

**APLICACIÓN DE TAREAS RELACIONADAS CON EL PROCESO DE
MINERÍA DE DATOS BAJO LA METODOLOGÍA CRISP-DM**

PRESENTADO POR:

ANIK VALERIA HERNÁNDEZ RONCANCIO

PRESENTADO A:

WILMER EDICSON GARZÓN ALFONSO

MINERÍA DE DATOS

DECANATURA DE INGENIERÍA INDUSTRIAL

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

BOGOTÁ D.C.

2018

TABLA DE CONTENIDO

INTRODUCCIÓN	4
OBJETIVOS	5
Objetivo general	5
Objetivos específicos	5
1. Descripción del problema	6
2. Descripción del conjunto de datos	7
3. Descripción y justificación imputación de valores faltantes	9
4. Descripción de las técnicas de preprocesamiento aplicadas	12
5. Predicción de la clase	14
5.1. Árbol de decisión	14
5.2. Vecino más cercano	17
5.3. Modelo elegido	20
6. Clustering	20
CONCLUSIONES	23
BIBLIOGRAFÍA	24

TABLA DE ILUSTRACIONES

Ilustración 1 Diagrama de cajas y bigotes variable "erl"	7
Ilustración 2 Diagrama de cajas y bigotes variable "pox"	7
Ilustración 3 Información estadística del conjunto de datos original.	8
Ilustración 4 Conjunto de datos con 10% de valores faltantes	9
Ilustración 5 Comparación del valor del MSE para los diferentes valores de k.	9
Ilustración 6 Comparación de los vectores de error para las técnicas del vecino más cercano y la media.	10
Ilustración 7 Comparación de los vectores de error para la imputación por media y moda.....	10
Ilustración 8 Información estadística del conjunto de datos después de la imputación.	11
Ilustración 9 Diagrama de cajas y bigotes antes de normalizar con Z-Score.....	13
Ilustración 10 Diagrama de cajas y bigotes después de normalizar con Z-Score.	13
Ilustración 11 Matriz de confusión árbol de decisión datos discretizados.....	14
Ilustración 12 Árbol de decisión conjunto de datos discretizado.	15
Ilustración 13 Matriz de confusión árbol de decisión datos sin discretizar.....	16
Ilustración 14 Árbol de decisión conjunto de datos sin discretizar.	17
Ilustración 15 Precisión y matriz de confusión kNN con k=7.....	18
Ilustración 16 Precisión y matriz de confusión kNN con k=21.	19
Ilustración 17 Variación de la precisión con kNN para diferentes valores de k según la literatura encontrada.	19
Ilustración 18 Resumen de resultados de las técnicas de clasificación.	20
Ilustración 19 Matriz de correlación.....	21
Ilustración 20 Matriz de diagramas de dispersión k=2.	22
Ilustración 21 Matriz de diagramas de dispersión k=3.	22
Ilustración 22 Matriz de diagramas de dispersión k=4.	23

INTRODUCCIÓN

La minería de datos se puede definir como el proceso de encontrar patrones y tendencias que no se conocían previamente en conjuntos de datos y haciendo uso de esa información, construir modelos de predicción. La minería de datos puede mejorar los procesos de toma de decisiones descubriendo estos patrones y tendencias en grandes cantidades de datos.

“Cross Industry Standard Process for Data Mining” o CRISP-DM propone la siguiente metodología para la minería de datos: comprensión del negocio, estudio y comprensión de los datos, preparación de los datos, modelado, evaluación de los datos y despliegue. La comprensión del negocio es una etapa crucial porque identifica los objetivos del negocio y así los criterios de éxito de los proyectos de minería de datos que se van a realizar; la comprensión y preparación de los datos que en otras palabras es el muestreo y transformación de los datos son procesos antecedentes esenciales para la construcción del modelo; la etapa del modelado es en sí el análisis de los datos actuales; la evaluación de los datos habilita la comparación de modelos y resultados y finalmente, el despliegue se refiere a la implementación y operacionalización de los modelos de minería de datos.

OBJETIVOS

Objetivo general

El objetivo general de este trabajo consiste en extraer información del conjunto de datos llamado “Protein Localization Sites”¹ y transformarla en una estructura comprensible para su uso posterior.

Objetivos específicos

- Transformar la información contenida en el conjunto de datos “Protein Localization Sites” mediante técnicas de preprocesamiento.
- Aplicarle técnicas supervisadas de clasificación a la información contenida en el conjunto de datos “Protein Localization Sites”.
- Clasificar cada proteína dentro de los 9 diferentes componentes celulares de la levadura con el fin de que a través de un análisis predictivo sea posible identificar el sitio de localización de las proteínas contenidas en la levadura a partir de los valores de los atributos presentes.
- Realizar un proceso de clasificación no supervisada de “clustering” agrupando instancias de modo que estas pertenezcan al mismo “cluster” y las que no tengan características similares pertenezcan a clusters diferentes.
- Encontrar el número óptimo de clusters en el que la clasificación de los datos sea clara, mediante intentos de prueba y error.

¹Kenta Nakai. Institute of Molecular and Cellular Biology. <https://archive.ics.uci.edu/ml/machine-learning-databases/yeast/yeast.names>

1. Descripción del problema

El conjunto de datos que se trabajará en este caso tiene como título “Protein Localization Sites”, el objetivo del conjunto de datos es predecir la localización de las proteínas dentro de la célula de levadura. Consta de 9 atributos de los cuales 8 son predictivas y 1 es de nombre, 1484 instancias, no contiene valores faltantes y la variable categórica tiene 10 clases.

El contenido de cada uno de los atributos es el siguiente:

- Sequence Name: Número de acceso para la base de datos SWISS-PROT².
- mcg: Método de McGeoch para el reconocimiento de secuencias de señales.
- gvh: Método de von Heijne para el reconocimiento de secuencias de señales.
- alm: Puntuación del programa de predicción ALOM³ para la región que abarca la membrana.
- mit: Puntuación del análisis discriminante del contenido de aminoácidos de la región N-terminal (también conocida como amino-terminal, NH₂-terminal, extremo amina) de las proteínas mitocondriales y no mitocondriales.
- erl: Presencia de la cadena HDEL. Atributo binario.
- pox: Señal de orientación del peroxisoma en el extremo C-terminal.
- vac: Puntuación del análisis discriminante del contenido de aminoácidos de las proteínas vacuolar y extracelular.
- nuc: Puntuación del análisis discriminante de las señales de localización nuclear de proteínas nucleares y no nucleares.

La clase es el sitio de localización, los detalles de cada clase son los siguientes:

- CYT (citoplasmático o citoesquelético)
- NUC (nuclear)
- MIT (mitocondrial)
- ME3 (proteína de membrana, sin señal N-terminal)
- ME2 (proteína de membrana, señal no dividida)
- ME1 (proteína de membrana, señal dividida)

² SWISS-PROT es una base de datos de secuencias de proteínas curada que se esfuerza por proporcionar un alto nivel de anotación, un nivel mínimo de redundancia y alto nivel de integración con otras bases de datos.

³ Método computacional que detecta posibles segmentos transmembrana.

- EXC (extracelular)
- VAC (vacuolar)
- POX (peroxisoma)
- ERL (luz del retículo endoplasmático)

2. Descripción del conjunto de datos

Para la preparación de los datos es importante analizar la información contenida en el conjunto de datos original y cuáles de estos aportan o no información al ejercicio, con el fin de que los resultados obtenidos al final tengan la mayor precisión posible.

Se decide remover el atributo “Sequence Names” ya que al ser una variable de nombre no afecta el comportamiento del conjunto de datos. Además, los atributos “erl” y “pox” no varían en su contenido, esto se puede observar gráficamente por medio de los siguientes diagramas de cajas y bigotes:

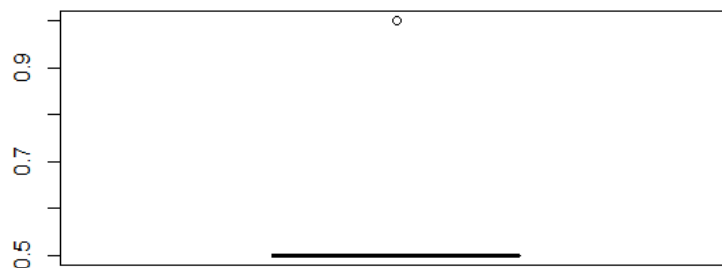


Ilustración 1 Diagrama de cajas y bigotes variable "erl"

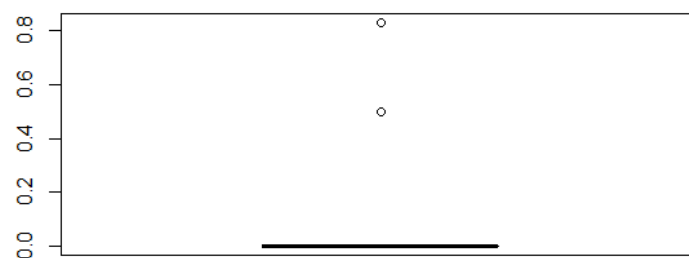


Ilustración 2 Diagrama de cajas y bigotes variable "pox"

En la ilustración 1 se muestra que el atributo “erl” solamente una toma un valor diferente al de las demás instancias, a pesar de ser binario toma el mismo valor la mayoría de las veces. En la ilustración 2, el atributo no es de tipo binario y únicamente cambia su valor en dos instancias. Por lo tanto, se toma la decisión

de remover también estos dos atributos, ya que no están aportando información al ejercicio.

Luego de esto, se muestra un resumen estadístico de cada uno de los atributos del conjunto de datos original con el fin de conocer su comportamiento antes de realizar las técnicas de preprocesamiento requeridas:

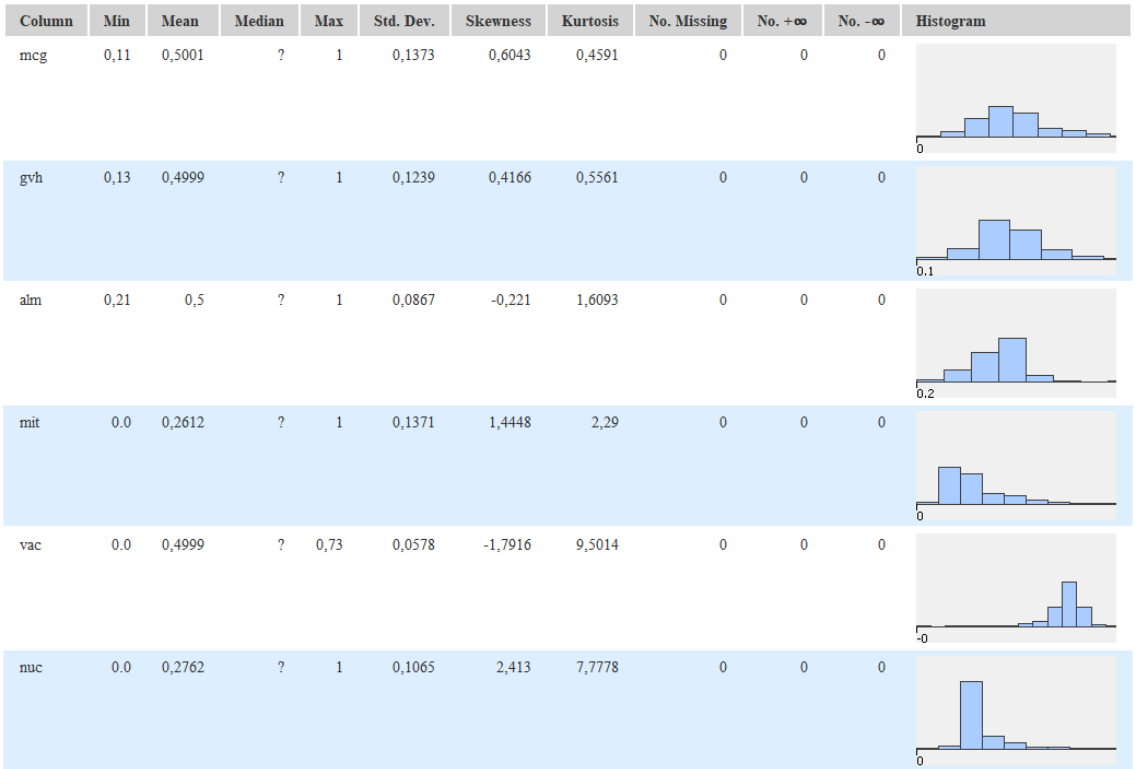


Ilustración 3 Información estadística del conjunto de datos original.

En la imagen se observa la media, la desviación estándar, el sesgo y otra información relevante que presenta originalmente cada una de las variables del conjunto, estas medidas estadísticas son de utilidad para conocer el comportamiento de los datos antes de ser sometidos a las técnicas de preprocesamiento. En la imagen no se muestra el atributo de la clase ya que al ser categórico no es posible presentar información estadística de este.

Luego de conocer el comportamiento original se introduce aleatoriamente un 10% de valores faltantes al conjunto de datos, para esto se muestra gráficamente a continuación los atributos del nuevo conjunto de datos con sus respectivos valores faltantes de color azul claro:

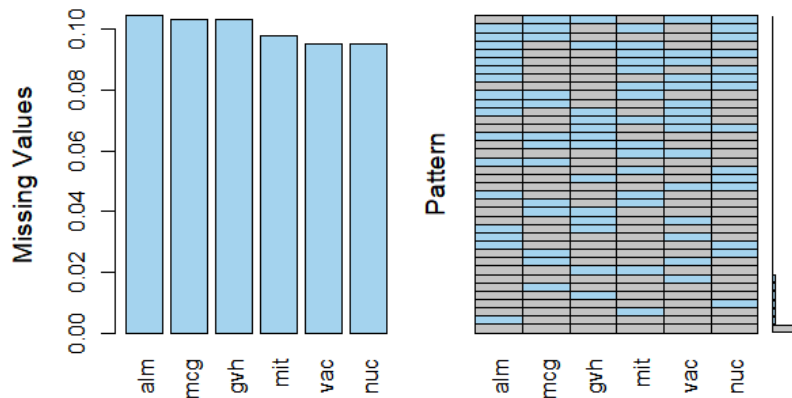


Ilustración 4 Conjunto de datos con 10% de valores faltantes

3. Descripción y justificación imputación de valores faltantes

Para aplicar las técnicas de imputación fue necesario remover la columna correspondiente a la variable de la clase. Teniendo en cuenta que no se encontró literatura relacionada a la imputación de valores faltantes para el conjunto de datos “Protein Localization Sites”, se realizaron pruebas con diferentes valores de k para la imputación con el método del vecino más cercano, estos valores fueron $k=7$, $k=8$ y $k=9$ y para las imputaciones resultantes de estos tres valores se calculó el error cuadrático medio MSE y se compararon los vectores que contienen la información del MSE, de los 3 valores evaluados para k , el que tiene menor error es $k=8$. El resultado de las comparaciones se obtuvo de la siguiente forma, donde se compara si el vector de errores para todos los atributos con cada valor de k es menor a los otros:

```
> table(Error7<Error8)#Verificar si el error de k=7 es menor a k=8
FALSE TRUE
   5    1
> table(Error7<Error9)#Verificar si el error de k=7 es menor a k=9
FALSE TRUE
   3    3
> table(Error8<Error7)#Verificar si el error de k=8 es menor a k=7
FALSE TRUE
   1    5
> table(Error8<Error9)#Verificar si el error de k=8 es menor a k=9
TRUE
   6
> table(Error9<Error7)#Verificar si el error de k=9 es menor a k=7
FALSE TRUE
   3    3
> table(Error9<Error8)#Verificar si el error de k=9 es menor a k=8
FALSE
   6
```

Ilustración 5 Comparación del valor del MSE para los diferentes valores de k .

En la ilustración se observa que de los 6 atributos que tiene el conjunto de datos, en 5 de ellos el MSE es menor para $k=8$ en comparación con $k=7$, para el caso de $k=9$ el MSE en todos los atributos del conjunto de datos es mayor que para $k=8$, por esta razón se decide trabajar con un valor k de 8 para la imputación de valores faltantes.

Luego se hicieron imputaciones con la media, la mediana y la moda y partiendo de estas técnicas y el resultado de la imputación por el vecino más cercano se calculó nuevamente el MSE y se compararon los vectores de error y la técnica con menor error respecto al conjunto de datos original fue la del vecino más cercano con $k=8$. Estos resultados se verificaron comparando los vectores de error de cada técnica, las ilustraciones 5 y 6 muestran los resultados de estas comparaciones:

```
> table(Errork8<ErrorMean)#Verificar si el error de k=8 es menor a mean
TRUE
6
> table(Errork8<ErrorMedian)#Verificar si el error de k=8 es menor a median
TRUE
6
> table(Errork8<ErrorMode)#Verificar si el error de k=8 es menor a mode
TRUE
6
> table(ErrorMean<Errork8)#Verificar si el error de mean es menor a k=8
FALSE
6
> table(ErrorMean<ErrorMedian)#Verificar si el error de mean es menor a median
FALSE TRUE
4 2
> table(ErrorMean<ErrorMode)#Verificar si el error de mean es menor a mode
FALSE TRUE
2 4
```

Ilustración 6 Comparación de los vectores de error para las técnicas del vecino más cercano y la media.

```
> table(ErrorMedian<Errork8)#Verificar si el error de median es menor a k=8
FALSE
6
> table(ErrorMedian<ErrorMean)#Verificar si el error de median es menor a mean
FALSE TRUE
2 4
> table(ErrorMedian<ErrorMode)#Verificar si el error de median menor a mode
FALSE TRUE
2 4
> table(ErrorMode<Errork8)#Verificar si el error de mode es menor a k=8
FALSE
6
> table(ErrorMode<ErrorMean)#Verificar si el error de mode es menor a mean
FALSE TRUE
4 2
> table(ErrorMode<ErrorMedian)#Verificar si el error de mode es menor a median
FALSE TRUE
4 2
```

Ilustración 7 Comparación de los vectores de error para la imputación por media y moda.

En la ilustración 6 se puede ver que el MSE en la imputación del vecino más cercano es menor para todos los atributos en comparación con las otras técnicas, por lo que se decide trabajar con el conjunto de datos resultante de la imputación del vecino más cercano con un valor de $k=8$.

Luego de determinar la técnica de imputación con menor error para el conjunto de datos, se muestra nuevamente un resumen estadístico de cada una de las variables que se tuvieron en cuenta durante esta etapa de preprocesamiento del conjunto de datos, con el fin de comparar el comportamiento de los datos antes y después de la imputación de valores faltantes.

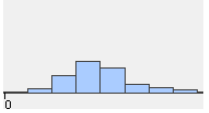
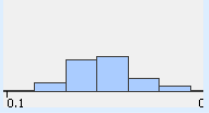
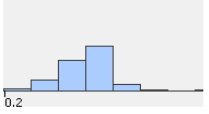
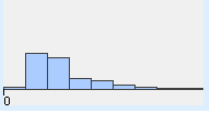
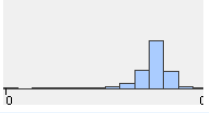
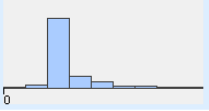
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
mcg	0,11	0,5004	?	1	0,134	0,6318	0,6349	0	0	0	
gvh	0,13	0,5007	?	0,94	0,1212	0,404	0,5789	0	0	0	
alm	0,21	0,4996	?	1	0,0829	-0,2369	2,0684	0	0	0	
mit	0.0	0,2606	?	1	0,1328	1,4917	2,6202	0	0	0	
vac	0.0	0,5007	?	0,73	0,056	-1,9343	10,9645	0	0	0	
nuc	0.0	0,2734	?	1	0,1014	2,4375	7,7865	0	0	0	

Ilustración 8 Información estadística del conjunto de datos después de la imputación.

Teniendo en cuenta los resultados de la ilustración 3 y los de la ilustración 8, se puede concluir que las medidas de tendencia central del conjunto de datos original no varían drásticamente si se comparan con los del conjunto de datos resultante de la imputación por el vecino más cercano con $k=8$, esto era precisamente lo que se quería lograr buscando la técnica de imputación con menor error para el conjunto de datos, que las medidas de tendencia central no se vieran fuertemente afectadas.

4. Descripción de las técnicas de preprocesamiento aplicadas

Para la clasificación con árbol de decisión se discretiza el conjunto de datos original ya que la mayoría de los atributos son numéricos y las técnicas de clasificación serían bastante extensas teniendo en cuenta que el proceso de clasificación busca asignar cada atributo “x” a una clase específica “y”, y, por lo tanto, independientemente de la técnica que se esté ejecutando se tendrían que evaluar una a una las instancias del conjunto de datos para llevar a cabo una clasificación correcta. El proceso de discretización se realizó en RStudio, haciendo uso de la librería “arules” las variables numéricas fueron convertidas en variables categóricas donde cada una tiene tres clases, cada clase es un intervalo en el que están contenidos los datos del conjunto original. Esta librería tiene una función para discretizar un conjunto de datos completo, teniendo en cuenta la distribución de los datos que se observa en la ilustración 3 se tomó la decisión de utilizar esta función ya que determina por defecto el número de “breaks” en los que se dividirá cada atributo.

Para la clasificación con el método del vecino más cercano se normalizaron los datos en un intervalo entre [0,1] siguiendo con los procedimientos que realizaron por (Horton & Nakai, 1997) [1], quienes utilizaron validación cruzada para encontrar una precisión óptima de clasificación y establecen parámetros para la clasificación del conjunto de datos de Yeast, el mismo que se está trabajando en este proyecto, estos parámetros serán citados más adelante en la clasificación del vecino más cercano.

Para la clasificación no supervisada con clustering se normaliza el conjunto de datos teniendo en cuenta que se tienen valores atípicos, esta decisión se toma en base en la ilustración 9 la cual muestra un diagrama de cajas y bigotes. La normalización esta vez se hace con el método Z-Score, el cual normaliza los datos tomando como base la media y la desviación estándar de cada uno de los datos, ya que no se encontró literatura referente a la preparación de este tipo de conjunto de datos para la clasificación con clustering.

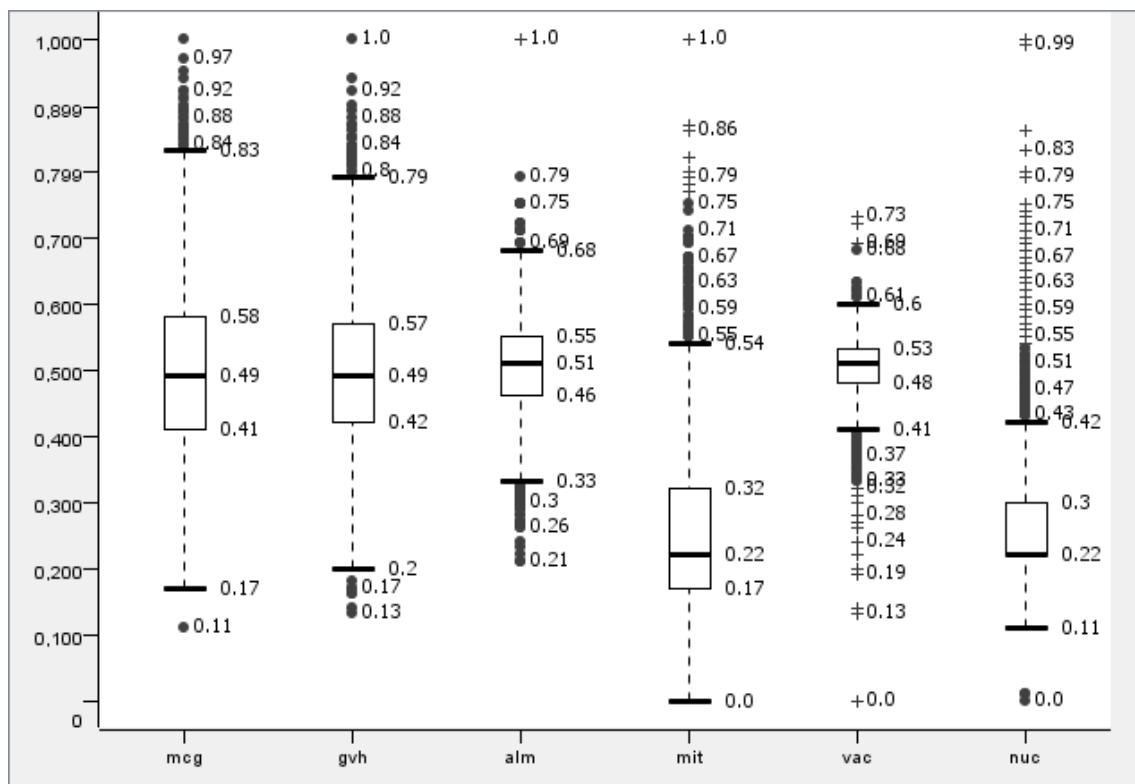


Ilustración 9 Diagrama de cajas y bigotes antes de normalizar con Z-Score.

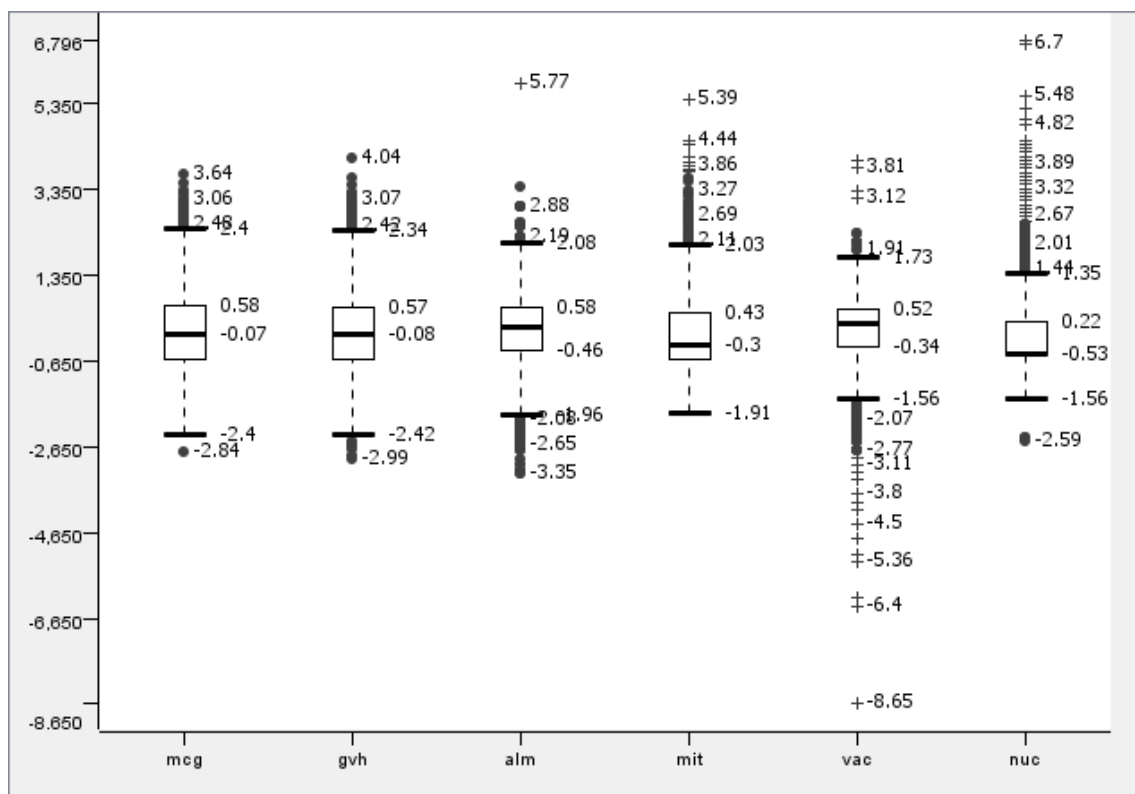


Ilustración 10 Diagrama de cajas y bigotes después de normalizar con Z-Score.

0En la ilustración 10 se observa el comportamiento de los datos después de normalizar.

5. Predicción de la clase

5.1. Árbol de decisión

Un árbol de decisiones es un sistema de soporte de decisiones que utiliza un gráfico similar a un árbol y sus posibles efectos secundarios, incluidos los resultados de eventos de azar, los costos de recursos y la utilidad. Un árbol de decisión, se usa para obtener una función de clasificación que estima el valor de un atributo dependiente dados los valores de los atributos independientes. Este es un tipo de como clasificación supervisada porque se tienen el atributo de la clase y los atributos de clasificación.

Los árboles de decisión son uno de los enfoques más poderosos en el descubrimiento de conocimiento y la minería de datos. Estos incluyen la tecnología de investigación de grandes volúmenes de datos complejos para descubrir patrones útiles. Los árboles de decisión ofrecen varios beneficios a la minería de datos ya que son fáciles de entender por el usuario final, pueden manejar una variedad de datos de entrada y tienen un alto rendimiento con un escaso número de esfuerzos.

El árbol de decisión se hizo en KNIME y para la clasificación se particionó el conjunto de datos con un 70% de datos de entrenamiento y un 30% de datos de prueba. Como se mencionó antes, para la clasificación con árbol de decisión se discretizaron los datos con el fin de obtener una cantidad moderada de niveles en el árbol teniendo en cuenta el tipo de datos de entrada con los que se cuenta.

Localizatio...	MIT	NUC	CYT	ME1	ME3	EXC	ME2	VAC	POX	ERL
MIT	36	16	8	11	2	0	0	0	0	0
NUC	9	69	42	2	7	0	0	0	0	0
CYT	9	31	85	7	7	0	0	0	0	0
ME1	0	0	2	11	0	0	0	0	0	0
ME3	2	11	3	3	30	0	0	0	0	0
EXC	2	2	3	4	0	0	0	0	0	0
ME2	0	5	1	7	2	0	0	0	0	0
VAC	2	0	6	0	1	0	0	0	0	0
POX	1	0	4	1	0	0	0	0	0	0
ERL	1	0	1	0	0	0	0	0	0	0
Correct classified: 231										
Wrong classified: 215										
Accuracy: 51,794 %										
Error: 48,206 %										
Cohen's kappa (κ) 0,371										

Ilustración 11 Matriz de confusión árbol de decisión datos discretizados.

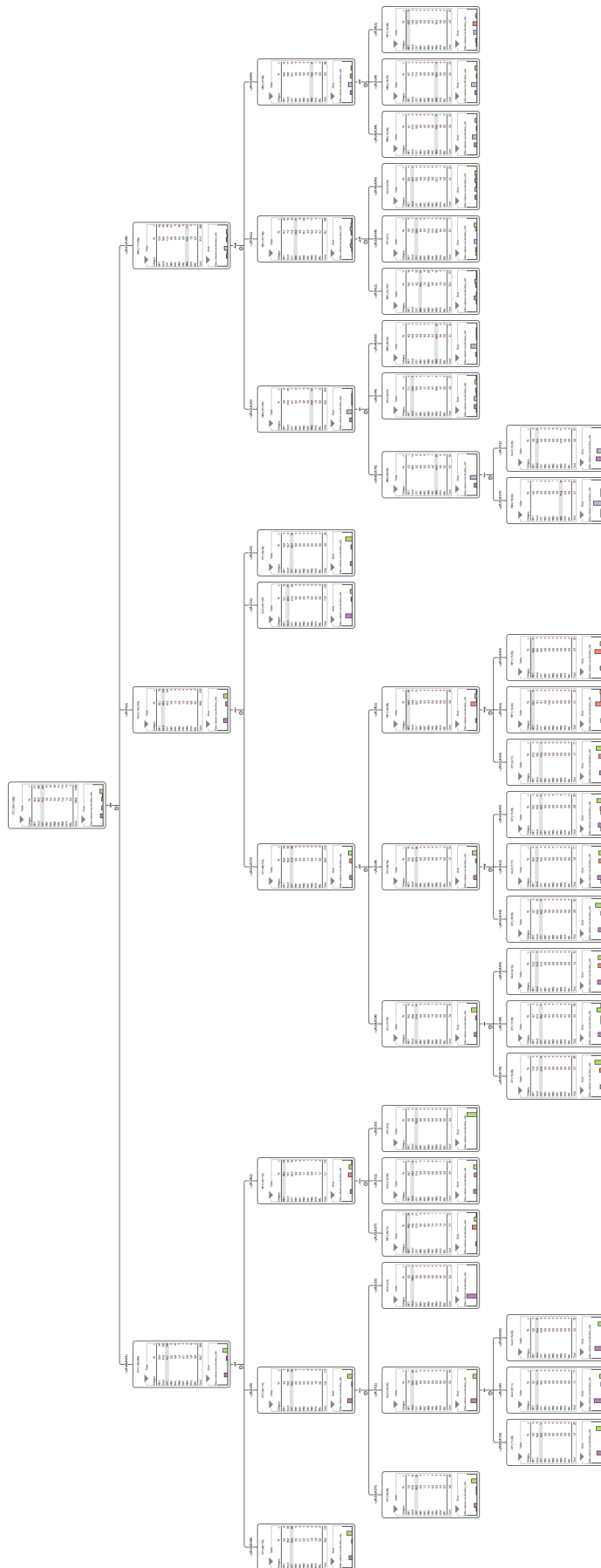


Ilustración 12 Árbol de decisión conjunto de datos discretizado.

En este tipo de clasificación se observó que la entrada de número mínimo de registros del nodo “Decision Tree Learner” afecta drásticamente la precisión de la clasificación, siguiendo con esto el valor de esta entrada con el que se alcanzó el valor de 20, y con este valor se logró una precisión de 51,4% clasificando correctamente 231 datos de un total de 446. En las ilustraciones xx y xx se muestra la matriz de confusión y el árbol de decisión respectivamente.

Luego de observar el porcentaje de precisión obtenido se decide realizar el mismo procedimiento, pero con el conjunto de datos original, es decir sin discretizar para observar si este porcentaje mejora. Se tuvo en cuenta el mismo porcentaje de partición y valor del número mínimo de registros con el fin de hacer una comparación bajo los mismos parámetros.

Los resultados obtenidos con los datos sin discretizar tienen bastante mejoría en comparación con los que se obtuvieron con los datos discretizados, se consiguió una precisión de 58,2% lo que corresponde a una clasificación correcta de 260 datos de un total 446 datos de prueba. La ilustración 11 muestra la matriz de confusión resultante.

Localizatio...	MIT	NUC	CYT	ME1	EXC	ME2	ME3	VAC	POX	ERL
MIT	39	6	20	0	1	1	6	0	0	0
NUC	2	71	52	0	0	0	4	0	0	0
CYT	9	42	87	0	1	0	0	0	0	0
ME1	1	0	0	10	2	0	0	0	0	0
EXC	2	2	2	1	4	0	0	0	0	0
ME2	1	5	2	0	0	5	2	0	0	0
ME3	3	2	0	0	0	0	44	0	0	0
VAC	1	1	4	0	0	1	2	0	0	0
POX	1	1	3	0	0	0	1	0	0	0
ERL	0	0	0	0	1	1	0	0	0	0

Correct classified: 260

Wrong classified: 186

Accuracy: 58,296 %

Error: 41,704 %

Cohen's kappa (κ) 0,45

Ilustración 13 Matriz de confusión árbol de decisión datos sin discretizar.

El árbol de decisión obtenido con los datos sin discretizar se despliega en 11 niveles, un número de niveles moderado teniendo en cuenta el tipo de atributos con el que se está trabajando ya que son numéricos y podrían haber resultado en un árbol con mayor cantidad de hojas.

Si se compara con el número de hojas que tiene el árbol de decisión de los datos discretizados no es muy grande y en cambio, la precisión si mejora considerablemente por lo que se decide trabajar con datos sin discretizar con

el fin de obtener una clasificación más precisa en comparación con el conjunto de datos originales.

La ilustración 12 muestra el árbol de decisión resultante del conjunto de datos original, es decir sin discretizar.

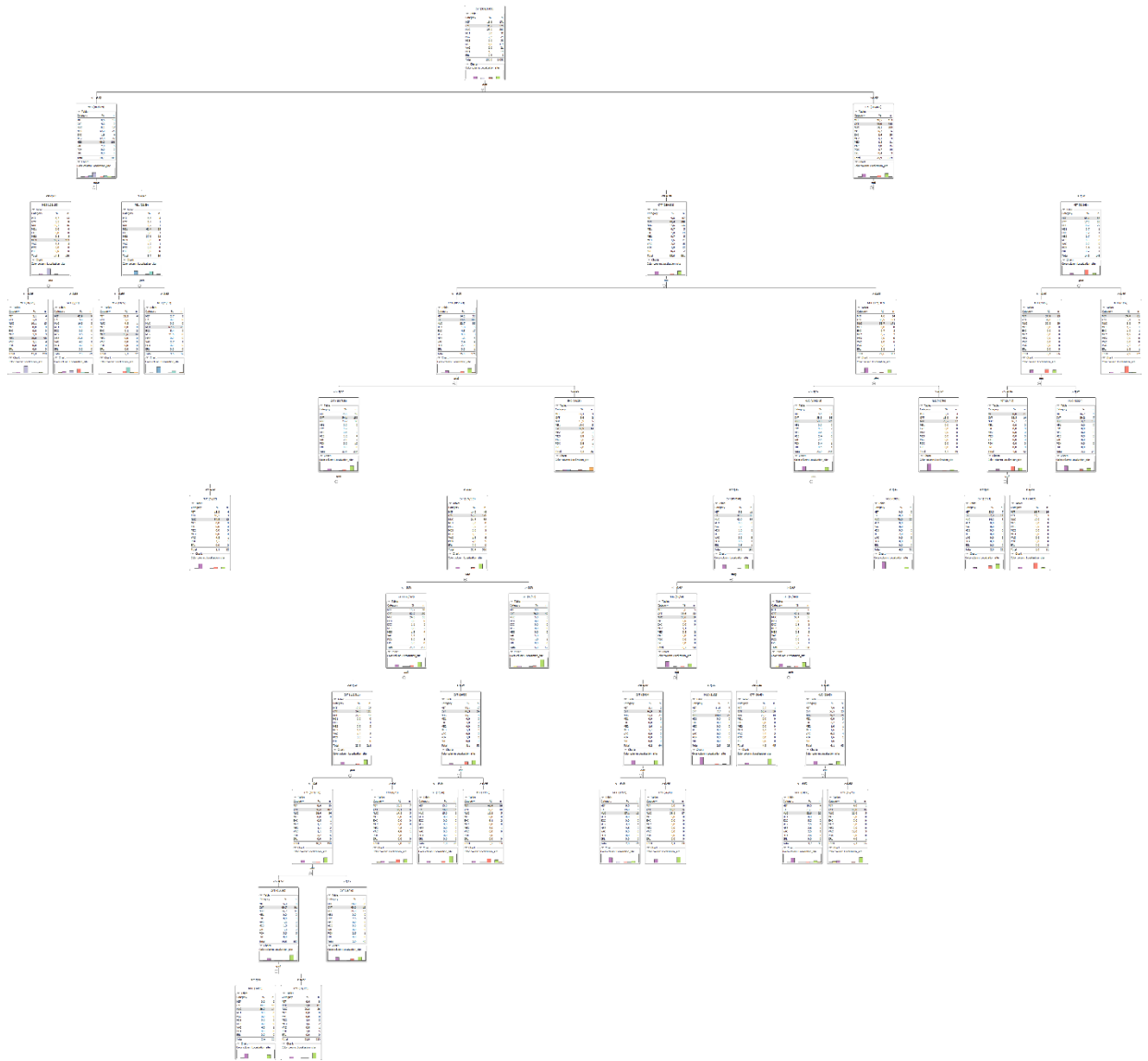


Ilustración 14 Árbol de decisión conjunto de datos sin discretizar.

5.2. Vecino más cercano

Para la clasificación con el vecino más cercano se precisaban de dos valores diferentes para el valor de k . Por lo que se buscó literatura existente referente a la clasificación de datos con el método del vecino más cercano, esperando poder encontrar un soporte para la elección de estos dos valores. (Horton &

Nakai, 1997) [1] estudiaron y compararon métodos de clasificación para los conjuntos de datos E.coli⁴ y Yeast, incluyendo la del vecino más cercano.

Según (Horton & Nakai, 1997) [1] para la precisión de la estimación utilizaron valores de k para los conjuntos de datos E.coli y Yeast 7 y 21 respectivamente, ellos determinaron estos valores haciendo validación cruzada en cada partición de entrenamiento y tomando el mejor valor global y encontraron que para el conjunto de datos Yeast, la mayor precisión es alcanzada con valores de k entre 21 y 25 y para el caso del conjunto de datos E.coli los valores de que alcanzan una precisión mayor están entre 5 y 7. Teniendo en cuenta esto, se decide que los valores con los que se va a trabajar en este ejercicio son k=7 y k=21 con el fin de demostrar lo que los autores de la literatura exponen.

Esta clasificación se hizo en RStudio y como se mencionó antes, los datos fueron normalizados. Los resultados obtenidos para cada uno de los valores de k se presentan a continuación:

```
> sum(test.site==knn.7)
[1] 236
> 100*sum(test.site==knn.7)/446 #Porcentaje de clasificacion correcta k=7
[1] 52.9148
> table(knn.7,test.site)#matriz de confusión k=7
      test.site
knn.7 CYT  ERL  EXC  ME1  ME2  ME3  MIT  NUC  POX  VAC
CYT   72   0   1   0   2   3  16  40   1   0
ERL   0   0   0   0   0   0   0   0   0   0
EXC   0   2   5   0   1   0   0   0   0   0
ME1   0   0   4  10   3   0   1   0   0   0
ME2   0   0   0   0   7   0   2   0   0   0
ME3   6   0   0   0   1  39   5   5   0   1
MIT  17   0   0   0   1   1  49  17   0   0
NUC  55   0   0   0   0   5   9  54   4   3
POX   0   0   0   0   0   0   0   0   0   0
VAC   2   0   0   0   0   0   0   1   1   0
```

Ilustración 15 Precisión y matriz de confusión kNN con k=7.

Se clasificaron correctamente 236 datos de 446, lo que corresponde a un 52,9% de precisión en la clasificación de los datos como se muestra en la ilustración 13 con un k=7.

⁴ Kenta Nakai, Institute of Molecular and Cellular Biology. Osaka.

Los resultados obtenidos con un valor de $k=21$ se muestran en la ilustración 14. En esta se muestra que se clasificaron correctamente 257 datos, obteniendo una precisión de 57,6%.

```
> sum(test.site==knn.21)
[1] 257
> 100*sum(test.site==knn.21)/446 #Porcentaje de clasificacion correcta
[1] 57.62332
> table(knn.21,test.site)#matriz de confusión k=21
      test.site
knn.21 CYT  ERL  EXC  ME1  ME2  ME3  MIT  NUC  POX  VAC
CYT    91    0    1    0    3    5   20   38    4    1
ERL     0    0    0    0    0    0    0    0    0    0
EXC     1    2    7    1    3    0    0    0    0    0
ME1     0    0    2    9    4    0    3    0    0    0
ME2     0    0    0    0    4    0    3    0    0    0
ME3     6    0    0    0    1   37    4    5    0    1
MIT     9    0    0    0    0    1   47   12    1    0
NUC    45    0    0    0    0    5    5   62    1    2
POX     0    0    0    0    0    0    0    0    0    0
VAC     0    0    0    0    0    0    0    0    0    0
```

Ilustración 16 Precisión y matriz de confusión kNN con $k=21$.

De esto se puede concluir que los autores aciertan en que se alcanza una mayor precisión con un valor de $k=21$ y que la precisión de la clasificación es aproximadamente 60%. Sin embargo, en la literatura se encuentra una gráfica en la que se muestra la precisión alcanzada con diferentes valores de k y según eso para un valor de $k=7$ sería más o menos 57% y no 52%, la gráfica presentada por los autores se muestra a continuación:

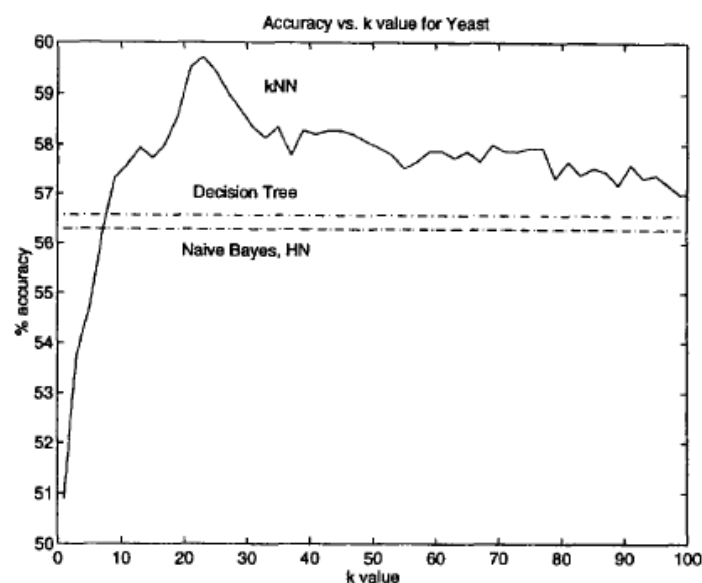


Ilustración 17 Variación de la precisión con kNN para diferentes valores de k según la literatura encontrada.

5.3. Modelo elegido

Después de analizar los resultados de cada una de las técnicas utilizadas anteriormente se hace una tabla para comparar fácilmente, y elegir el modelo que mejor se ajusta a las necesidades del ejercicio. En la ilustración x se muestra un resumen de cada una de las técnicas y sus resultados de clasificación.

Técnica de clasificación	Cantidad de instancias clasificadas correctamente	Porcentaje de precisión
Árbol de decisión	260	58,296%
Vecino más cercano "k=7"	236	52,914%
Vecino más cercano "k=21"	257	57,623%

Ilustración 18 Resumen de resultados de las técnicas de clasificación.

Teniendo en cuenta los resultados y buscando que la precisión de la clasificación de los datos sea lo más alta posible, se elige la el modelo de árbol de decisión con el conjunto de datos sin discretizar ya que de los 4 modelos evaluados fue el que presentó mayor número de datos correctamente clasificados, obteniendo la precisión más alta del ejercicio.

6. Clustering

Los algoritmos de agrupación en clúster son generalmente utilizados en la clasificación no supervisada. En este tipo de clasificación se presenta un conjunto y el objetivo es agrupar los que tienen características de similitud. El algoritmo tiene acceso solo al conjunto de características que describe un objeto; no se ha proporcionado ninguna información sobre dónde se debe colocar cada una de las instancias dentro de la partición. Sin embargo, en algunos casos la persona que está experimentando con el conjunto de datos puede poseer algún conocimiento de fondo que podría resultar útil para la agrupación de los datos.

Para esta clasificación se utilizó la técnica de k-Means el cual es un algoritmo particional que solicita un número de clusters y a partir de ese número busca la cantidad óptima de clusters para el conjunto de datos. Este algoritmo converge cuando no existen cambios en la asignación de instancias a clusters. Al no encontrar literatura relacionada a clasificación no supervisada

para este conjunto de datos o uno similar, se realiza un proceso de clustering con 3 valores diferentes, ya que el problema más común de esta técnica es encontrar el valor óptimo de k, y de estos valores se elige el que clasifique de una mejor manera los datos y que sea gráficamente fácil de analizar. Los valores con los que se trabajaron fueron escogidos aleatoriamente, se trabajó con valores de 2, 3 y 4.

Como apoyo para la elección del valor de k que mejor se ajusta a la clasificación de los datos se cuenta con una matriz de correlación que muestra qué atributos tienen una relación positiva y cercana a uno, negativa y cercana a menos uno o los que su factor de correlación es cercano a cero lo que luego del clustering nos puede indicar los atributos en los que debemos fijar en la matriz de gráficas de dispersión para identificar los clusters. En la siguiente gráfica las casillas de color azul indican que la relación entre las variables es positiva y a medida que se vuelve más oscuro el color indica que este valor es más cercano a uno, mientras que las casillas de color rojo muestran una relación negativa.

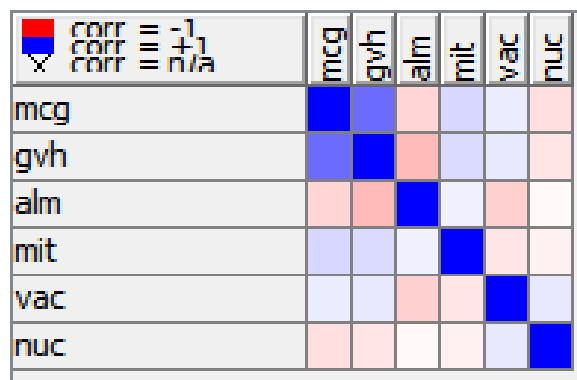


Ilustración 19 Matriz de correlación.

Analizando las matrices de diagramas de dispersión y su relación con el factor de correlación de cada atributo podría decirse que los atributos cuyo factor está más cercano a cero independientemente de si es positivo o negativo, no deberían ser tenidas en cuenta para la elección de la cantidad de clusters ya que no se distinguen a simple vista los grupos dentro de los diagramas.

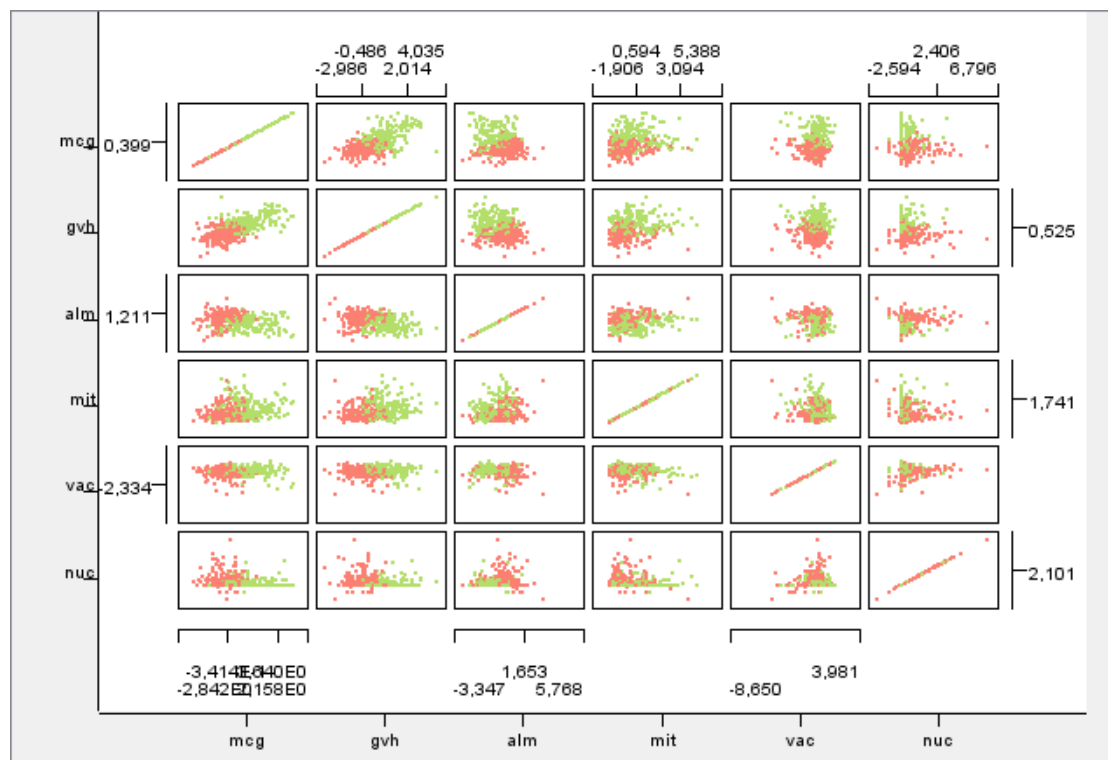


Ilustración 20 Matriz de diagramas de dispersión k=2.

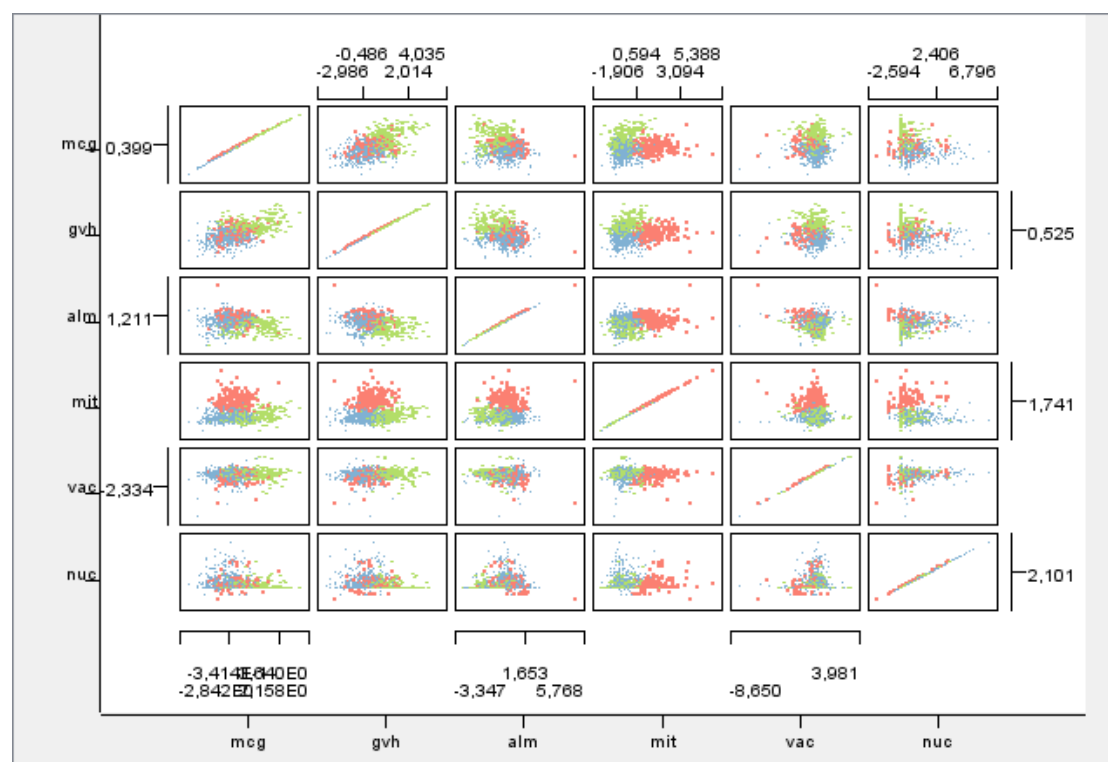


Ilustración 21 Matriz de diagramas de dispersión k=3.

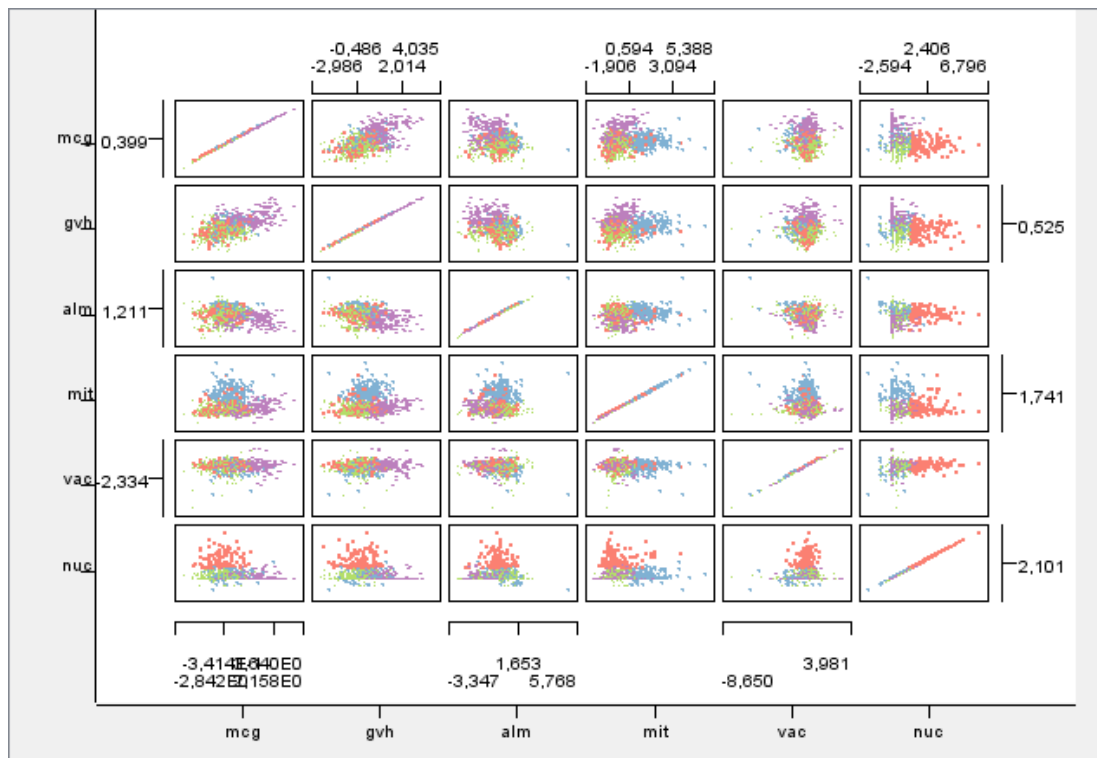


Ilustración 22 Matriz de diagramas de dispersión k=4.

Observando las ilustraciones 20, 21 y 22 es posible concluir que el atributo que más correlación tiene con los demás atributos es “mit” por lo que la toma de decisión del k que mejor se ajusta al conjunto de datos en base a los diagramas de dispersión de este atributo.

Se descarta el valor de k=4 ya que con este no es posible distinguir los clusters con claridad. Entre los valores 2 y 3 existe un equilibrio en los clusters que tienen relación con el atributo “mit” por lo que se analizan los demás atributos para poder llegar a una decisión acertada. Si se observa con detenimiento, en la ilustración 21 algunos de los diagramas de dispersión del atributo “gvah” y todos los del atributo “nuc” no permiten distinguir con claridad los clusters, mientras que estos atributos en la ilustración 20 sin más notorios los límites de los clusters, por lo que se toma la decisión de trabajar con un k=2 para la clasificación no supervisada de este conjunto de datos.

CONCLUSIONES

- El error medio cuadrático es menor en la técnica de imputación del vecino más cercano debido a que las otras técnicas imputan siempre el mismo valor y con esto alteran notablemente el comportamiento de los datos, mientras

que el vecino calcula la distancia dependiendo de la posición del valor faltante y esto puede contribuir a que el error calculado sea menor.

- La diferencia en los resultados de la precisión obtenidos en este ejercicio y los de la literatura presentada pueden deberse a que el conjunto de datos que se trabajó aquí sufrió una transformación debido a que se agregaron aleatoriamente valores faltantes y se imputaron mediante la técnica del vecino más cercano con un valor de $k=8$.
- Las medidas de tendencia central del conjunto de datos original no varían drásticamente si se comparan con los del conjunto de datos resultante de la imputación por el vecino más cercano con $k=8$.
- Es importante experimentar la clasificación con diferentes técnicas de pre procesamiento aplicadas al conjunto de datos, ya que teniendo en cuenta la naturaleza de los atributos no es posible predecir cuál técnica aportará mayor precisión al ejercicio.
- La literatura encontrada fue un apoyo importante en la determinación del valor de k , ya que (Horton & Nakai, 1997) [1] encontraron un valor en el que se alcanzaba un alto nivel de precisión en la clasificación supervisada con la técnica del vecino más cercano.
- Trabajar la clasificación en KNIME requiere de un nodo de predicción adicional para poder conocer la precisión el ejercicio ya que en los resultados de la técnica no es posible obtener esta información.
- Los flujos de trabajo de KNIME permiten al usuario un mejor entendimiento del proceso en comparación con otras herramientas utilizadas para la clasificación de conjuntos de datos, sea o no supervisada.

REFERENCIAS

- [1] Horton, P., & Nakai, K. (1997). Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier. *American Association for Artificial Intelligence*.

BIBLIOGRAFÍA

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, 3(6).

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 65.

Patil, P. H., Thube, S., Ratnaparkhi, B., & Rajeswari, K. (2014). Analysis of different data mining tools using classification, clustering and association rule mining. *International Journal of Computer Applications*, 93(8).

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In *ICML* (Vol. 1, pp. 577-584).

Mishra, R. (2018). KNN example in R. Tomado de https://rstudio-pubs-static.s3.amazonaws.com/123438_3b9052ed40ec4cd2854b72d1aa154df9.html

Example for Learning a Decision Tree | KNIME. (2018). Tomado de <https://www.knime.com/nodeguide/analytics/classification-and-predictive-modelling/example-for-learning-a-decision-tree>

Performing a k-Means Clustering | KNIME. (2018). Tomado de <https://www.knime.com/nodeguide/analytics/clustering/performing-a-k-means-clustering>

DIAGRAMA

