3.1. Diagramas de caja de Iris con valores $0.88 \leq p \leq 1.02$.

Figura 32: Precisión de CGS en Iris con $0.88 \leq p \leq 1.02$.

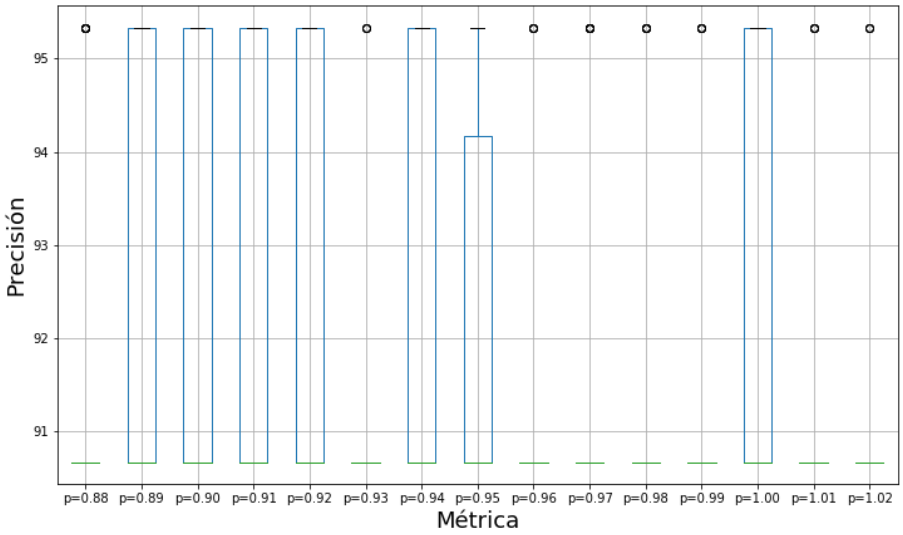
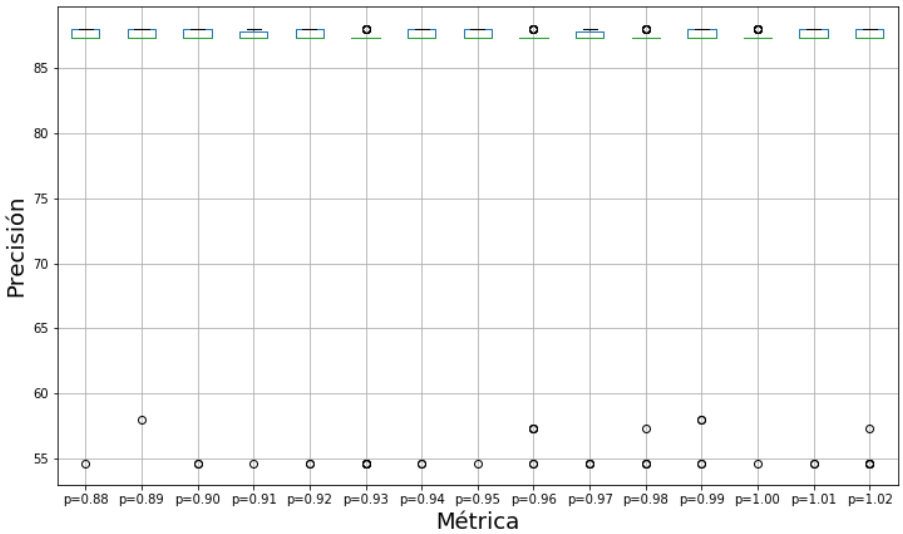


Figura 33: Precisión de K-medias en Iris con $0.88 \leq p \leq 1.02$.



A.3.2. Diagramas de caja de Wine con valores $0.88 \leq p \leq 1.02$.

Figura 34: Precisión de CGS en Wine con $0.88 \leq p \leq 1.02$.

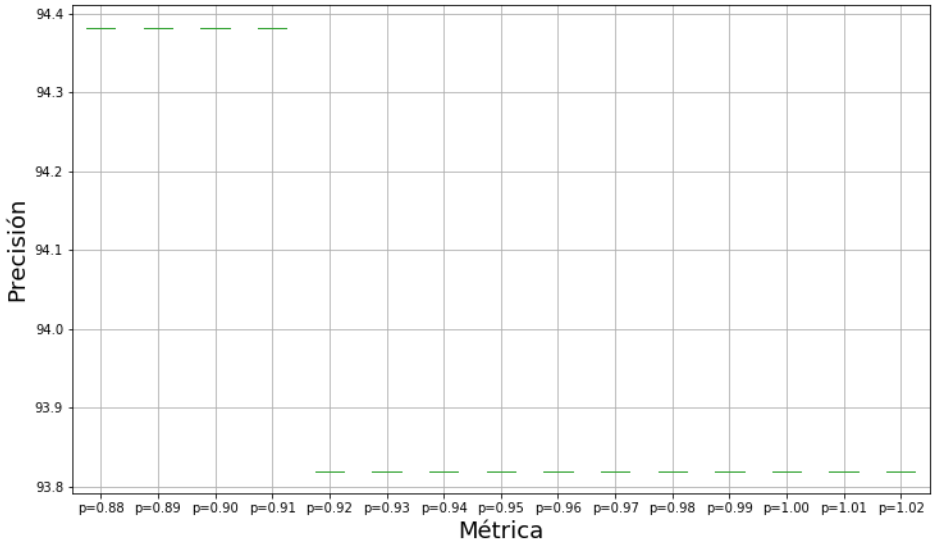
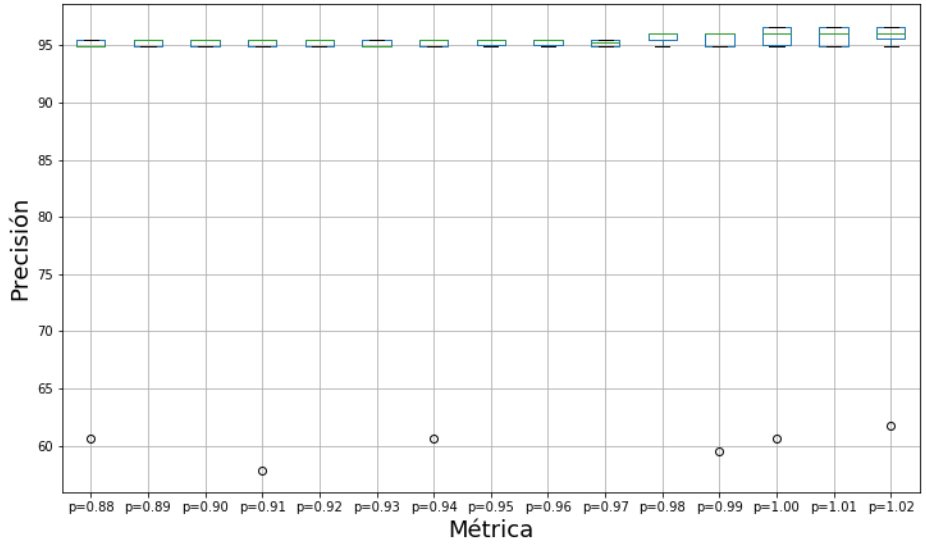


Figura 35: Precisión de K-medias en Wine con $0.88 \leq p \leq 1.02$.



A.3.3. Diagramas de caja de Zoo con valores $0.88 \leq p \leq 1.02$.

Figura 36: Precisión de CGS en Zoo con $0.88 \leq p \leq 1.02$.

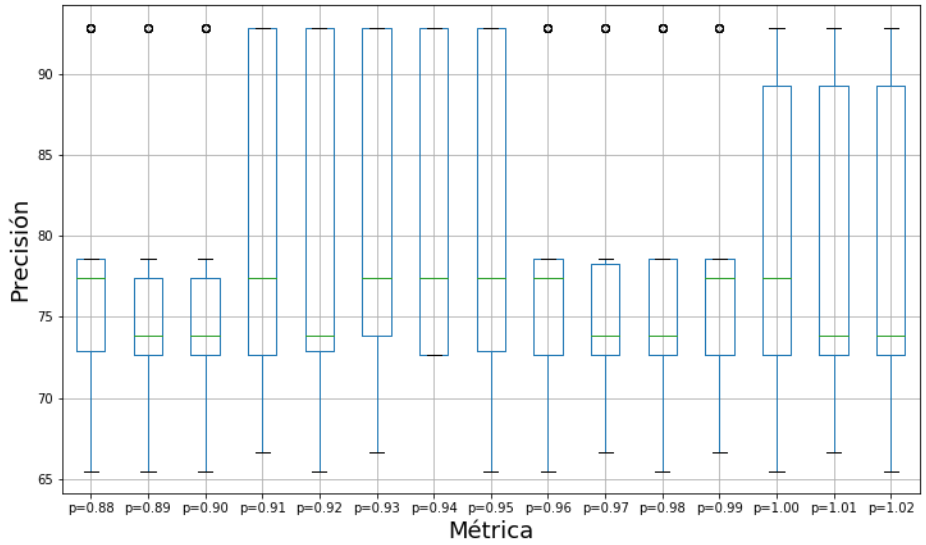
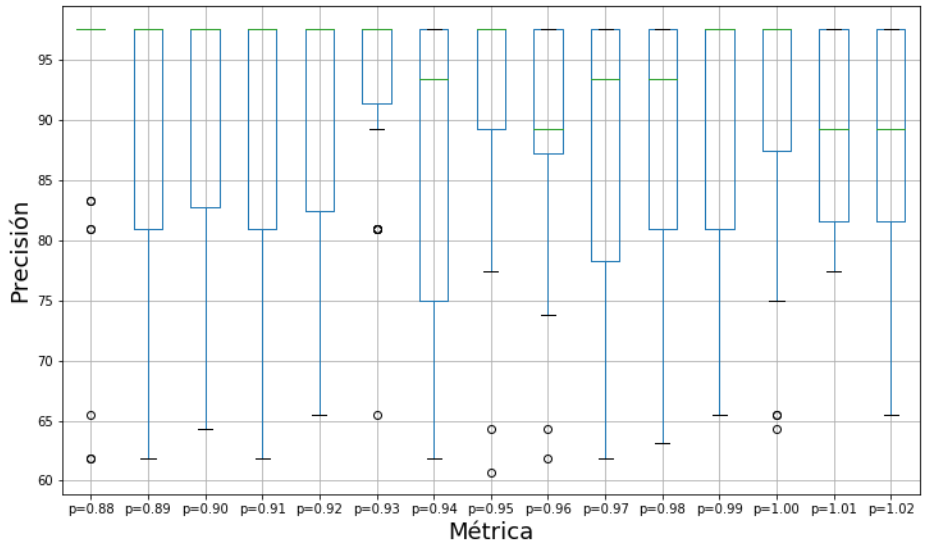


Figura 37: Precisión de K-medias en Zoo con $0.88 \leq p \leq 1.02$.



A.3.4. Diagramas de caja de Stone Flakes con valores $0.88 \leq p \leq 1.02$.

Figura 38: Precisión de CGS en Stone Flakes con $0.88 \leq p \leq 1.02$.

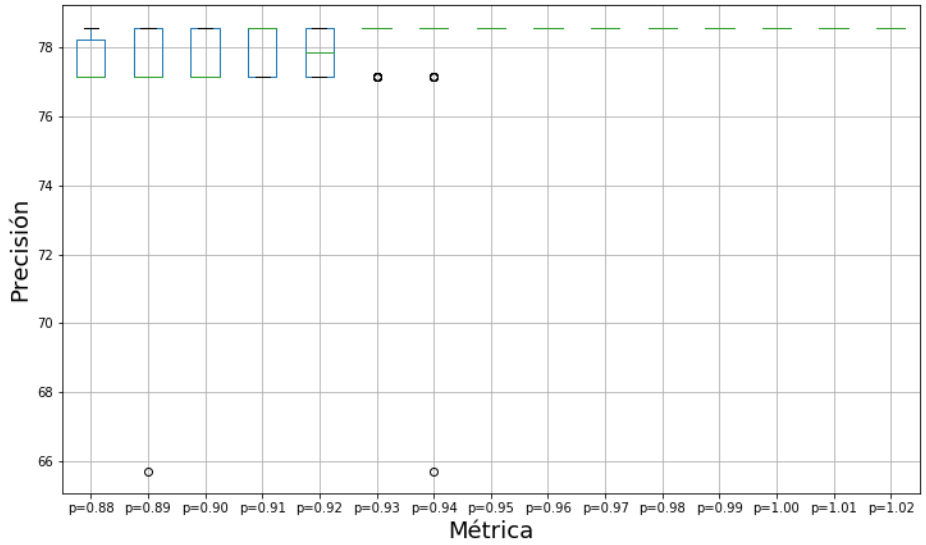
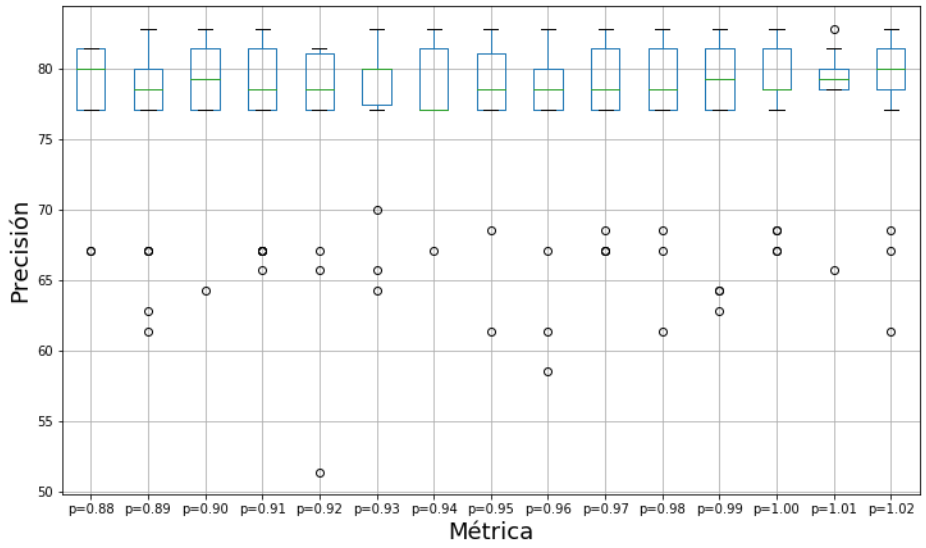


Figura 39: Precisión de K-medias en Stone Flakes con $0.88 \leq p \leq 1.02$.



A.3.5. Diagramas de caja de Titanic con valores $0.88 \leq p \leq 1.02$.

Figura 40: Precisión de CGS en Titanic con $0.88 \leq p \leq 1.02$.

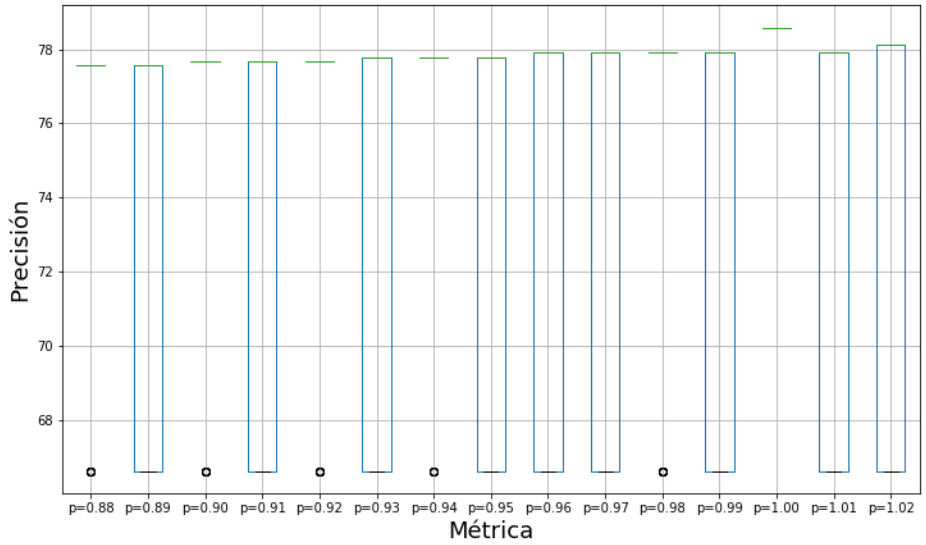
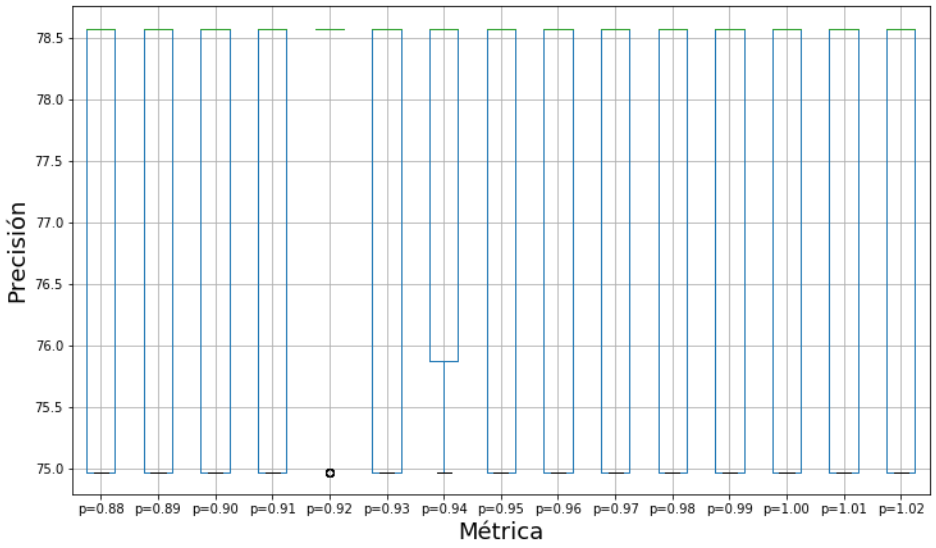


Figura 41: Precisión de K-medias en Titanic con $0.88 \leq p \leq 1.02$.



A.3.6. Diagramas de caja de Bach con valores $0.88 \leq p \leq 1.02$.

Figura 42: Precisión de CGS en Bach con $0.88 \leq p \leq 1.02$.

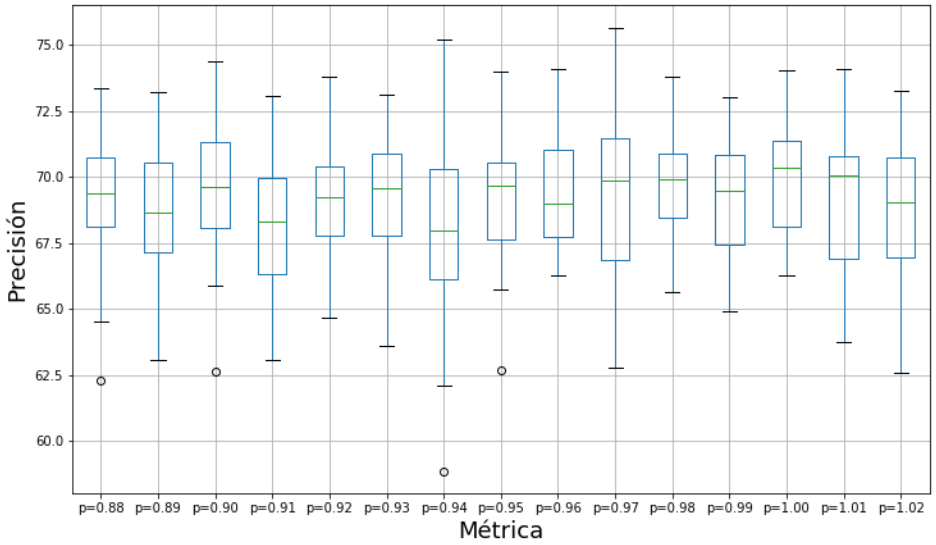
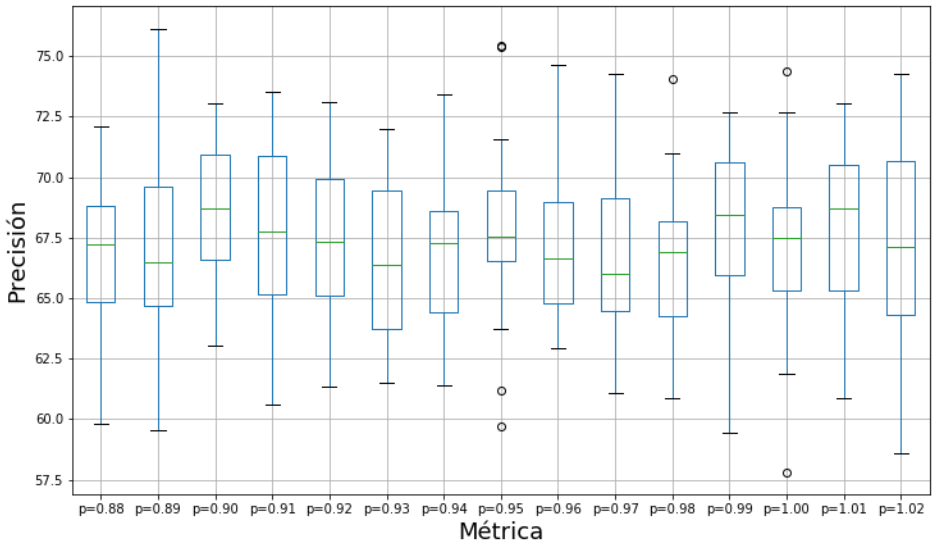


Figura 43: Precisión de K-medias en Bach con $0.88 \leq p \leq 1.02$.



A.3.7. Diagramas de caja de Celeb con valores $0.88 \leq p \leq 1.02$.

Figura 44: Precisión de CGS en Celeb con $0.88 \leq p \leq 1.02$.

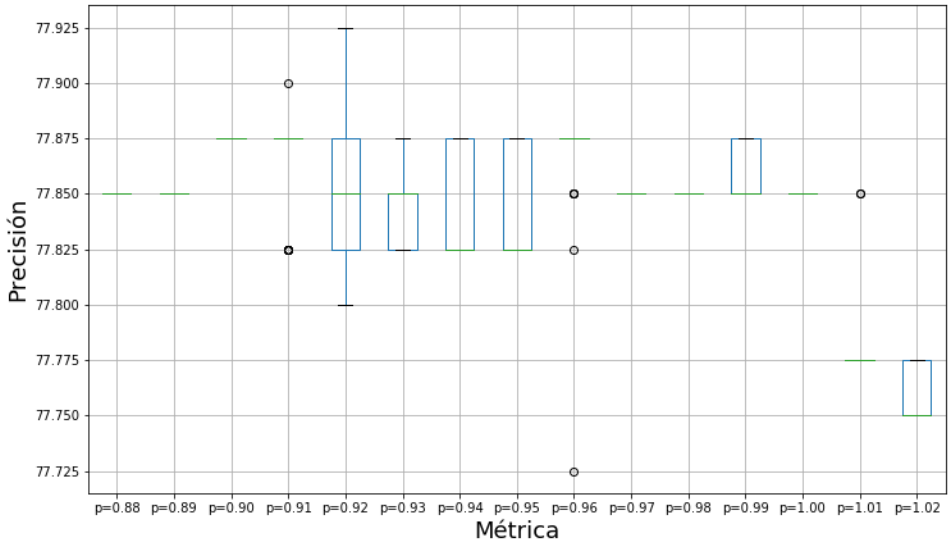
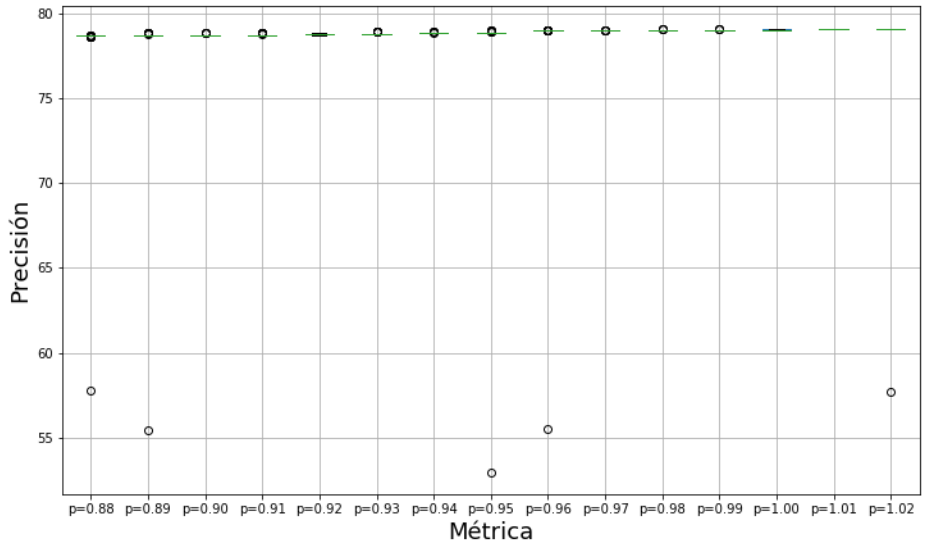


Figura 45: Precisión de K-medias en Celeb con $0.88 \leq p \leq 1.02$.



A.4. Diagramas de caja con distancia híbrida.

Se muestran los diagramas de caja con la precisión obtenida en cada base de datos, con ambos clasificadores al utilizar la distancia híbrida.

A.4.1. Diagramas de caja de Iris con distancia híbrida.

Figura 46: Precisión de CGS en Iris con distancia híbrida.

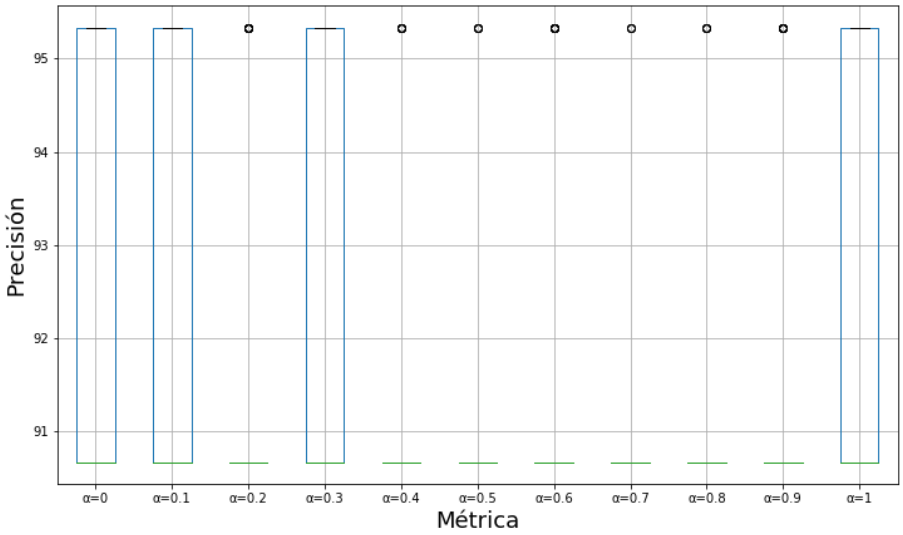
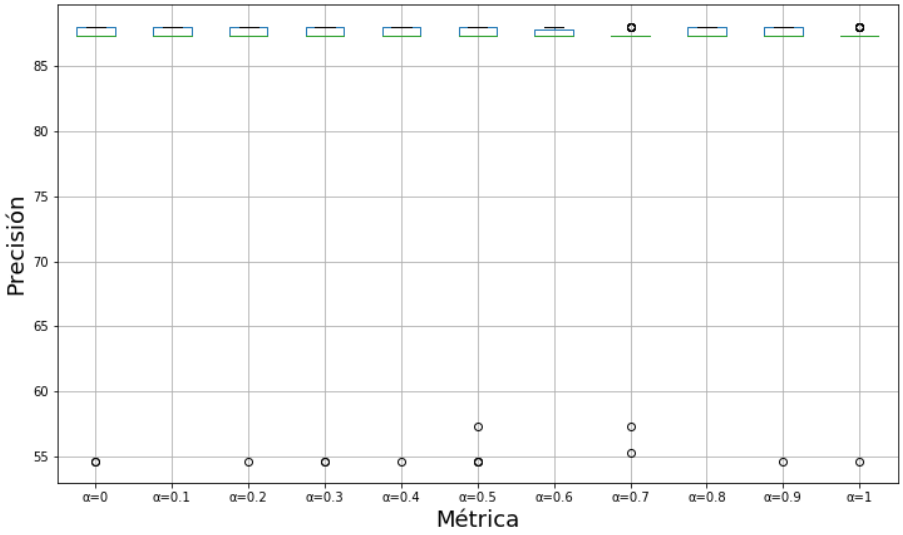


Figura 47: Precisión de K-medias en Iris con distancia híbrida.



A.4.2. Diagramas de caja de Wine con distancia híbrida.

Figura 48: Precisión de CGS en Wine con distancia híbrida.

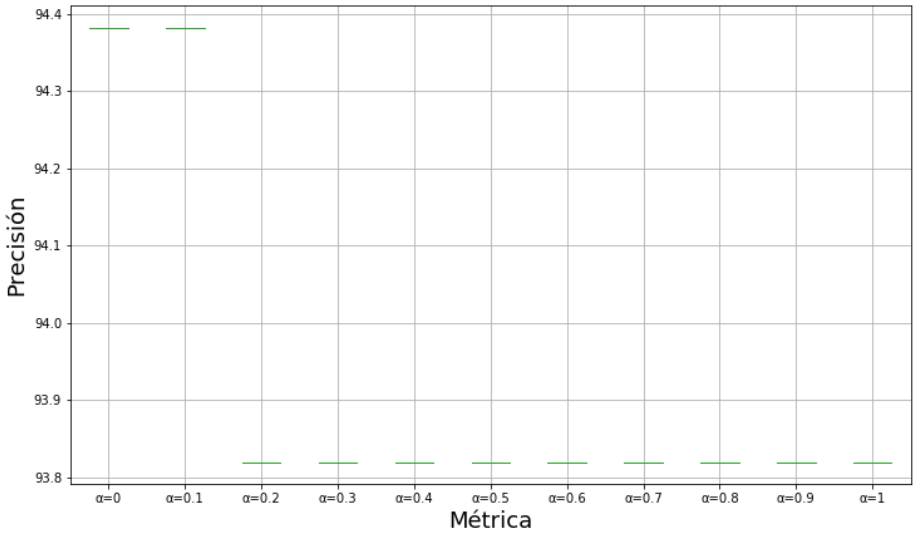
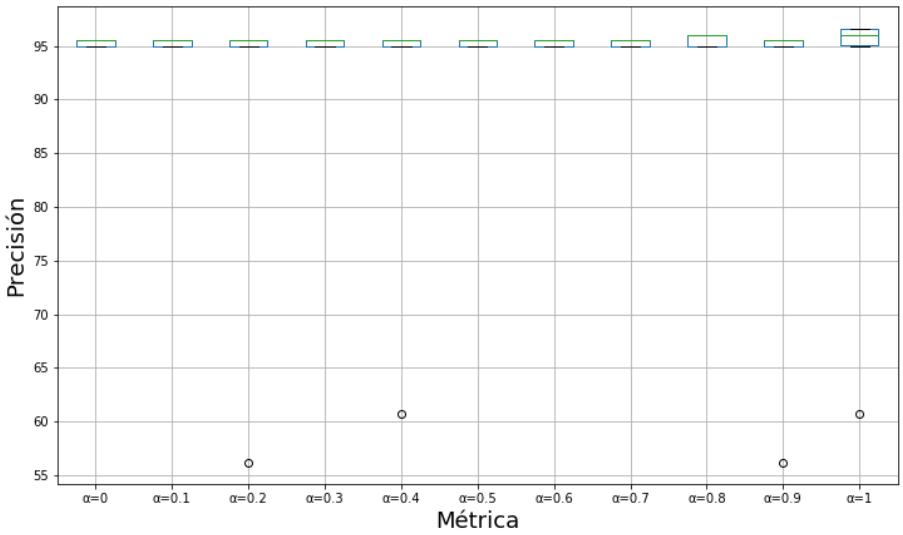


Figura 49: Precisión de K-medias en Wine con distancia híbrida.



A.4.3. Diagramas de caja de Zoo con distancia híbrida.

Figura 50: Precisión de CGS en Zoo con distancia híbrida.

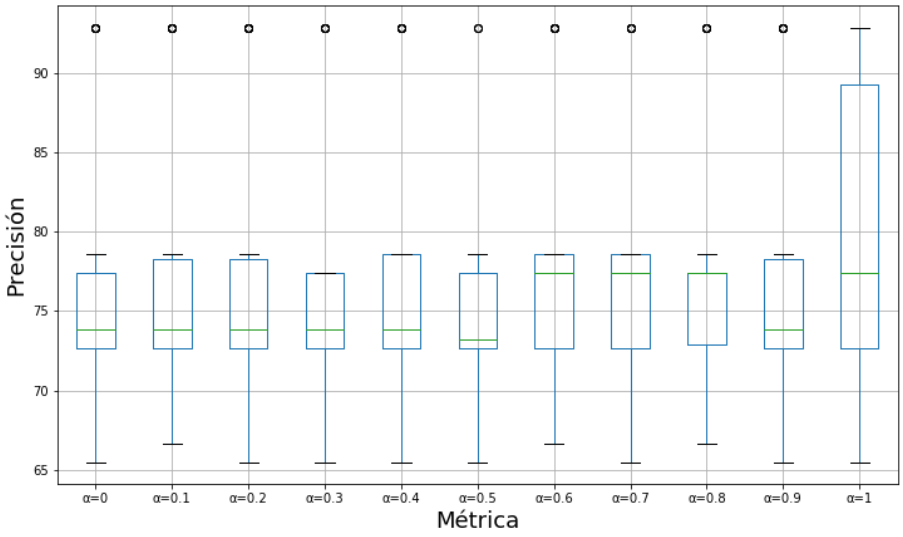
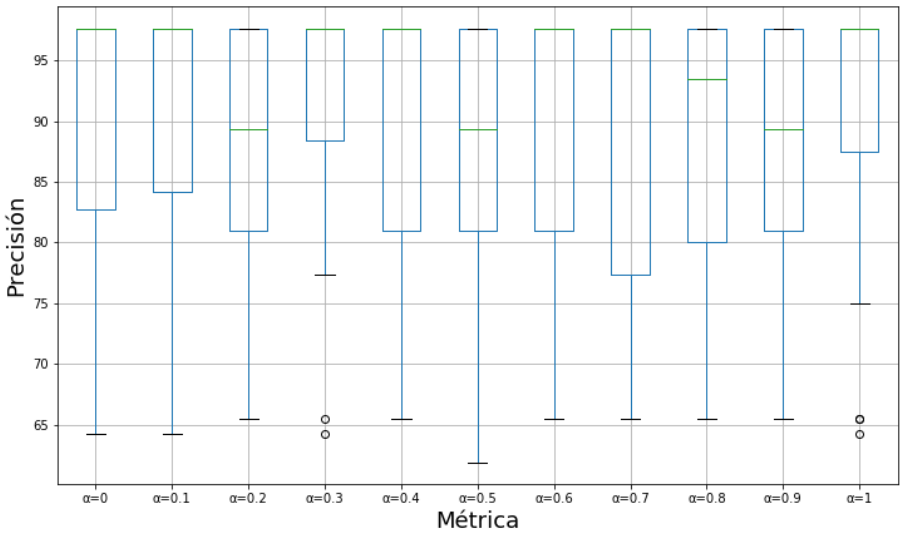


Figura 51: Precisión de K-medias en Zoo con distancia híbrida.



A.4.4. Diagramas de caja de Stone Flakes con distancia híbrida.

Figura 52: Precisión de CGS en Stone Flakes con distancia híbrida.

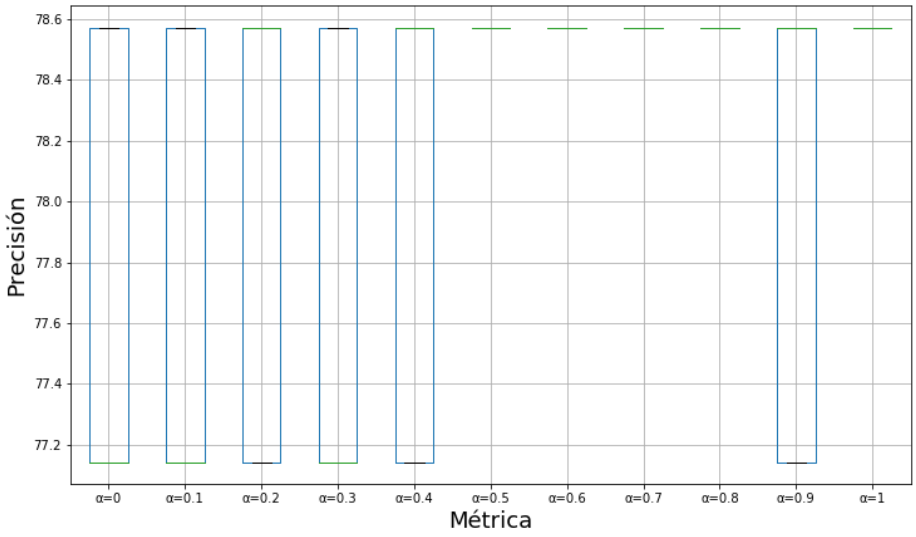
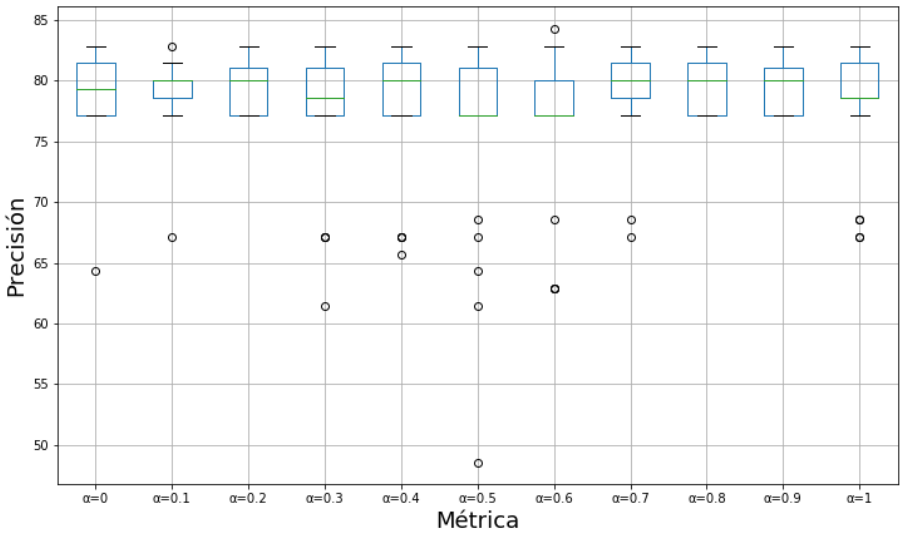


Figura 53: Precisión de K-medias en Stone Flakes con distancia híbrida.



A.4.5. Diagramas de caja de Titanic con distancia híbrida.

Figura 54: Precisión de CGS en Titanic con distancia híbrida.

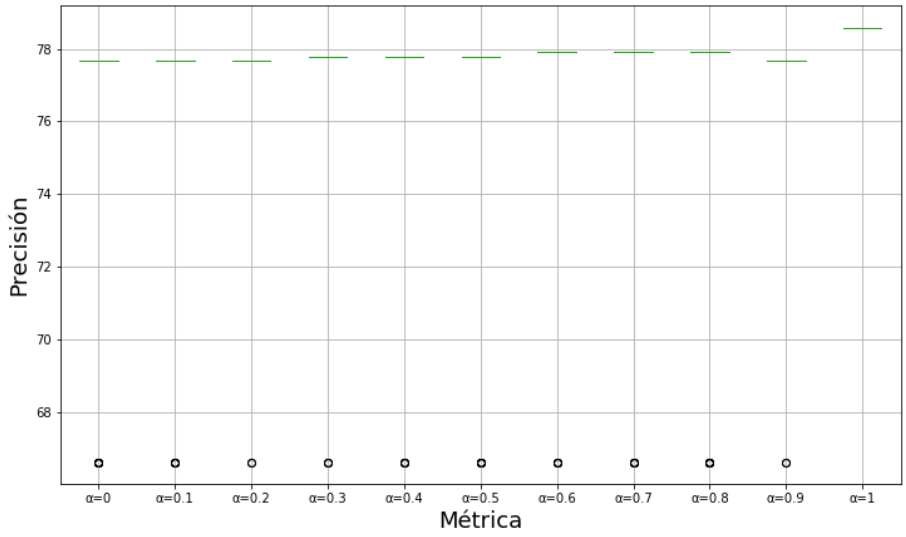
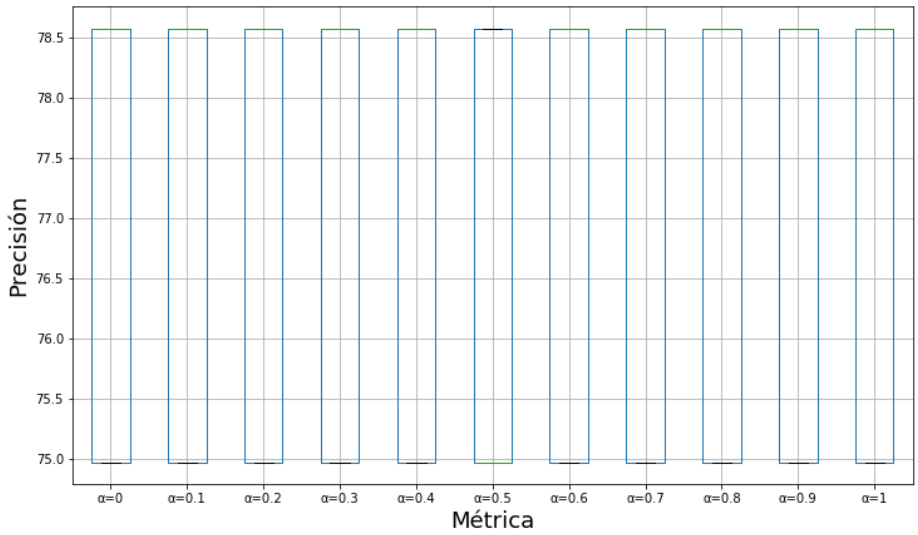


Figura 55: Precisión de K-medias en Titanic con distancia híbrida.



A.4.6. Diagramas de caja de Bach con distancia híbrida.

Figura 56: Precisión de CGS en Bach con distancia híbrida.

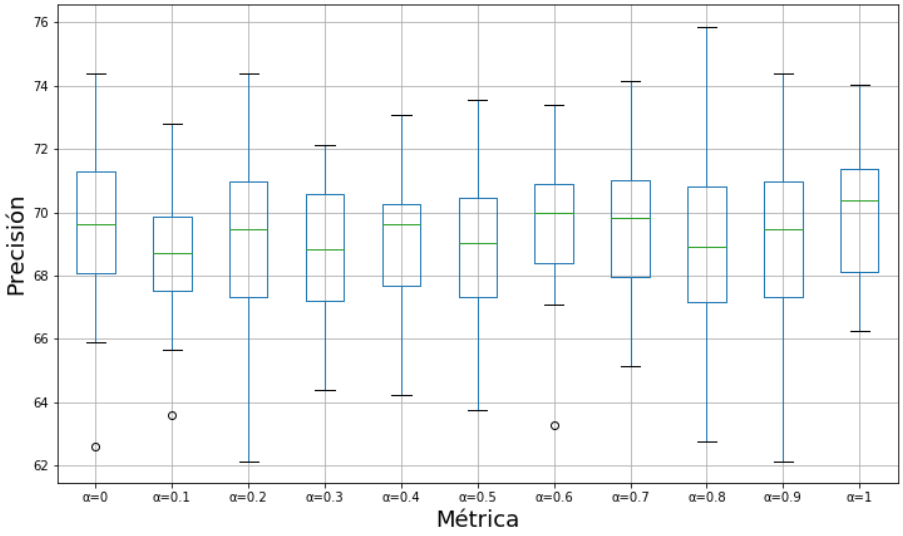
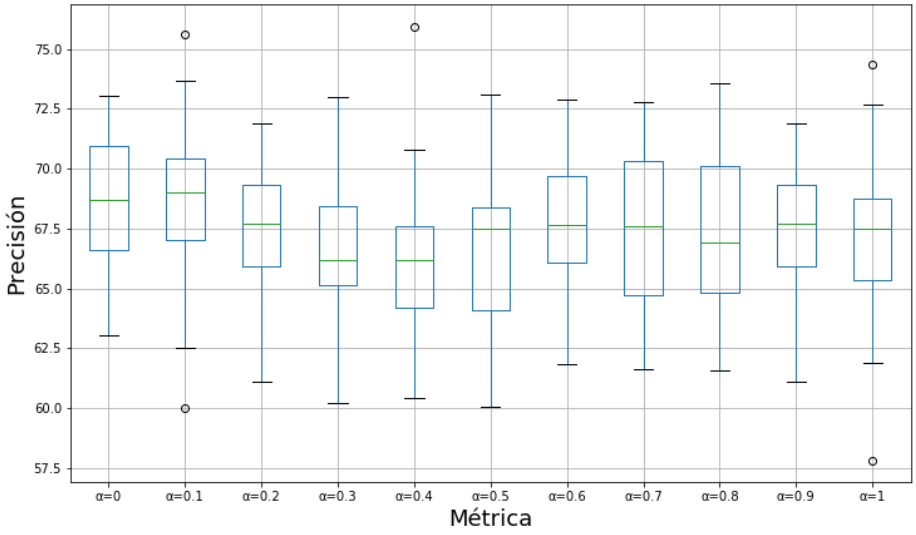


Figura 57: Precisión de K-medias en Bach con distancia híbrida.



A.4.7. Diagramas de caja de Celeb con distancia híbrida.

Figura 58: Precisión de CGS en Celeb con distancia híbrida.

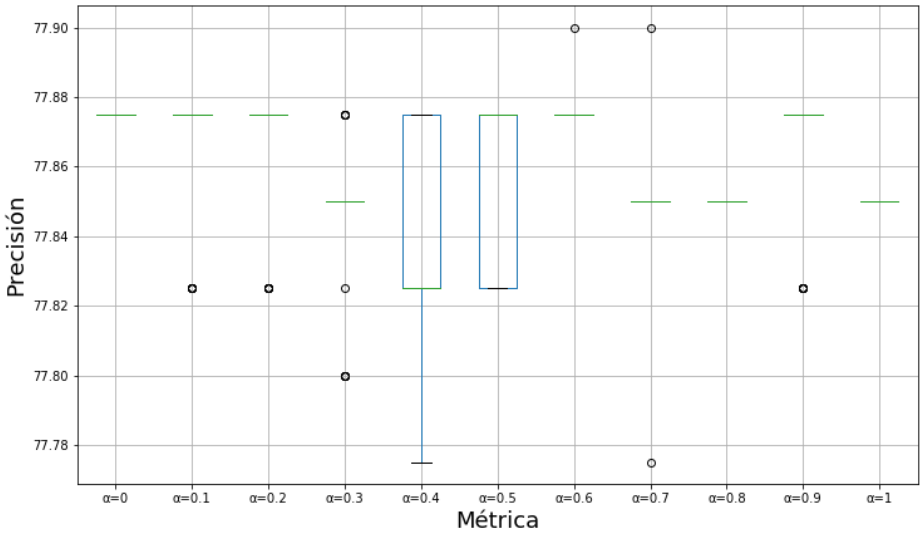
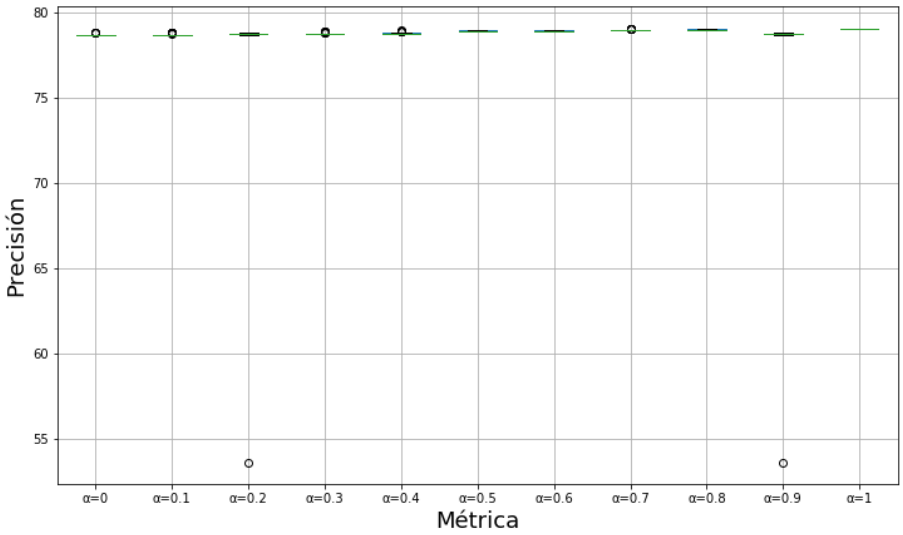


Figura 59: Precisión de K-medias en Celeb con distancia híbrida.



Referencias

[Aha et al., 1987] Aha, D., Murphy, P., Merz, C., Keogh, E., Blake, C., Hettich, S., and Newman, D. (1987). UCI Machine Learning Repository.

[Aparicio Reyes, 2017] Aparicio Reyes, J. L. (2017). Modelo filológico para lenguas francesas y germánicas utilizando coloración de gráficas suaves. Master's thesis, Universidad Autónoma Metropolitana Unidad Iztapalapa, Av. San Rafael Atlixco 186, Leyes de Reforma 1ra Secc, Iztapalapa, 09340 Ciudad de México, CDMX.

[Blanca et al., 2017] Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., and Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29(5):552–557.

[Bradley, 1978] Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2):144–152.

[Branciari, 2000] Branciari, A. (2000). A fixed point theorem of Banach-Caccioppoli type on a class of generalized metric spaces. *Publicationes Mathematicae*, 57.

[Dalatu et al., 2017] Dalatu, P. I., Fitrianto, A., and Mustapha, A. (2017). Hybrid distance functions for K-Means clustering algorithms. *Statistical Journal of the IAOS*, 33(4):989–996.

[Euler, 1741] Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Novi Commentarii Academiae Scientiarum Imperalis Petropolitanae*, 8:128–140.

[Fisher, 1918] Fisher, R. A. (1918). The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433.

[Flores-Cruz et al., 2017] Flores-Cruz, J., Lara-Velázquez, P., Gutiérrez-Andrade, M. A., De-Los-Cobos-Silva, S. G., and Rincón-García, E. A. (2017). Un sistema clasificador utilizando coloración de gráficas suaves. *Revista de Matemática Teoría y Aplicaciones*, 24:129–156.

[Galán, 2017] Galán, S. F. (2017). Simple decentralized graph coloring. *Computational Optimization and Applications*, 66(1):163–185.

[Goldbloom and Hamner, 2010] Goldbloom, A. and Hamner, B. (2010). Kaggle: Your Machine Learning and Data Science Community.

[Guccione and Guccione, 2017] Guccione, J. A. and Guccione, J. J. (2017). Espacios métricos. In *Espacios métricos*, chapter 1, pages 1–7.

[Hausner, 2010] Hausner, A. (2010). A new clustering algorithm for coordinate-free data. *Pattern Recognition*, 43(4):1306–1319.

[Jleli and Samet, 2015] Jleli, M. and Samet, B. (2015). A generalized metric space and related fixed point theorems. *Fixed Point Theory and Applications*, 2015(1):61.

[Kiris, 2017] Kiris, M. (2017). Multiplicative generalized metric spaces and fixed point theorems. *Journal of Mathematical Analysis*, 8(1):212–224.

[Kolay et al., 2019] Kolay, S., Pandurangan, R., Panolan, F., Raman, V., and Tale, P. (2019). Harmonious coloring: Parameterized algorithms and upper bounds. *Theoretical Computer Science*, 772:132–142.

[Lara-Velázquez et al., 2015] Lara-Velázquez, P., Gutiérrez-Andrade, M. Á., De-Los-Cobos-Silva, S. G., and Rincón-García, E. A. (2015). Coloración de gráficas suaves. *Revista de Matemática Teoría y Aplicaciones*, 22(2):311–323.

[Lloyd, 1982] Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[Love and Dowling, 1985] Love, R. F. and Dowling, P. D. (1985). Optimal Weighted Lp Norm Parameters for Facilities Layout Distance Characterizations. *Management Science*, 31(2):200–206.

[MacQueen, 1967] MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.*, pages 281–297.

[Mexicano et al., 2016] Mexicano, A., Rodríguez, R., Cervantes, S., Montes, P., Jiménez, M., Almanza, N., and Abrego, A. (2016). The Early Stop Heuristic: A New Convergence Criterion for K-means. *AIP Conference Proceedings*, 1738(1):310003–310004.

[Pannu, 2015] Pannu, A. (2015). Artificial Intelligence and its Application in Different Areas. *International Journal of Engineering and Innovative Technology*, 4(10):79–84.

[Pearson, 1931] Pearson, E. S. (1931). The analysys of variance in cases of non-normal variation. *Biometrika*, 23(1-2):114–133.

[Peña et al., 2018] Peña, M., Cerrada, M., Alvarez, X., Jadán, D., Lucero, P., Milton, B., Guamán, R., Sánchez, R.-V., Li, C., and de Oliveira, J. V. (2018). Feature engineering based on ANOVA, cluster validity assessment and KNN for fault diagnosis in bearings. *Journal of Intelligent & Fuzzy Systems*, 34(6):3451–3462.

[Pérez-Ortega et al., 2018] Pérez-Ortega, J., Almanza-Ortega, N. N., and Romero, D. (2018). Balancing effort and benefit of K-means clustering algorithms in Big Data realms. *PLoS ONE*, 13(9):1–19.

[Pham et al., 2019] Pham, T. T., Lobos, G. A., and Vidal-Silva, C. L. (2019). Innovación en Minería de Datos para el Tratamiento de Imágenes: Agrupamiento K-media para Conjuntos de Datos de Forma Alargada y su Aplicación en la Agroindustria. *Información tecnológica*, 30(2):135–142.

[Seijas, 2017] Seijas, S. P. (2017). El Problema de Coloración de Grafos. Master's thesis, Universidad de Santiago de Compostela.

[Steinhaus, 1956] Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bulletin de l'académie polonaise des sciences*, 4(12):801–804.

[Thangthong and Khemphet, 2018] Thangthong, C. and Khemphet, A. (2018). Coincidence Point Theorems for (α,β,γ)-Contraction Mappings in Generalized Metric Spaces. *International Journal of Mathematics & Mathematical Sciences*, pages 1–7.

[Uluçay et al., 2019] Uluçay, V., Kılıç, A., Şahin, M., and Deniz, H. (2019). A New Hybrid Distance-Based Similarity Measure for Refined Neutrosophic Sets and its Application in Medical Diagnosis. *Matematika*, 35(1):83–96.

[Urueta-Hinojosa et al., 2019] Urueta-Hinojosa, D. E., Lara-Velázquez, P., Gutiérrez-Andrade, M. Á., and De-Los-Cobos-Silva, S. G. (2019). Classic colouring problems as special cases of the soft graph colouring model. *Int. J. Technology, Policy and Management*, 19(2):131–148.

[Vásquez-Calderón et al., 2018] Vásquez-Calderón, H. E., Lara-Velázquez, P., De-Los-Cobos-Silva, S. G., and Gutiérrez-Andrade, M. Á. (2018). Scatter search for the soft graph colouring problem. *Int. J. Business Continuity and Risk Management*, 8(3):200–218.

[Xu and Wunsch, 2005] Xu, R. and Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678.

[Yang et al., 2017] Yang, J., Ma, Y., Zhang, X., Li, S., and Zhang, Y. (2017). An Initialization Method Based on Hybrid Distance for k -Means Algorithm. *Neural Computation*, 29(11):3094–3117.

[Zarinbal, 2009] Zarinbal, M. (2009). Distance Functions in Location Problems. In Reza Zanjirani, F. and Masoud, H., editors, *Facility Location: Concepts, Models, Algorithms and Case Studies*, chapter 1, pages 5–17. Springer Dordrecht Heidelberg London New York.

[Ziwei Liu and Tang, 2015] Ziwei Liu, Ping Luo, X. W. and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00090
Matrícula: 2183802157

Uso de métricas generalizadas en clasificadores.

Con base en la Legislación de la Universidad Autónoma Metropolitana, en la Ciudad de México se presentaron a las 12:30 horas del día 19 del mes de febrero del año 2021 POR VÍA REMOTA ELECTRÓNICA, los suscritos miembros del jurado designado por la Comisión del Posgrado:

DRA. HERICA SANCHEZ LARIOS
DR. ROMAN ANSELMO MORA GUTIERREZ
DR. ERIC ALFREDO RINCON GARCIA

Bajo la Presidencia de la primera y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (CIENCIAS Y TECNOLOGIAS DE LA INFORMACION)

DE: CARLOS ALBERTO HERNANDEZ NAVA

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, la presidenta del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

UNIVERSIDAD AUTONOMA METROPOLITANA
Casa abierta al tiempo
DIRECCIÓN DE SISTEMAS ESCOLARES

CARLOS ALBERTO HERNANDEZ NAVA
ALUMNO

REVISÓ

MTRA. ROSALIA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JESUS ALBERTO OCHOA TAPIA

PRESIDENTA

DRA. HERICA SANCHEZ LARIOS

VOCAL

DR. ROMAN ANSELMO MORA GUTIERREZ

SECRETARIO

DR. ERIC ALFREDO RINCON GARCIA

El presente documento cuenta con la firma –autógrafa, escaneada o digital, según corresponda- del funcionario universitario competente, que certifica que las firmas que aparecen en esta acta – Temporal, digital o dictamen- son auténticas y las mismas que usan los c.c. profesores mencionados en ella