Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Tarea 05. Preprocesamiento de datos

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495
Luis Enrique García Gómez	315063880

1. Lectura de Artículo

1.1. Resumen

Introducción

Se describe de la importancia del preprocesamiento de datos en el análisis de datos operativos de edificios, destacando los retos de trabajar con operaciones complejas de edificios y datos de baja calidad y a la vez introduce las diversas técnicas que se tratarán en la revisión.

Marco general para el preprocesamiento de datos operativos de edificios

Muestra un marco general para el preprocesamiento de datos en el análisis de datos operativos de edificios y describe las cinco tareas principales del preprocesamiento de datos: limpieza, reducción, escalado, transformación y partición.

Métodos de limpieza de datos para crear análisis de datos operativos

Trata sobre las técnicas de limpieza de datos, que se utilizan para mejorar la calidad de los datos brutos mediante la eliminación de valores atípicos y la imputación de valores perdidos, después describe varios métodos para la detección de valores atípicos y la imputación de valores perdidos y se analizan las ventajas y desventajas de cada enfoque.

Escalado de datos

Esta sección habla sobre la transformación de los datos originales en rangos similares para el modelado predictivo y describe tres enfoques principales para el escalado de datos, incluidos los métodos basados en rangos de datos, basados en la distribución y basados en la estructura, el objetivo del escalado de datos es hacer que los datos sean más adecuados para el análisis reduciendo el impacto de las diferencias de escala entre las variables.

Transformación de datos

Analiza el proceso de organización de los datos originales en formatos adecuados para diversos algoritmos de extracción de datos, que incluye dos tareas, la transformación de datos numéricos, que transforma los datos numéricos en datos categóricos, y la transformación de datos categóricos, que transforma los datos categóricos en datos numéricos. El objetivo de la transformación de datos es hacer que los datos sean más adecuados para el análisis convirtiéndolos en un formato que pueda analizarse más fácilmente.

Partición de datos

Describe el proceso de división del conjunto de datos en diferentes grupos en función de las características de funcionamiento del edificio. El objetivo de la partición de datos es aumentar la sensibilidad y fiabilidad del análisis de seguimiento. Se analizan varios métodos de partición de datos, como la agrupación, el muestreo estratificado y el muestreo aleatorio.

Aumento de datos

Presenta técnicas avanzadas de preprocesamiento de datos desarrolladas a partir del aumento de datos, como el aumento de datos, que es una técnica utilizada para aumentar el tamaño de un conjunto de datos mediante la creación de nuevos datos a partir de datos existentes, también se describen varios métodos de aumento de datos, incluido el aumento de datos de imágenes y texto, y se analizan las ventajas y limitaciones de cada enfoque.

Aprendizaje por transferencia

Habla acerca de el proceso de transferencia de conocimientos de un dominio a otro, como el aprendizaje por transferencia, que es una técnica utilizada para mejorar el rendimiento de un modelo de aprendizaje automático aprovechando los conocimientos adquiridos en una tarea o dominio relacionado y después describe varios métodos de aprendizaje por transferencia, como el ajuste fino, la extracción de características y la adaptación al dominio.

Aprendizaje semisupervisado

Describe el proceso de aprendizaje a partir de datos parcialmente etiquetados, como el aprendizaje semisupervisado, el cual es una técnica utilizada para mejorar el rendimiento de un modelo de aprendizaje automático aprovechando tanto los datos etiquetados como los no etiquetados y analiza varios métodos de aprendizaje semisupervisado, como el autoentrenamiento, el coentrenamiento y el aprendizaje multivista.

Conclusiones

Ofrece recomendaciones para futuras investigaciones y destaca la importancia del preprocesamiento de datos en la construcción de análisis de datos operativos, subrayando la necesidad de realizar más esfuerzos de investigación hacia la automatización de las tareas de preprocesamiento de datos para mejorar la eficiencia del análisis de datos. También se discuten los beneficios potenciales de las técnicas avanzadas de preprocesamiento de datos, como el aumento de datos, el aprendizaje por transferencia y el aprendizaje semisupervisado, para mejorar la calidad y la fiabilidad del análisis de datos operativos de edificios.

1.2. Adrian Aguilera Moreno

El pre-procesamiento de datos es el proceso de limpiar, transformar y preparar los datos para su análisis posterior. El objetivo del pre-procesamiento de datos es mejorar la calidad de los datos y hacerlos más adecuados para el análisis posterior. Es un paso indispensable en el análisis de datos operativos de edificios debido a la complejidad intrínseca de las operaciones de construcción y las deficiencias en la calidad de los datos. Este artículo sirve como una revisión exhaustiva de las técnicas de pre-procesamiento de datos para analizar datos operativos de edificios masivos.

El pre-procesamiento de datos operativos de negocios consta de cinco tareas principales: limpieza de datos, reducción, escalado, transformación y partición. La limpieza de datos tiene como objetivo mejorar la calidad de los datos mediante la imputación de valores faltantes y la eliminación de valores atípicos. Los datos operativos de negocios suelen almacenarse en tablas de datos bidimensionales, donde cada fila representa una

observación recopilada en un momento específico y cada columna representa una variable de negocio.

La reducción de datos es una técnica de pre-procesamiento de datos que se utiliza para reducir la cantidad de datos en un conjunto de datos. La reducción de datos se realiza típicamente en dos direcciones: reducción de muestras de datos (filas) y reducción de variables de datos (columnas). La reducción de muestras de datos se puede lograr mediante técnicas de muestreo aleatorio o estratificado, mientras que la reducción de variables de datos se puede lograr mediante técnicas de selección de características o extracción de características. El objetivo de la reducción de datos es reducir la complejidad del conjunto de datos y, por lo tanto, reducir los costos computacionales asociados con el análisis de datos.

Las técnicas de vanguardia en la ciencia de datos son:

- **Aumento de datos:** esta técnica implica la generación de datos sintéticos a partir de los datos existentes para aumentar el tamaño del conjunto de datos y mejorar la capacidad del modelo para generalizar a nuevos datos.
- **Aprendizaje por transferencia:** esta técnica implica el uso de un modelo pre-entrenado en un conjunto de datos relacionado para mejorar el rendimiento en un conjunto de datos objetivo.
- **Aprendizaje semi-supervisado:** esta técnica implica el uso de un conjunto de datos que contiene tanto datos etiquetados como no etiquetados para mejorar el rendimiento del modelo en la tarea de clasificación.

La normalización max-min escala los datos a un rango entre cero y uno restando el valor mínimo de la variable a cada valor y luego dividiendo por el rango de la variable (es decir, la diferencia entre los valores máximo y mínimo). Este método es sensible a los valores atípicos de los datos, ya que su presencia puede cambiar drásticamente el rango de los datos. Por el contrario, la estandarización z-score es menos afectada por los valores atípicos. Escala los datos para que tengan una media de cero y una desviación estándar de uno restando la media de la variable a cada valor y luego dividiendo por la desviación estándar de la variable. Este método se utiliza típicamente para reformar la variable para que tenga una distribución normal con una media de cero y una desviación estándar de uno.

La transformación de datos numéricos se realiza para transformar los datos numéricos en datos categóricos, mientras que la transformación de datos categóricos se realiza para transformar los datos categóricos en datos numéricos. Una técnica comúnmente utilizada para transformar variables categóricas en numéricas es la codificación one-hot, en la que se genera una matriz de L_1 columnas para una variable categórica con L niveles.

La partición de datos se puede realizar mediante técnicas de clustering, que agrupan los datos en grupos o clústeres según su similitud. Los métodos de clustering comunes incluyen k-means, EWKM, PAM y fuzzy c-means clustering.

La idea principal detrás del aumento de datos es generar datos sintéticos que tengan distribuciones de datos similares a los datos reales. Estos datos sintéticos se pueden integrar con los datos reales para mejorar el rendimiento de los modelos de predicción basados en datos. En el campo de la construcción, el aumento de datos se utiliza para abordar el problema potencial de escasez de datos en edificios individuales. Por ejemplo, si los datos de operación del edificio son escasos debido a la falta de tiempo de acumulación de datos para edificios nuevos o la falta de sistemas de recolección de datos automatizados para edificios existentes, se

pueden generar datos sintéticos adicionales para mejorar la calidad y cantidad de los datos disponibles. Es importante tener en cuenta que el aumento de datos debe realizarse con precaución, ya que la generación de datos sintéticos que no reflejen adecuadamente la distribución de los datos reales puede afectar negativamente el rendimiento del modelo.

1.3. Marco Antonio Rivera Silva

El pre-procesamiento de datos es una fase crucial en el análisis de datos que comprende una serie de tareas destinadas a limpiar, transformar y preparar los datos para su posterior análisis. Su objetivo fundamental es mejorar la calidad de los datos y hacerlos adecuados para su uso en modelos y algoritmos de análisis. Esta etapa es especialmente esencial en el contexto de los datos operativos de edificios, debido a la complejidad inherente de las operaciones de construcción y las posibles deficiencias en la calidad de los datos recopilados.

El pre-procesamiento de datos operativos de edificios abarca cinco tareas fundamentales. La primera es la "limpieza de datos", que busca mejorar la calidad de los datos mediante la imputación de valores faltantes y la eliminación de valores atípicos o erróneos. Dado que estos datos suelen ser recopilados de múltiples fuentes y sistemas, la limpieza es esencial para garantizar la integridad de los datos.

La segunda tarea es la "reducción de datos", una técnica que busca disminuir la cantidad de datos en un conjunto sin sacrificar información crítica. La reducción puede ocurrir en dos dimensiones: reducción de muestras de datos (filas) o reducción de variables de datos (columnas). Esto ayuda a reducir la complejidad del conjunto de datos y los costos computacionales asociados al análisis.

La tercera tarea se enfoca en el "escalado de datos". Dos métodos comunes son la normalización "max-min", que escala los datos al rango entre cero y uno, y la estandarización "z-score", que los ajusta para tener una media de cero y una desviación estándar de uno. Estos métodos permiten comparar y analizar variables en la misma escala, pero su elección depende de la naturaleza de los datos y la sensibilidad a valores atípicos.

La cuarta tarea es la "transformación de datos", que se realiza para cambiar la naturaleza de las variables. Esto puede ser la transformación de datos numéricos en datos categóricos o viceversa. Por ejemplo, la codificación "one-hot" es una técnica común para convertir variables categóricas en numéricas.

La quinta tarea, "partición de datos", se lleva a cabo utilizando técnicas de clustering que agrupan los datos según similitudes. Métodos de clustering como k-means, EWKM, PAM y fuzzy c-means ayudan a organizar los datos en grupos significativos, lo que puede ser útil en análisis posteriores.

En el ámbito de la ciencia de datos, se han desarrollado técnicas avanzadas, entre las que destacan:
Aumento de datos: Consiste en generar datos sintéticos a partir de los datos existentes para incrementar el tamaño del conjunto de datos y mejorar la capacidad de los modelos para generalizar a nuevas observaciones.
Aprendizaje por transferencia: Implica el uso de modelos pre-entrenados en conjuntos de datos relacionados para mejorar el rendimiento en un conjunto de datos objetivo.
Aprendizaje semi-supervisado: Utiliza conjuntos de datos que contienen tanto datos etiquetados como no etiquetados para mejorar el rendimiento del modelo en tareas de clasificación.
En la transformación de datos, es importante mencionar que la normalización y la estandarización son técnicas clave. La normalización "max-min" escala los datos al rango entre cero y uno, lo que es sensible a los valores atípicos. En contraste, la estandarización "z-score" es menos afectada por los valores atípicos, ya que

transforma los datos para tener una media de cero y una desviación estándar de uno, lo que es útil para lograr una distribución normal.
El aumento de datos es una estrategia valiosa en el campo de la construcción para abordar la posible escasez de datos en edificios. Esta técnica implica la generación de datos sintéticos que se asemejen a los datos reales, lo que puede mejorar la calidad y cantidad de los datos disponibles. Sin embargo, es esencial realizar el aumento de datos con precaución para que los datos sintéticos reflejen con precisión la distribución de los datos reales, ya que datos sintéticos mal generados pueden afectar negativamente el rendimiento del modelo.

1.4. Sebastían Alejandro Gutiérrez Medina

El artículo es una revisión exhaustiva de las técnicas de preprocesamiento de datos para el descubrimiento de conocimiento eficiente y confiable a partir de datos operativos de edificios. El rápido desarrollo de la ciencia de datos y la creciente disponibilidad de datos operativos de edificios han proporcionado grandes oportunidades para desarrollar soluciones impulsadas por datos para la gestión inteligente de la energía de los edificios.

El autor proporciona una revisión exhaustiva de las técnicas de preprocesamiento de datos para mejorar la calidad y la eficiencia del análisis de datos operativos de edificios entre los temas laterales se incluyen la eliminación de valores atípicos, la imputación de valores faltantes, la reducción de datos, el escalado, la transformación, el particionamiento, el aumento de datos, la transferencia de aprendizaje y el aprendizaje semi-supervisado.

Además, el articulo expone y enfatiza la importancia de estas técnicas para mejorar la sensibilidad y confiabilidad del análisis posterior y proporciona recomendaciones para futuras investigaciones en este campo.

En conclusión, es de suma importancia utilizar tecnicas de preprocesamiento para poder asegurar y mejorar la calidad del analisis de datos y , al mismo tiempo, reducir el tiempo que lleva realizar el análisis, sin embargo tambien es importante validar el resultado de estas técnicas de preprocesamiento para poder comprobar y asegurar su utilidad y correctez.

1.5. Israel Hernández Dorantes

En el artículo se realiza un análisis sobre las **técnicas de preprocesamiento de datos** mostrando las características y aplicaciones de cada una de las técnicas. El autor plantea *resaltar la importancia del preprocesamiento de datos* como base para un análisis válido de datos operativos. Asimismo, hace énfasis en el rápido desarrollo en la ciencia de datos y la creciente disponibilidad de datos operativos, los cuales ofrecen grandes oportunidades para desarrollar soluciones impulsadas por datos en la gestión energética de edificios. Además, recalca que el preprocesamiento de datos es una etapa indispensable en este análisis debido a la complejidad de las operaciones de edificaciones y las posibles deficiencias que puedan existir en la calidad de los datos.

La **temática central** del artículo es el preprocesamiento de datos y su papel fundamental en el análisis de datos operativos de edificaciones, donde se identifican una serie de técnicas, como la eliminación de valores atípicos y la imputación de valores faltantes, que son fundamentales para mejorar la calidad de los datos. Entre otras más técnicas como el aprendizaje por transferencia y el aprendizaje semi-supervisado para abordar

desafíos prácticos en el campo del diseño de edificaciones.

En cuanto a la **práctica profesional**, la importancia del preprocesamiento de datos se relaciona evidentemente en la *organización energética de edificaciones*, donde el preprocesamiento es esencial para garantizar la *validez* y *confiabilidad* de los resultados del análisis. Pues la presencia de valores faltantes y valores atípicos puede ser común debido a fallas en la recolección, transmisión y almacenamiento de datos, por lo cual realizando las técnicas de preprocesamiento de datos permite abordar estos problemas y realizar un análisis confiable.

Como conclusión, debemos de reconocer que el preprocesamiento de datos desempeña un papel importante en la organización energética de edificaciones; y la cantidad de datos disponibles, junto con su calidad, influyen directamente en la efectividad de las soluciones planteadas. Técnicas como la imputación de valores faltantes, la detección de valores atípicos, entre otra más, son fundamentales para garantizar la confiabilidad en los resultados analíticos. Sin embargo, es importante mencionar que el preprocesamiento de datos no es una tarea trivial pues requiere de una gran parte en el esfuerzo total en la tarea de minería de datos. Por lo que, la elección de las **técnicas de preprocesamiento de datos** debe ser cuidadosa, pues se debe de considerar un balance entre la precisión y los costos o complejidades computacionales que resulten.

1.6. Alejandra Ortega García

Este artículo resalta la importancia del preprocesamiento de datos en el análisis de datos operativos de edificios, subrayando los desafíos inherentes a trabajar con datos complejos y variables. A través de una revisión exhaustiva, el autor proporciona una visión más clara de esta etapa crucial en el proceso de análisis de datos operativos de edificios.

El preprocesamiento de datos comprende una serie de técnicas destinadas a mejorar la calidad de los datos brutos, incluyendo la eliminación de valores atípicos y la imputación de valores faltantes. La calidad y la confiabilidad de los resultados finales dependen en gran medida de cómo se aborden las tareas de limpieza, reducción, escalado, transformación y partición de datos. Además, la aplicación de técnicas avanzadas, como el aumento de datos, el aprendizaje por transferencia y el aprendizaje semisupervisado, ofrece el potencial de mejorar aún más la calidad y la eficiencia de este proceso.

En general, se hace hincapié en la importancia de mantenerse al tanto de las últimos métodos para el preprocesamiento de datos operativos de edificios y destaca su capacidad para transformar la forma en que se gestionan y operan los edificios en el futuro. La gestión eficiente de edificios es fundamental en la búsqueda de la sostenibilidad y la optimización de recursos, y el preprocesamiento de datos desempeña un papel crítico en este proceso.

1.7. Luis Enrique García Gómez

El articulo hace gran énfasis que el pre-procesamiento de datos es una fase esencial en el análisis de datos por lo que esta destinada a limpiar, transformar y preparar los datos para su análisis subsecuente. Su propósito primordial es mejorar la calidad de los datos y adaptarlos para su uso en modelos y algoritmos. En el contexto de almacenes y minería de datos, esta etapa es crítica debido a la complejidad inherente de las operaciones de almacenamiento y la posible falta de calidad en los datos recopilados.

Tambien el articulo mensiona que estas se dividen en cinco tareas fundamentales: limpieza de datos para mejorar la calidad mediante la imputación y eliminación de valores atípicos, reducción de datos para dis-

minuir la complejidad, escalado para comparar variables en una misma escala, transformación para cambiar la naturaleza de las variables y partición de datos utilizando técnicas de clustering.

Las técnicas avanzadas, como el aumento de datos, aprendizaje por transferencia y aprendizaje semi-supervisado, se han desarrollado para abordar desafíos específicos en almacenes y minería de datos. Además, la normalización "max-min" y la estandarización "z-score" son técnicas esenciales en el escalado de datos, siendo la primera más sensible a valores atípicos y la segunda menos afectada por ellos.

En ambito profesional, es de suma importancia del preprocesamiento de datos ya que se relaciona directamente en la construccion de modelos que permitan generar modelos predictivos o modelos de clasificacion, que por consiguiente nos permitirian hacer un analisis del comportamiento de los datos para finalmente encontrar patrones y generar conocimiento. Por lo que el preprocesamiento es esencial para garantizar una confianza en que los resultados del análisis sean correctos.

En conclusion un gran impedimento son la presencia de valores faltantes y valores atípicos siendo situaciones muy recurrentes debido a errores e inconsistencias de los datos, entre los principales problemas se encuentran, la falta de estándares, valores perdidos, información no consolidada, comparación compleja e integración, datos fuera de dominio, homónimos, datos faltantes, por lo cual realizando las técnicas de preprocesamiento de datos permite abordar estos problemas y realizar un mejor análisis.

2. Preparación de datos

a. ¿Cuáles son las **principales técnicas** de **preparación de datos**?

> "Las principales técnicas de esta etapa de minado de datos son:
> - Limpieza de datos . Permite rellenar valores perdidos, remover el ruido, resolver inconsistencias en los datos, identificar o eliminar valores atípicos.
> - Integración de datos . Permite mezclar datos provenientes de múltiples fuentes de datos heterogéneas en un repositorio coherente de datos.
> - Transformación de datos . Permite escalar los datos y lograr que caigan en un rango peque~no (0.0 a 1.0).
> - Reducción de datos . Permite reducir el tamaño de los datos (eliminado características redundantes o agrupándolas), obteniendo una representación reducida en volumen que produce los mismos resultados analíticos (o similares)."
>
> Lopez, L. G., & Avilés, G. A. (2021). Minería de datos con R (Primera edición) [Libro Digital], Página 59-60

En cuanto a la limpieza de datos, se debe de poner en practica los procesos de limpieza y mejoramiento de la calidad de los datos, estas acciones permiten detectar, remover redundancia, errores e inconsistencias de los datos, entre los principales problemas se encuentran, la falta de estándares, valores perdidos, información no consolidada, comparación compleja e integración, datos fuera de dominio, homónimos, datos faltantes y datos predeterminados, elementos a considerar al realizar la limpieza en las etapas de : adquisición de datos y metadatos, reformateo, identificación de valores atípicos, suavizado de datos con ruido y corrección de datos inconsistentes.

b. ¿Cuáles son las formas que se tienen para **tratar los valores faltantes** en los datos? Explica algunas de **ventajas** y **desventajas** con respecto al **tamaño del conjunto de datos**.

Tenemos las siguientes formas para tratar los valores faltantes en los datos:

- **Ignorar la tupla**: Es efectivo cuando la tupla no contiene varios atributos con valores faltantes, pero el gran problema es cuando el porcentaje de valores faltantes es variado en relación a los atributos que contiene cada tupla.
- **Llenar el valor manualmente**: Para esta técnica tenemos como ventajas que se tiene un *control total* sobre los valores que se está ingresando y se tiene una gran *flexibilidad* ya que se puede adaptar de una manera más fácil el proceso a las necesidades del análisis que se esté haciendo. Y como desventajas tenemos que toma mucho tiempo y que no es factible cuando se tienen grandes bases de datos con varios valores faltantes.
- **Usar una constante global**: Para esta técnica tenemos como ventajas que es más simple agregar datos constantes ya que no requeriremos de cálculos, y que además la estructura original de los datos se va a preservar. Y como desventaja tenemos que nos puede llevar a conclusiones incorrectas debido a que se pueden distorsionar los datos.
- **Utilizar una medida de tendencia central (imputación)**: Para esta técnica tenemos como ventajas que se preserva la estructura general de los datos, y que además es un proceso simple de implementar. Y como desventajas tenemos que pueden ocurrir sesgos en la distribución de los datos, ya que los valores imputados se basan en valores existentes, y la imputación tiende a modificar la variabilidad de los datos, pues todos los valores imputados serán los mismos.

- **Utilizar la media (o mediana) para todas las muestras que pertenezcan a la misma clase (imputación)**: Para esta técnica tenemos como ventajas que se preserva la estructura de los datos *por las clases*, y además, a diferencia de las otras técnicas antes mencionadas, se reduce la posibilidad de introducir un sesgo en la imputación. Y las desventajas que tenemos son que si existe una gran variabilidad dentro de las clases, puede suceder que la imputación no sea precisa y que lleve a estimaciones poco confiables.
- **Utilizar el valor más probable**: Para esta técnica tenemos como ventajas que se mantiene la estructura general de los datos al reemplazar los valores faltantes con el valor más probable; y es simple de implementar y entender. Y como desventajas tenemos que existiría una falta de variabilidad real en los datos, afectando a los análisis posteriores que se vayan a a realizar, pues todos los valores faltantes se llenan con el mismo valor más probable.

c. Supón que se tiene el **dataset** que se muestra a continuación:

Age	Income	Student	Credit_rating	Buy_computer
21	60,000	Yes	3	No
30	70,000	No	5	No
38		No	2	Yes
45	45,000	Yes	3	Yes
46	25,000	No	2	Yes
47	30,000	Yes	6	No
39	28,000	Yes	5	No
29	48,000	Yes	3	No
50	75,000	Yes	2	No
48		Yes	3	No
30		Yes	6	Yes
51	46,000	No	4	Yes
32	80,000	Yes	2	No
45	50,000	No	4	No

- **Rellena** los **valores perdidos**, indicando el **criterio que utilizaste**. Justifica tu respuesta.

  (a) **Verificar datos faltantes :** Primero de tratar los datos que faltan, es bueno verificar la cantidad de datos que faltan.

  ```
  df_Computer <- read.csv(ruta,header=T)
  colSums(is.na(df_Computer))
  ```

  Tenemos 3 valores faltantes de la columna Income.

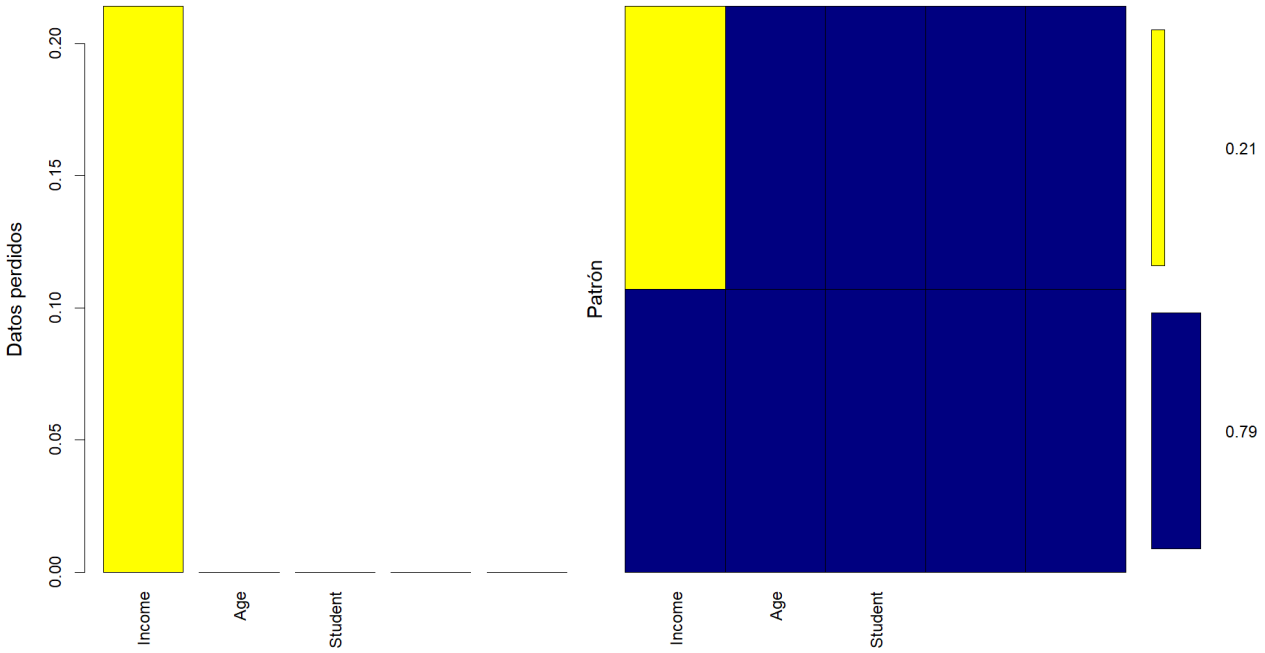


Figure 1: Graficos que muestran el % de Datos faltantes

(b) **Imputación de la media/moda/mediana:** la imputación es un método para completar los valores faltantes con valores estimados, utilizando relaciones conocidas que puedan identificarse en los valores válidos del conjunto de datos para ayudar a estimar los valores faltantes.

Aplicaremos una **Imputación generalizada:** en este caso, calculamos la media o la mediana para todos los valores que no faltan de esa variable y luego reemplazamos el valor faltante con la media o la mediana.

```
df_Computer_imputed <- mice(df_Computer, m = 5, maxit = 50,
                    method = 'pmm', seed = 500)
```

- **m** — Se refiere a 5 conjuntos de datos imputados
- **maxit** — Se refiere al no. de iteraciones realizadas para imputar valores perdidos
- **método** — Hace referencia al método utilizado en la imputación. utilizamos emparejamiento predictivo de medias.

```
df_Computer_imputed$imp$Income
datos_completos <- complete(df_Computer_imputed, 2)
```

```
> df_Computer_imputed$imp$Income
      1     2     3     4     5
3  60000 70000 25000 60000 48000
10 45000 28000 28000 48000 50000
11 70000 46000 48000 48000 45000
```

Figure 2: Conjuntos de datos imputados

El primer comando nos permite verificar los valores imputados, como hay 5 conjuntos de datos imputados, puede seleccionar cualquiera usando la función *complete()*.

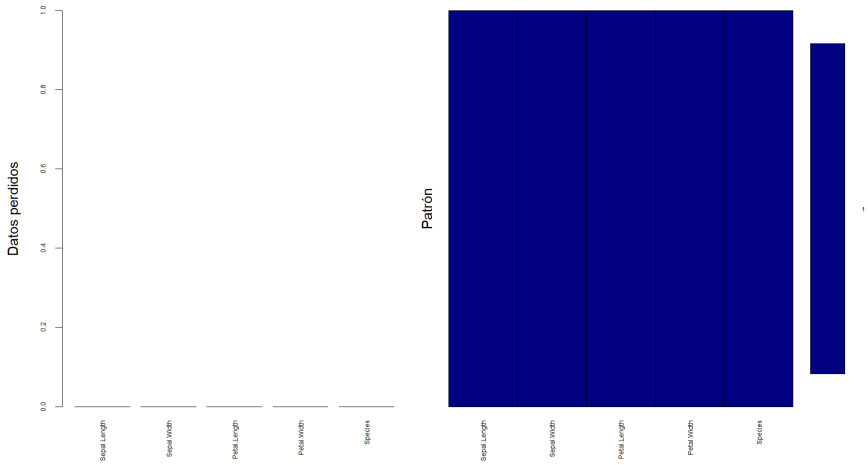


Figure 3: Graficos que muestran el % de Datos faltantes

Dataframe resultante usando imputación por media generalizada. La imputación por media, empleada para reemplazar valores faltantes, se justifica por su simplicidad y eficiencia. Sin embargo, esta estrategia no captura la variabilidad de los datos originales y puede introducir sesgos, especialmente si los valores faltantes siguen un patrón específico.

Age	Income	Student	Credit_rating	Buy_computer
21	60,000	Yes	3	No
30	70,000	No	5	No
38	70,000	No	2	Yes
45	45,000	Yes	3	Yes
46	25,000	No	2	Yes
47	30,000	Yes	6	No
39	28,000	Yes	5	No
29	48,000	Yes	3	No
50	75,000	Yes	2	No
48	28,000	Yes	3	No
30	46,000	Yes	6	Yes
51	46,000	No	4	Yes
32	80,000	Yes	2	No
45	50,000	No	4	No

Por otro lado para eliminar los sesgos podemos construir un modelo predictivo utilizando regresión lineal, aunque esto dependera de nuestro modelo y la cantidad de datos para costruirlo, por lo que para este caso puede que no sea tan viable.

```
models <- with(df_Computer_imputed, lm(Income ~ Age + Credit_rating))
combinado <- pool(models)
df_Computer_imputed_complete <- complete(df_Computer_imputed)
```

- Empleando alguna de las **técnicas revisadas en el capítulo**, realiza una **discretización** de los atributos **Age**, **Income** y **Credit_Rating**

Para este momento de realizar la discretización es preferible no tener valores perdidos, ya que buscamos convertir los datos cuantitativos en cualitativos, es decir variables continuas en categorías o rangos. Utilizaremos la tecnica de Análisis de Clústeres con K-Means, para agrupar los valores en grupos o clústeres discretos basados en la similitud entre observaciones.

Empezamos con la columna **Credit_rating**, para la cual definimos 3 cluteres y que dichos valores los agrege al data frame discretizado, para posteriormente se le asingen sus respectivos nombres:

```
# Extraer 'Credit_rating' en una matriz para el analisis de clusteres
data_to_cluster <- datos_completos$Credit_rating
# Aplicar K-Means para formar 3 clusteres
kmeans_model <- kmeans(data_to_cluster, centers = 3)
# Agregar los clusteres al dataframe original
df_discretized <- df_discretized %>%
  mutate(Credit_Cluster = kmeans_model$cluster)
# Asigna nombres descriptivos a los clusters
df_discretized <- df_discretized %>%
  mutate(Credit_Discretized = cut(Credit_Cluster,
  breaks = c(-Inf, 1, 2, 3, Inf),
  labels = c("Low", "Medium", "High", "No Cluster")))
```

Esto lo aplicamos de acuerdo a la cantidad de cluster-categorias que queremos, la forma de saber un cluter optimo puede ser con el Método del Codo que busca el punto de inflexión en la gráfica de la varianza explicada por el número de clusters. Selecciona el número de clusters donde la disminución de la varianza explicada comienza a aplanarse, pareciendo un "codo". Aplicamos estos comandos para las demás columnas **Age** y **Income**, considerando que podemos dar una discretizacion optima o podemos darla.

	Age	Income	Student	Credit_rating	Buy_Computer	Credit_Cluster	Credit_Discretized	Income_Cluster	Income_Discretized	Age_Cluster	Age_Discretized
1	21	60000	Yes	3	No	2	Medium	1	Low	1	Young
2	30	70000	No	5	No	3	High	1	Low	1	Young
3	38	70000	No	2	Yes	2	Medium	1	Low	2	Middle-aged
4	45	45000	Yes	3	Yes	2	Medium	2	High	3	Elderly
5	46	25000	No	2	Yes	2	Medium	2	High	3	Elderly
6	47	30000	Yes	6	No	1	Low	2	High	3	Elderly
7	39	28000	Yes	5	No	3	High	2	High	2	Middle-aged
8	29	48000	Yes	3	No	2	Medium	2	High	1	Young
9	50	75000	Yes	2	No	2	Medium	1	Low	3	Elderly
10	48	28000	Yes	3	No	2	Medium	2	High	3	Elderly
11	30	46000	Yes	6	Yes	1	Low	2	High	1	Young
12	51	46000	No	4	Yes	3	High	2	High	3	Elderly
13	32	80000	Yes	2	No	2	Medium	1	Low	1	Young
14	45	50000	No	4	No	3	High	2	High	3	Elderly

Figure 4: **discretización** de los atributos **Age**, **Income** y **Credit_Rating**

En resumen las categorías obtenidas mediante la discretización con K-Means para cada variable: 'Age': Tres clústeres identificados. Representan grupos de edades, 'Joven', 'Adulto' o 'Mayor'. 'Income': Dos clústeres formados. Reflejan diferentes niveles de ingresos„ como 'Bajo', y 'Alto'. 'Credit_rating': Tres clústeres obtenidos.Representan niveles de calificación crediticia, corresponden a 'Baja', 'media' y 'Alta'

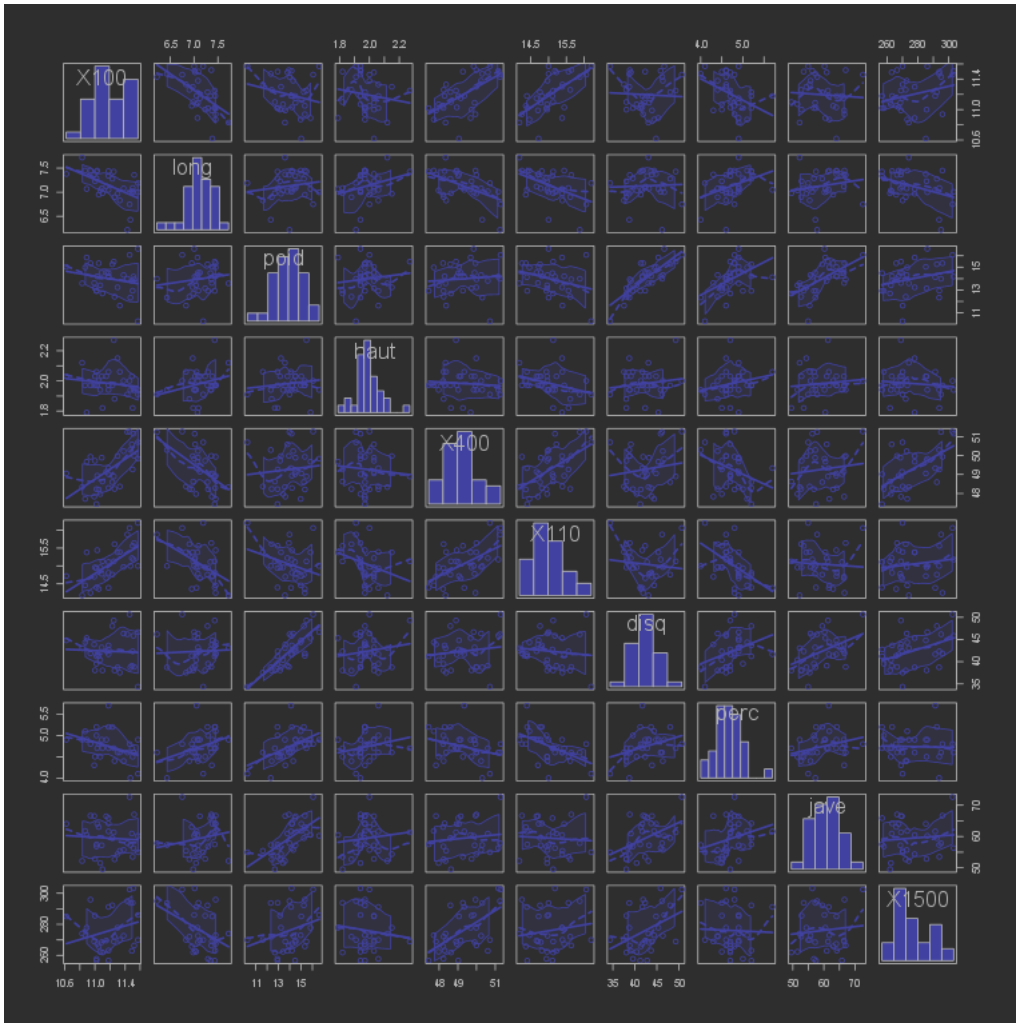
d. Toma el **dataset olympic** que se encuentra en el paquete ade4, investiga si es candidato para aplicar la técnica de PCA. Si es el caso, realiza el análisis correspondiente. Las variables que se encuentran en el dataset y que corresponde a eventos de un decatlón son: 100 metros (100), salto de longitud (long), lanzamiento de peso (poid), salto de altura (haut), 400 metros (400), obstáculos 110 metros (110), lanzamiento de disco (disq), salto de garrocha (perc), jabalina (jave) and 1500 metros (1500).

Para poder saber si el dataset es candidato para aplicar PCA, necesitamos analizar la composición de los datos, si hay datos extremos y si existe una fuerte correlación entre ellos.

Primero obtenemos un resumen del dataset :

```
> summary(decathlon)
      100             long            poid            haut            400             110             disq      
 Min.   :10.62   Min.   :6.220   Min.   :10.27   Min.   :1.790   Min.   :47.44   Min.   :14.18   Min.   :34.36  
 1st Qu.:11.02   1st Qu.:7.000   1st Qu.:13.15   1st Qu.:1.940   1st Qu.:48.34   1st Qu.:14.72   1st Qu.:39.08  
 Median :11.18   Median :7.090   Median :14.12   Median :1.970   Median :49.15   Median :15.00   Median :42.32  
 Mean   :11.20   Mean   :7.133   Mean   :13.98   Mean   :1.983   Mean   :49.28   Mean   :15.05   Mean   :42.35  
 3rd Qu.:11.43   3rd Qu.:7.370   3rd Qu.:14.97   3rd Qu.:2.030   3rd Qu.:49.98   3rd Qu.:15.38   3rd Qu.:44.80  
 Max.   :11.57   Max.   :7.720   Max.   :16.60   Max.   :2.270   Max.   :51.28   Max.   :16.20   Max.   :50.66  
      perc            jave            1500      
 Min.   :4.000   Min.   :49.52   Min.   :256.6  
 1st Qu.:4.600   1st Qu.:55.42   1st Qu.:266.4  
 Median :4.700   Median :59.48   Median :272.1  
 Mean   :4.739   Mean   :59.44   Mean   :276.0  
 3rd Qu.:4.900   3rd Qu.:64.00   3rd Qu.:286.0  
```

Como podemos observar, no hay datos extremos y todas las columnas corresponden a valores numéricos, ahora obtenemos el diagrama de dispersión correspondiente :



Y por ultimo obtenemos el índice de correlación de cada columna :

```
> round(cor(decathlon),2)
       100  long  poid  haut   400   110  disq  perc  jave  1500
100   1.00 -0.54 -0.21 -0.15  0.61  0.64 -0.05 -0.39 -0.06  0.26
long -0.54  1.00  0.14  0.27 -0.52 -0.48  0.04  0.35  0.18 -0.40
poid -0.21  0.14  1.00  0.12  0.09 -0.30  0.81  0.48  0.60  0.27
haut -0.15  0.27  0.12  1.00 -0.09 -0.31  0.15  0.21  0.12 -0.11
400   0.61 -0.52  0.09 -0.09  1.00  0.55  0.14 -0.32  0.12  0.59
110   0.64 -0.48 -0.30 -0.31  0.55  1.00 -0.11 -0.52 -0.06  0.14
disq -0.05  0.04  0.81  0.15  0.14 -0.11  1.00  0.34  0.44  0.40
perc -0.39  0.35  0.48  0.21 -0.32 -0.52  0.34  1.00  0.27 -0.03
jave -0.06  0.18  0.60  0.12  0.12 -0.06  0.44  0.27  1.00  0.10
1500  0.26 -0.40  0.27 -0.11  0.59  0.14  0.40 -0.03  0.10  1.00
```

Por el diagrama de dispersión y los índices de correlación podemos observar que sólo existe una correlación media entre la mayoría de columnas, siendo el lanzamiento de peso y el lanzamiento de disco las únicas columnas con una correlación fuerte, por lo que podemos concluir que el dataset no es un candidato

apto para aplicar la técnica de PCA en su forma actual.

e. Toma el **dataset olympic** que se encuentra en el paquete ade4, realiza una selección de características, probando algunos de los métodos descritos y compara las distintas elecciones que realiza cada uno de ellos. Normaliza las variables, utilizando la normalización **min-max** de forma que los datos se ajusten al intervalo [0,1]. Muestra una matriz de diagramas de dispersión para contrastar los resultados antes y después de normalizar.

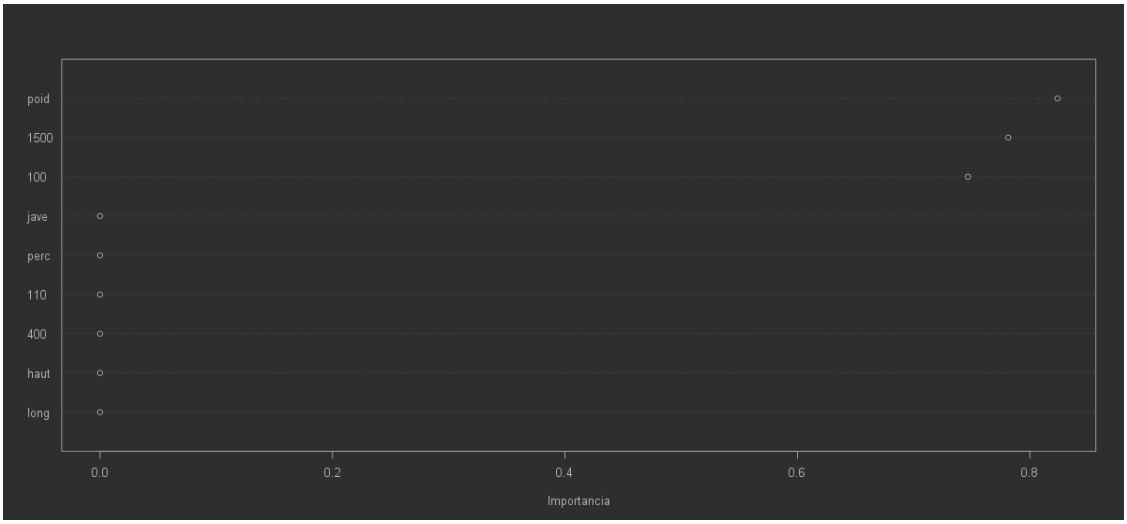
Utilizamos la variable disq y cuatro métodos distintos, *chi.squared, information.gain, gain.ratio* y *oneR*, los resultados fueron muy similares a traves de los cuatro metodos :

```
> chi.squared(disq~.,data=decathlon)
     attr_importance
100        0.7467744
long       0.0000000
poid       0.8237545
haut       0.0000000
400        0.0000000
110        0.0000000
perc       0.0000000
jave       0.0000000
1500       0.7815244
> information.gain(disq~.,data=decathlon)
     attr_importance
100        0.3593793
long       0.0000000
poid       0.4234930
haut       0.0000000
400        0.0000000
110        0.0000000
perc       0.0000000
jave       0.0000000
1500       0.3300437
```

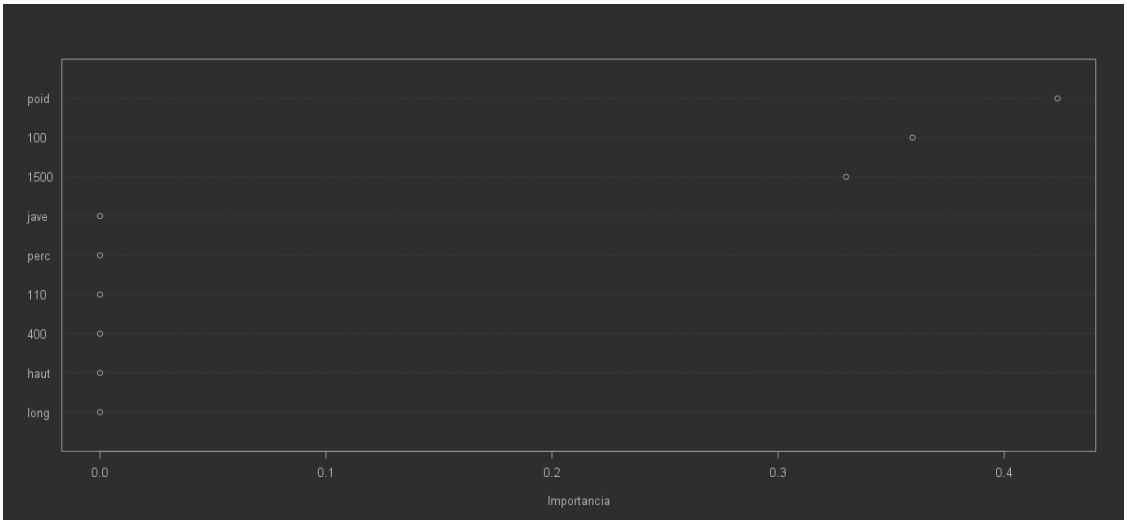
```
> gain.ratio(disq~.,data=decathlon)
     attr_importance
100        0.5215886
long       0.0000000
poid       0.6653317
haut       0.0000000
400        0.0000000
110        0.0000000
perc       0.0000000
jave       0.0000000
1500       0.5632600
> oneR(disq~.,data=decathlon)
     attr_importance
100        0.7575758
long       0.5757576
poid       0.7878788
haut       0.5757576
400        0.5757576
110        0.5757576
perc       0.5757576
jave       0.5757576
1500       0.7575758
```

Para representar mejor los resultados hicimos las siguientes graficas :

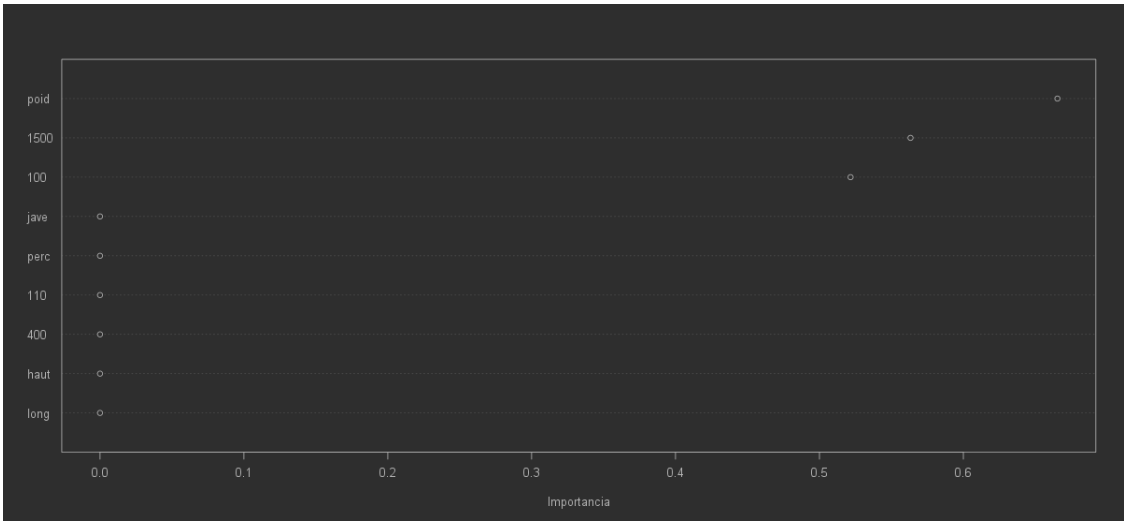
- ***chi.squared***



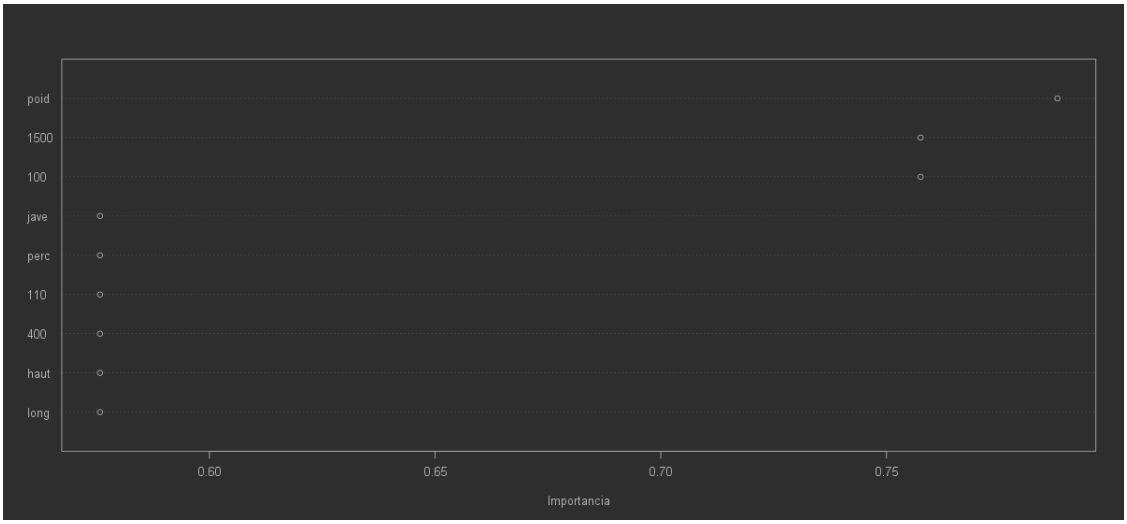
- ***information.gain***



- ***gain.ratio***



- ***oneR***



Para los diagramas de dispersión antes de normalizar obtuvimos :

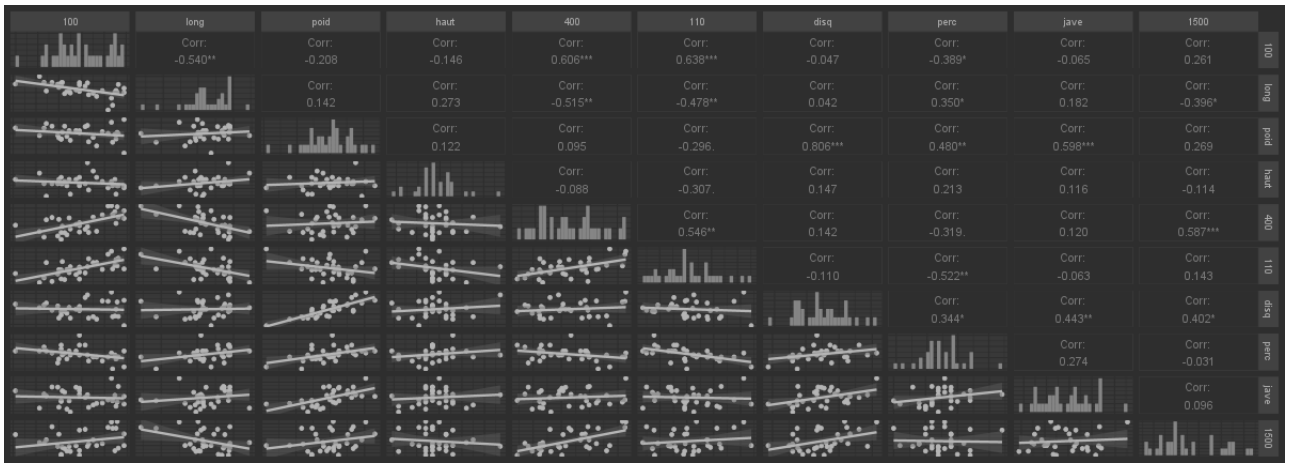


Diagrama después de normalizar :