# Pontificia Universidad Javeriana - Wikipedia, la enciclopedia libre

**APRENDIZAJE DE MÁQUINA** - **PARCIAL 2**

**Fecha: mayo 30 de 2025. Duración: 3 horas**

# REGLAMENTO DE ESTUDIANTES de la PONTIFICIA UNIVERSIDAD JAVERIANA

*Faltas disciplinarias graves y gravísimas:*

Artículo 123 inciso (d): El fraude en actividades, trabajos y evaluaciones académicos y la posesión o utilización de material no autorizado en los mismos. (Falta disciplinaria grave)

Artículo 124 inciso (b): Todas las modalidades de plagio. (Falta disciplinaria gravísima)

Artículo 124 inciso (e): La suplantación en una evaluación académica, en exámenes preparatorios, en trabajos de grado y tesis. (Falta disciplinaria gravísima)

*Sanciones Disciplinarias:*

Artículo 127: Las faltas graves serán sancionadas con amonestación escrita con cargo a la hoja de vida del estudiante y la imposición de matrícula condicional durante el tiempo necesario para cumplir la condición.

Artículo 128: Adicional a la sanción disciplinaria, el fraude en actividades, trabajos y evaluaciones académicos se sancionará académicamente con la pérdida de la asignatura, la cual será calificada con nota definitiva de cero punto cero (0.0).

NO SE PERMITE EL USO DE HERRAMIENTAS DE CORREO, CHAT, DRIVES, o HERRAMIENTAS DE IA COMO CHATGPT, DEEPSEEK, ETC. SOLO PUEDE USAR LA PLATAFORMA DE BS ASÍ COMO GOOGLE PARA CONSULTA DE INFORMACIÓN GENERAL.

NO SE PERMITE USO DE CELULARES O TABLETS. PUEDE USER APUNTES DE CLASE.

IMPORTANTE

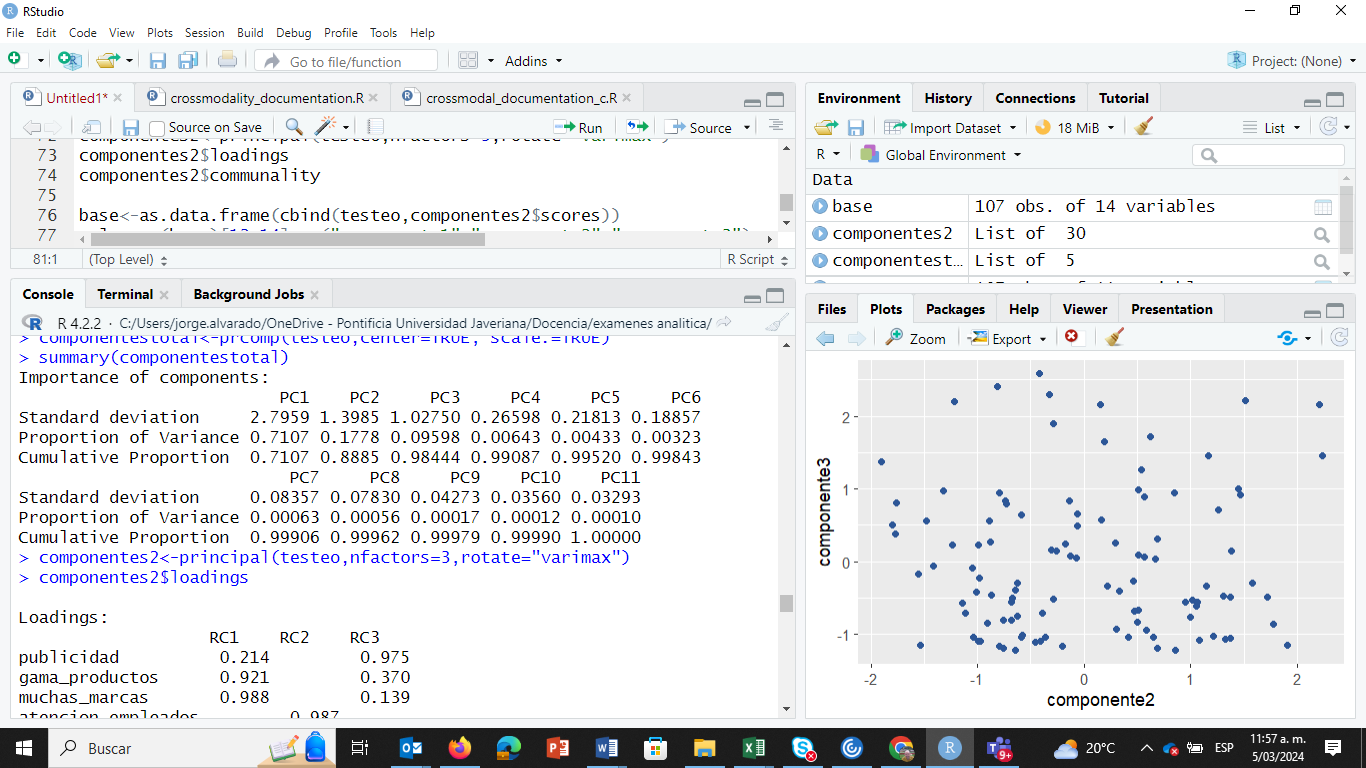
* Responda las preguntas del Punto 1 dentro de este mismo archivo de Word. Mientras responde este punto, vaya ejecutando el cargue de librerías para el punto 2.
* Para el punto 2, debe generar un notebook y plasmar allí las respuestas a las preguntas planteadas.
* Solo puede utilizar los comandos vistos en clase en R.
* Para todos los comandos que involucren una selección aleatoria, por ejemplo, dividir el dataset entre entrenamiento y prueba, o seleccionar registros de un dataset, fije siempre la semilla con los **últimos 3 números de su cédula** de ciudadanía.
* No puede usar herramientas de Inteligencia Artificial como ChatGPT o DeepSeek. Tampoco puede usar herramientas de correo, chats, o drives.

**PUNTO 1 (20). PCA**

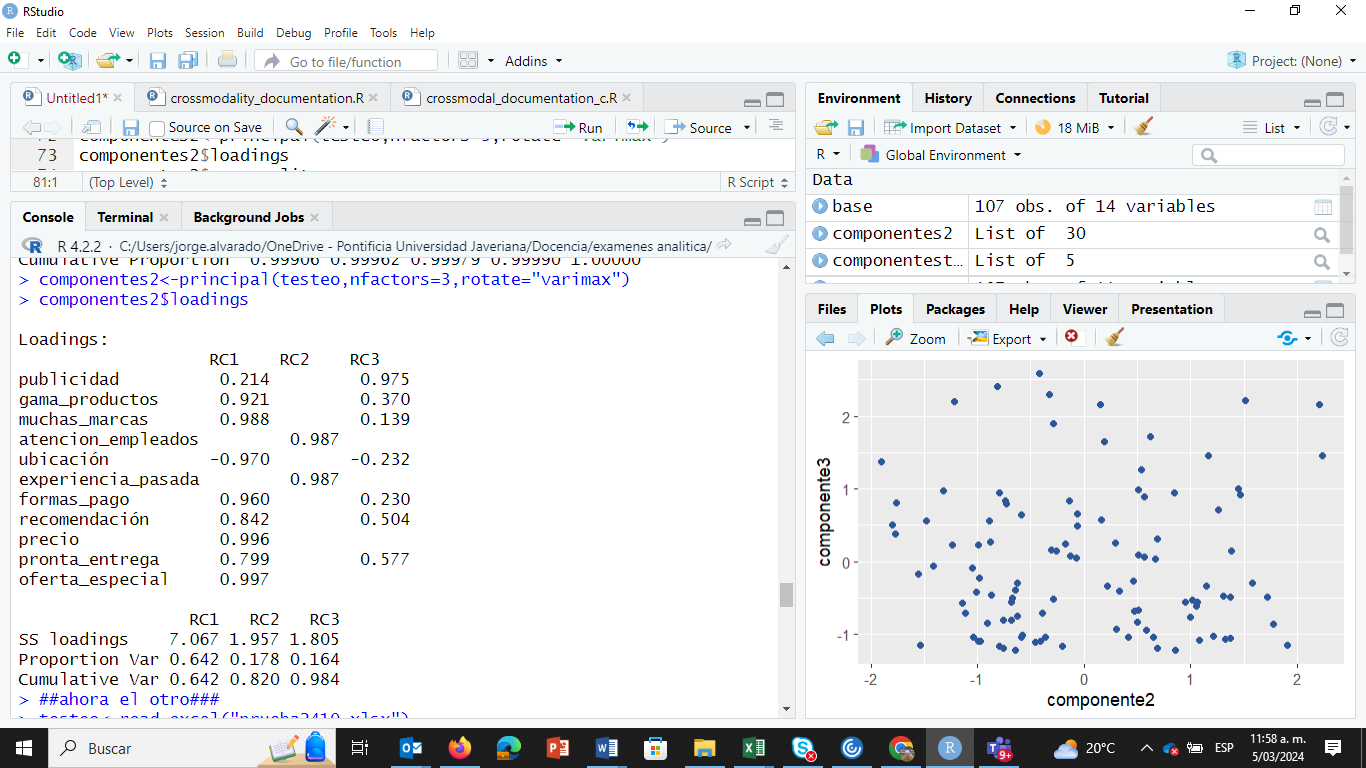
Se realizó una encuesta aleatoria a 1200 personas en la ciudad de Manizales con respecto a los motivadores fundamentales para dirigirse a un punto de venta para la compra de productos de comida y aseo (11 motivadores principales). Cada motivador fue calificado en una escala de 0 a 100. Los motivadores son:

|  |  |
| --- | --- |
| **Variable** | **Definición (importancia de 0 a 100)** |
| Publicidad | Haber recibido publicidad |
| Gama\_productos | Encontrar una amplia gama de productos |
| Muchas\_marcas | Encontrar una amplia gama de marcas |
| Atención\_empleados | Atención por parte de los empleados |
| Ubicación | Ubicación del punto de venta |
| Experiencia\_pasada | Experiencia pasada en el punto de venta |
| Formas\_pago | Existencia de múltiples formas de pago |
| Recomendación | Recomendación del punto por conocidos |
| Precio | Precio de los artículos |
| Pronta\_entrega | Entrega rápida de domicilios |
| Oferta\_especial | Existencia de ofertas y descuentos especiales |

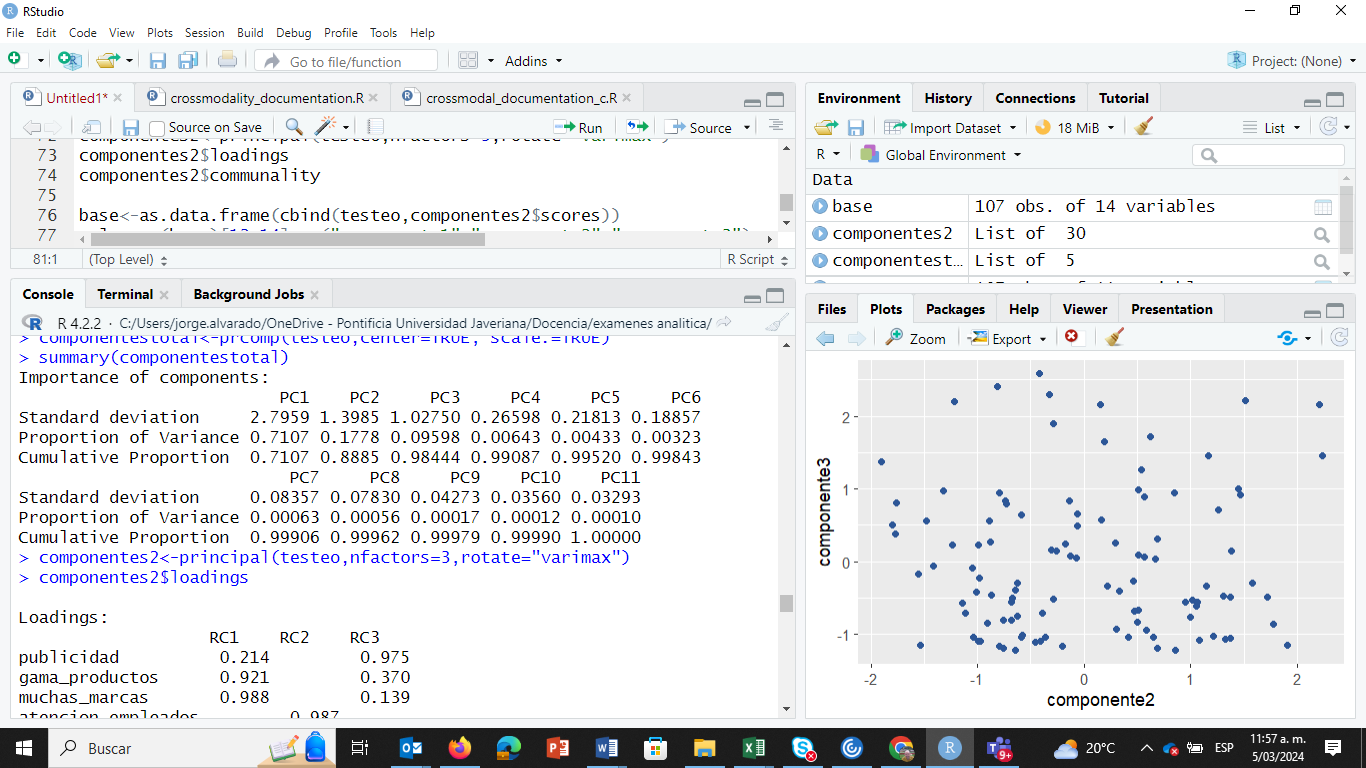
Se aplicó el algoritmo de componentes principales para tratar de encontrar factores que impulsan/motivan la compra de productos de comida y aseo en el punto específico de venta, obteniendo el siguiente resumen:



Posteriormente se aplicó una rotación varimax a los componentes principales con una solución de 3 componentes, obteniendo las siguientes cargas factoriales:



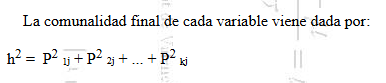
1. [4] Explique si la solución de 3 componentes representa adecuadamente la base de datos; en caso contrario, proponga y explique por cuántos componentes la cambiaría.

R/ En los componentes principales existen diferentes maneras para seleccionar el numero de componente principales para realizar el análisis, en este caso según la información suministrada. La primera forma es revisando los eigenvalues: usar componentes que expliquen mas de una variable, o sea, de autovalor (eigenvalue) superior a 1. A través de este método se seleccionarían 3 componentes principales, ya que: , ya del 3 componente en adelante el auto valor es menor a 1.

La segunda forma es a traves de la Ley de los pocos vitales usa el 80% de proporcion acumulada como valor empirico de corte. Con este metodo se selecionarian 2 componentes principales, ya que en el segundo componente ya tiene una varianza acumulada del 88.85%.

Por lo tanto en mi opinion 3 componentes puede ser una opcion, sin embargo yo selecionaria 2 componentes ya que solo con 2 componentes ya tiene el 88% de la varianza acumulada, sumado a que al realizar el metodo con los autovalores la tercera comonente es mayor a uno por muy poco (1.02750) y finalmente es mas facil analizar 2 componentes que 3 componentes.

1. [4] Calcule las comunalidades (una para cada variable) a partir de las cargas de cada variable en los 3 componentes mostrados. Si no recuerda cómo calcularlas, puede usar Google para buscar este cálculo. ¿Cuál es la variable mejor representada? ¿Cuál es la variable menos representada?



R/ Por lo tanto, la comunalidad final de cada variable es:

Publicidad: 0.214^2 + 0.975^2  = 0.996421

Gama\_productos: 0.921^2 + 0.370^2 = 0.985141

Muchas\_marcas: 0.988^2 + 0.139^2 = 0.995465

Atención\_empleados: 0.987^2 = 0.974169

Ubicación: 0.970^2 + 0.232^2 = 0.994724

Experiencia\_pasada: 0.987^2 = 0.974169

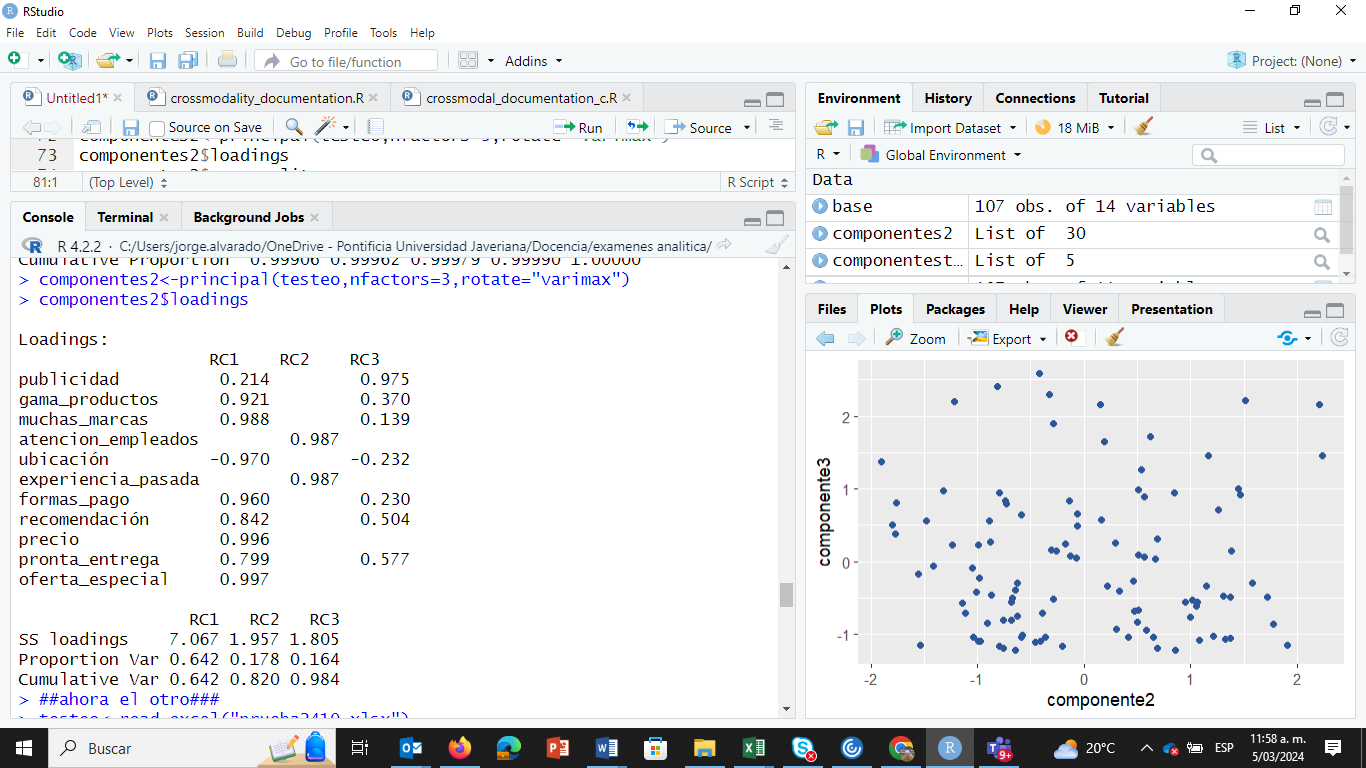
Formas\_pago: 0.960^2 + 0.230^2 = 0.9745

Recomendación: 0.842^2 + 0.504^2 =0.96298

Precio: 0.996^2 = 0.992016

Pronta\_entrega: 0.799^2 + 0.577^2 = 0.97133

Oferta\_especial: 0.997^2 = 0.994009



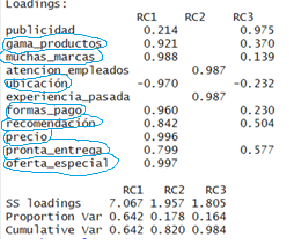
1. [4] Durante la reunión de presentación de resultados, un empleado indicó que este modelo de los motivadores de compra tiene muy poco en cuenta la ubicación del punto de venta. Responda a esta cuestión, soportado de manera técnica en cifras.

R/ En primer lugar en el análisis factorial analizando las comunalidades la variable ubicación que es la ubicación del punto de venta si esta representada en el análisis, el porcentaje de su varianza que la solución factorial representa es de -0.970; es decir que es muy importante en el análisis, pero es inversa a las otras variables que estén en su componente. Es decir, por ejemplo, que si la variable muchas marcas tiene un valor muy alto, entonces la variables ubicación va a tener un valor muy bajo.

Para verificar si una variable aporta al análisis antes de hacer este, se puede hacer la prueba de KMO, en donde en dentro de esta cada variable tiene un MSA para detectar si una variable no esta relacionado con las otras. Puntualmente es este caso si al hacer el calculo si la variable ubicación tiene MSA menor a.70 significa que es mediocre y debe ser eliminada del análisis, si por lo contrario es igual o mayor de 0.70, la variable estará representada por alguno de loc componetes principales que se haga dentro del análisis

1. [4] Explique el primer componente principal en cuanto a variables más representadas, y después explique su significado en términos del problema de negocio.

R/ Las variables mas representadas en el componente uno son:



En donde las mas representadas con el porcentaje de su varianza que la solución factorial representa sea mayor de 0.90 son: Oferta especial, precio, muchas marcas, ubicación, formas\_pago y gama\_productos. Es decir que este componente caracteriza muy bien las variables que contienen información sobre todo lo que tiene que ver con los precios de los productos(si tienen alguna oferta, su precio, el método de pago) y también representa bien las características de los productos(su gama o su marca) y finalmente representa bien su ubicación.

1. [4] Explique qué caracteriza a los clientes de la zona marcada en rojo en el siguiente gráfico.



R/ Los clientes que caigan en la zona marcada en rojo son caracterizados muy bien por el componente 1, por lo tanto, estos clientes tienen como motivaciones fundamentales: Si tiene alguna oferta especial, el precio, si cuentan con una amplia gama de marcas y productos, la ubicación del punto de venta y la forma de pago. Esos son los motivadores de esos clientes para dirigirse a un punto de venta y comprar productos de comida y aseo.

**PUNTO 2 (30). BOOSTING**

Con la temporada navideña acercándose rápidamente, la organización liderada por Santa Claus está en proceso de preparar su equipo de manufactura para determinar qué niños recibirán regalos este año. El departamento encargado de tomar esta decisión, Delivery Gifts, ha buscado la ayuda de expertos y expertas de la Universidad Javeriana para abordar este desafío para la temporada de 2025, utilizando datos históricos recopilados en los años 2023 y 2024. Con un conjunto de datos compuesto por 15 variables de predicción, 1 variable de respuesta y 250 registros históricos (**gifts\_historico.xslx**) se le ha encomendado la tarea de ayudar a Delivery Gifts a clasificar a 25 niños para la próxima Navidad (**gitfs\_proximos.csv**). Usted debe implementar un algoritmo de predicción sobre quienes debe recibir regalo e indicar y evaluar las métricas obtenidas.

1. (6) Prepare los datos con miras al desarrollo de modelos para predecir la variable **reciberegalo**
2. (12) Desarrolle los siguientes modelos para predecir la variable de respuesta, usando los parámetros por defecto:
   1. Gradient Boosting (GBM)
   2. XGBoost
   3. LightGBM
3. (8) Seleccione alguno de los modelos desarrollados en (2) y tómelo como el modelo base.
   1. Haga un cambio adecuado en los parámetros para provocar un sobre-ajuste del modelo. Justifique la respuesta.
   2. Haga un cambio adecuado en los parámetros para provocar un sub-ajuste del modelo. Justifique la respuesta.
4. (4) Seleccione el mejor modelo desarrollado en (2), úselo para predecir la variable de respuesta en el archivo **gitfs\_proximos.csv**. Muestre sus resultados en el mismo *dataframe* donde están las variables de dicho archivo.