



# Análisis de datos

## Laboratorio 2: Análisis estadístico

Profesor Max Chacón

Ayudante Adolfo Guzmán Leiva

### Objetivos

- Extraer el conocimiento del problema asignado, mediante el uso del software R, utilizando el algoritmo de clustering K-means y realizar el análisis respectivo.
- Comparar los resultados con lo expuesto en la literatura encontrada y ver si se sustenta el conocimiento obtenido.
- Analizar por grupo e identificar aquellas características más relevantes, si clasifica mejor a una clase que otra e inferir conocimiento respecto a ello.

### Informe

El informe se debe regir por el reglamento de titulación v 1.3, apéndice C, apartado C.3 y contener los siguientes puntos:

Informe	Puntos a evaluar	Porcentaje
	Presentación, ortografía y redacción	5%
	Introducción (Máximo 1 página)	
	Marco Teórico. Clustering, algoritmo K-means y distancias utilizadas (Máximo 2 páginas)	15%
	Pre-procesamiento: Se deben definir criterios para eliminar registros o columnas que presenten datos perdidos, outliers o que no aporten información relevante para el estudio del problema. En caso de normalizar datos fundamentar la decisión (Máximo 6 páginas)	15%
	Obtención del Clúster: Variar parámetros de la función a utilizar de manera que se genere un clúster adecuado en base a las métricas de eficiencia en la clasificación, justifique la utilización de aquellos parámetros, además del criterio para seleccionar el clúster más adecuado. Decida y fundamente el criterio de proximidad de acuerdo a los datos (Máximo 6 páginas)	10%
	Análisis de los resultados: Analizar el clúster e identificar aquellas características que sean más interesantes, entregando su significado en el dominio del problema y contrastar esta información con lo expuesto en la literatura.	30%
	Conclusiones: Respecto a los resultados obtenidos, el desarrollo del laboratorio y el método utilizado. Menciona aspectos positivos y a mejorar en este desarrollo. (Máximo 2 páginas)	20%
	Referencias: usar formato APA 6	5%
	Anexo: código fuente en R	

## Observaciones

---

- Todas las consultas deben ser realizadas al mail [adolfo.guzman@usach.cl](mailto:adolfo.guzman@usach.cl)
- La entrega debe ser subida al sitio Web Usach-Virtual hasta las 23.55 hr del día Viernes 16 de Septiembre.
- Es necesario realizar TODAS las experiencias para aprobar el laboratorio
- Cualquier página más allá del máximo permitido no será revisada.  
Para trabajar con R deben descargar el package “Cluster” y utilizar la función “pam”.  
<http://cran.r-project.org/web/packages/cluster/cluster.pdf>
- Deben decidir y justificar la normalización de las variables.
- Deben justificar el criterio de proximidad a usar con el algoritmo.