PESQUIMAT 21(1): 11–21 (2018) ISSN:1560-912X/ ISSN-E:1609-8439
http://dx.doi.org/10.15381/pes.v21i1.15077 Facultad de Ciencias Matemáticas – UNMSM

# K-Vecino más próximo en una aplicación de clasificación y predicción en el Poder Judicial del Perú

***Nel Quezada Lucio*** [1]

**Resumen:** Este artículo resume las contribuciones principales de la tesis con el título "K-vecino más próximos en una aplicación de clasificación y predicción en el Poder Judicial del Perú". En esta tesis se construye un modelo utilizando el método de los k-vecinos más próximos que permite clasificar y predecir las Cortes Superiores de Justicia del Perú. Mediante un análisis descriptivo de datos se excluye la Corte de Lima del estudio. Con las restantes 30 Cortes Superiores, se genera un modelo de tres grupos fundado en clasificación no supervisada, para ello se deduce la matriz de distancia euclidiana que origina el árbol de clasificación. Se construye el modelo de clasificación de tres vecinos más próximos, con partición y pliegues de validación cruzada aleatoria, que indica; el modelo de espacio de predictores, el error cuadrático o índice de error que valida el valor óptimo de k = 3 vecinos, el error del modelo y el índice global de pronostico que miden la precisión o exactitud del modelo encontrado, importancia del predictor, mapas de cuadrantes y tabla de vecinos.
**Palabras clave:** clasificación supervisada y no supervisada; k-vecinos más próximos; clasificación no paramétrica; partición; validación cruzada aleatoria.

# K-Nearest neighbor in a classification and prediction application in the Judicial Power of Peru

**Abstract:** This article summarizes the main contributions of the thesis with the title "K-Nearest neighbor in a classification and prediction application in the Judicial Power of Peru". In this thesis a model is constructed using the method of the nearest k-neighbors that allows classifying and predicting the Superior Courts of Justice of Peru. Through a descriptive data analysis, the Lima Court is excluded from the study. With the remaining 30 Superior Courts, a three-group model based on unsupervised classification is generated, for which the Euclidean distance matrix that originates the classification tree is deduced. The classification model of three nearest neighbors is constructed, with partition and random cross-validation folds, which indicates; the predictor space model, the quadratic error or error index that validates the op-timal value of k = 3 neighbors, the model error and the global forecast index that measure the accuracy or accuracy of the model found, importance of the predictor, maps of quadrants and table of neighbors.
**Keywords:** supervised and unsupervised classification; nearest k-neighbors; non-parametric classification; partition; random cross-validation..

*Recibido*: 20/02/2018. *Aceptado*: 16/05/2018. *Publicado online*: 30/06/2018.



[1]UNMSM, Facultad de Ciencias Matemáticas, e-mail:nquezadal@unmsm.edu.pe

## 1. Introducción

Con la finalidad de poder optimizar los recursos económicos, logísticos, humanos; evaluar los estándares de producción judicial y determinar las principales características de las Cortes Superiores de Justicia del Poder Judicial del Perú, se construye el modelo de clasificación y predicción, basado en el método no paramétrico de los k−vecinos más próximos.

En efecto para asociar y determinar las características de las Cortes Superiores de Justicia en conglomerados se utilizó el análisis exploratorio de datos y agrupación jerárquica fundamentada en vecinos más próximos. Para el modelo de clasificación y predicción se estableció el modelo construido (espacio de predictores) de tres vecinos más próximos con validación cruzada de pliegues aleatorios, definida mediante el error cuadrático. La verificación de la validez y precisión del modelo se define mediante el error de clasificación y el índice global de pronóstico del modelo encontrado. Para ello se utilizó las variables. Pendientes: procesos que faltan resolver. Ingresos: procesos ingresados en el año. Resueltos: procesos resueltos en el año. Población: personas que viven en la Corte Superior. Órganos jurisdiccionales: cantidad dependencias judiciales. Personal: cantidad de trabajadores.

Además el modelo señalado permite actualizar los conglomerados de las Cortes Superiores de Justicia de acuerdo al momento actual de sus variables, permitiendo determinar el estado actual de las cortes mediante la lectura de los gráficos de homólogos y mapa de cuadrantes. El modelo también identifica la importancia de cada una de las variables para realizar un pronóstico, independiente de que si éste pronóstico es preciso o no.

## 2. Metodología

El nivel de medición y análisis de los datos es descriptivo, correlacional e inferencial, dado que para el análisis se recurren a gráficos, indicadores de relación y estimación. Para identificar y evaluar las relaciones existen entre 31 Cortes Superiores de Justica y sus respectivas variables (Pendientes, Ingresados, Resueltos, Personal, Dependencias, Población), se realiza análisis descriptivo que permite evaluar el número de grupos y verificar la variación de los datos atípicos. Se utiliza clasificación no supervisada de conglomerados jerárquicos para crear el modelo de agrupamiento (tres grupos), que valora la matriz de distancia euclidiana mediante encadenamiento simple, que mide el grado jerarquía y se construye el árbol de clasificación. Con tres vecinos más próximos de validación cruzada de pliegues aleatorios se construye el modelo más eficaz y preciso, que revela el modelo llamado espacio de predictores, el error cuadrático o Índice de error que determina el óptimo valor de k igual a tres vecinos, el error del modelo o de clasificación y el índice global de pronostico miden la precisión o exactitud del modelo encontrado, los gráficos de homólogas sitúa al caso focal y sus vecinos, la importancia del predictor indica la importancia de las variables, tabla de vecinos y distancias valora las distancias, mapas de cuadrantes analiza el promedio y la variación, finalmente se presenta el modelo predictivo.

De otro lado se debe precisar que los datos se recogieron sobre la base de las hipótesis y teorías planteadas, con la intención de exponer la información de manera cuidadosa y analizar minuciosamente los resultados obtenidos, con la finalidad de extraer generalizaciones significativas que contribuyan al conocimiento dentro y fuera del Poder Judicial. Todos estos resultados y análisis se fundamentan en la información obtenida de una base de datos del Poder Judicial del Perú.

El propósito principal del estudio es que el modelo de clasificación y predicción encontrado sea innovador y permita resolver los problemas de manera precisa y oportuna.

# 3. Resultados y discusión

## 3.1. Análisis exploratorio de datos

La figura 1, muestra la gráfica de las 31 Cortes Superior de Justicia construido sobre un círculo, igualmente espaciados, cuyas magnitudes de las variables nacen desde el centro del círculo que permite agrupar a priori. Es así que Lima se distinguir claramente de los demás con magnitudes más amplias, esto sugiere la conformación de un grupo. Las 30 Cortes Superiores pueden formar, el grupo uno: Arequipa, Cusco, Ica, Junín, La Libertad, Lambayeque, Lima Norte, Piura. Grupo dos: Cajamarca, Callao, Lima Sur, Loreto, Puno, San Martín. Grupo tres las demás Cortes Superiores.

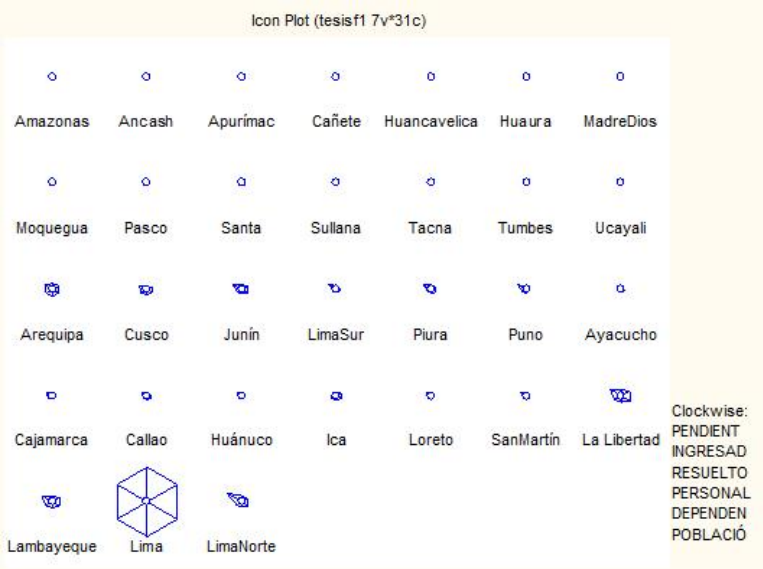


Figura 1. Grupos a priori de las Cortes Superiores de Justicia

La figura 2, presenta los valores atípicos (outlier) para cada variable, aquí se advierte que existe un valor extremo en las seis variables muy representativo que es Lima y es numéricamente distante del resto de las Cortes Superiores, lo que puede tener un efecto desproporcionado en los resultados estadísticos, así como conducir a interpretaciones engañosas.

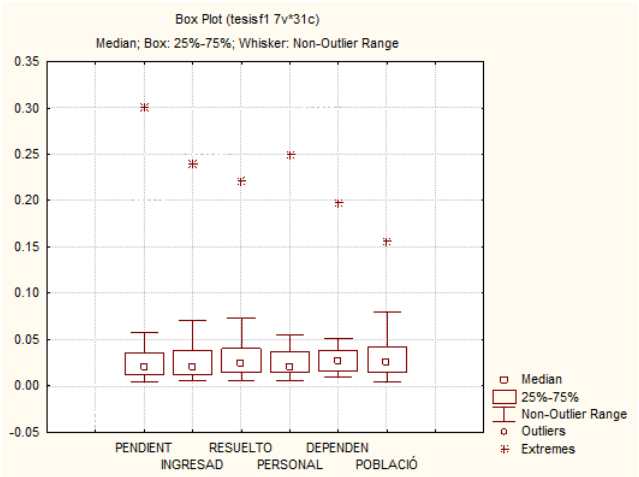


Figura 2. Valores extremos en cada variable.

Las características de la información que presenta Lima difieren significativamente en cuanto a magnitud y variabilidad en cada una de las variables en comparación con las demás Cortes Superiores de Justicia. Esto evidencia no considerar Lima para desarrollar el modelo buscado.

## 3.2. Análisis de conglomerados a posteriori

La figura 3, despliega el modelo a posteriori de tres grupos: pequeño ($n_1 = 11$), mediano ($n_2 = 9$) y grande ($n_3 = 10$). Construido mediante el método de conglomerados jerárquicos, que calcula la matriz distancia euclidiana que valora las distancias existentes entre cada una de las Cortes Superiores, mediante encadenamiento simple o vecino más próximo se encuentra el historial de conglomeración que mide el grado de jerarquía de cada una de las Cortes Superiores, en un árbol de clasificación (dendrograma).

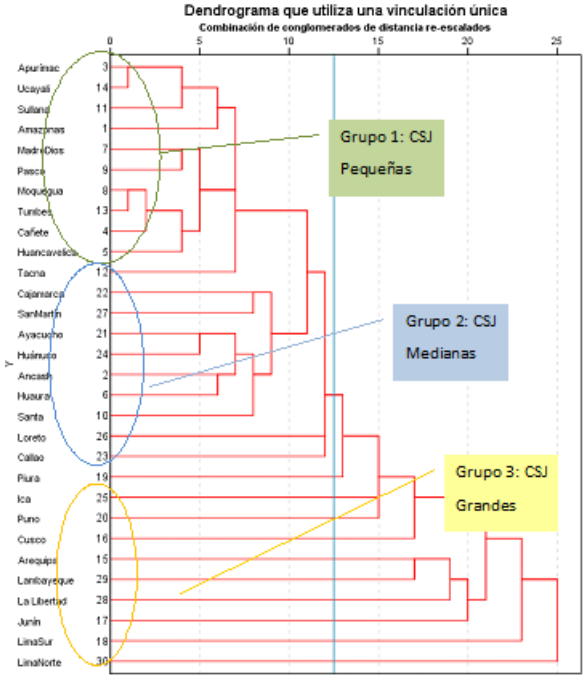


Figura 3. Árbol jerárquico de Cortes Superiores de Justicia.

## 3.3. Modelo de tres vecinos más próximos

La estimación a priori del valor de k depende del tamaño del grupo. Pequeño; $K \equiv \sqrt{n_1 = 11} =$ 3,3, Mediano;$k \equiv \sqrt{n_2 = 9} = 3$, Grande; $k \equiv \sqrt{n_3 = 10} = 3{,}2$. De otro lado la partición de entrenamiento y reserva se realizó mediante asignación aleatoria, la muestra de entrenamiento es 80 % (24) utilizado para entrenar y obtener el modelo; y la muestra de reserva 20 % (6) que se utiliza para evaluar el modelo final encontrado, dado que el tamaño de las muestras son pequeñas, utilizamos validación cruzada de pliegues aleatorios que divide la muestra en tres pliegues (sub muestras), de este modo se agranda la muestra y se mejora la estimación.

### 3.3.1. Modelo construido (Espacio de predictores)

La figura 4, presenta el modelo construido para los predictores (variables), para tres vecinos más próximo ($k = 3$), con sus respectivas particiones de entrenamiento y reserva, el caso focal en

esta ocasión es Junín, con sus tres vecinos más próximos, dos de ellos (Ica, Lambayeque) pertenece al grupo Grande y un vecino más próximos (Callao) pertenece al grupo Mediana, apoyado en el método utilizado para el modelo, Junín se clasifica en el grupo que tiene mayor probabilidad de clasificación, dado que la probabilidad de grupo grande es mayor que la probabilidad del grupo mediano. Junín se clasifica como Grande.

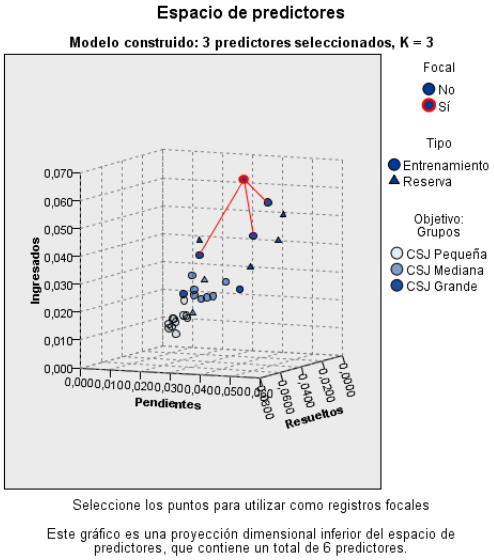


Figura 4. Modelo construido $k = 3$

### 3.3.2. Error cuadrático o Índice de error (registro de errores de selección)

La figura 5, muestra el registro de errores de selección para el modelo encontrado, donde se observa que para tres vecinos más próximos (k=3) el índice de error (error cuadrático) es de 12 %, mientras que para valores de cuatro y cinco vecinos más próximos el índice de error o de error cuadrático es superior al 20 %. En consecuencia se afirmar que el modelo con tres (k=3) vecinos más próximo es el más adecuado debido que tiene el índice de error menor.

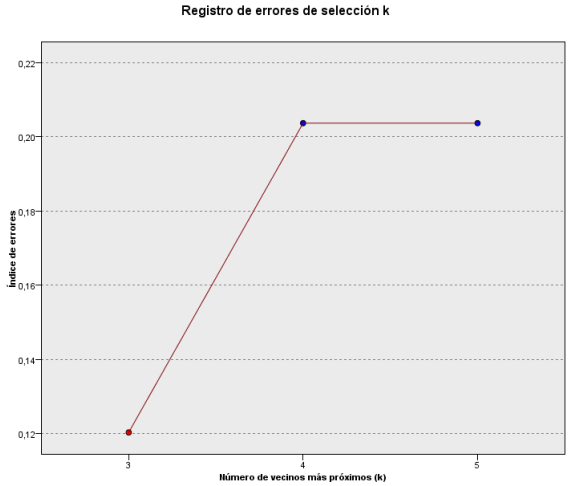


Figura 5. Error cuadrático o Índice de error de k

### 3.3.3. Error del Modelo o de clasificación (Resumen de error)

El cuadro 1, muestra el resumen de los errores asociado con el modelo, es decir, es el porcentaje de Cortes Superiores de Justicia incorrectamente clasificados. Para la muestra de entrenamiento le corresponde una tasa de error de 12.5 %, mientras que para la muestra de reserva la tasa es de 0 %. En consecuencia el modelo encontrado es considera idóneo, debido que la tasa de error encontrado es pequeña, esto es, de cada 10 Cortes Superiores Justicia clasificadas en uno de los grupos una de ellas es clasificada erróneamente.

Cuadro 1. Error del Modelo o de clasificación

| Resumen de errores | |
|---|---|
| Partición | Porcentaje de registros clasificados incorrectamente |
| Entrenamiento | 12,5% |
| Reserva | 0,0% |

### 3.3.4. Precisión o Exactitud (Tabla de clasificación)

El cuadro 2, revela la tabla de clasificación cruzada de los valores observados en comparación con los valores pronosticados de los grupos, en función de la partición (entrenamiento y reserva). Aquí, se evidencia respecto a la muestra de entrenamiento la Precisión (tasa de clasificación) es de 87.5 %, mientras que para la muestra de reserva es de 100 %. En consecuencia se puede afirmar que el modelo es aceptable, dado que la tasa de clasificación global del modelo (presión) es 100 %, esto implica que el modelo para tres vecinos más próximo es el más es adecuado.

Cuadro 2. Precisión (Índice global pronosticado)

| Tabla de clasificación | | |
|---|---|---|
| Partición | Observado | Pronosticado |
| | | Porcentaje correcto |
| Entrenamiento | Porcentaje global | 87,5% |
| Reserva | Porcentaje global | 100,0% |

### 3.3.5. Gráfico de homólogos

La figura 6, muestra el caso focal (Junín) y sus tres vecinos más próximos Ica, Lambayeque y Callao. La barra «Grupos» evidencia que Junín, Ica y Lambayeque pertenecen al grupo grande, mientras que Callao al mediano. Las barras de variables indica el valor que tiene Junín y sus tres vecinos más próximos en cada variable. Por ejemplo, la barra «pendientes» detalla la ubicación de cada una de las Cortes Superiores respecto al valor de la variable pendiente, es así que Junín está más cerca de Lambayeque, pero más lejos que Callao e Ica.

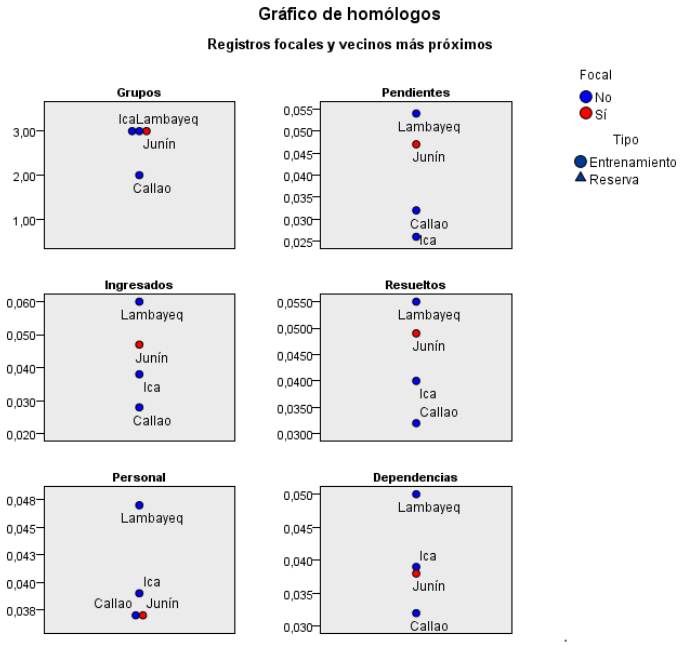


Figura 6. Gráfico de homólogos k=3.

### 3.3.6. Importancia del predictor

La figura 7, presenta la importancia relativa de las variables, la variable pendiente es la más importante en la estimación del modelo ya que su índice de importancia es cercano al 20 %, seguida por las demás variables, con índice cercano al 16 %. Además la suma de los indicadores de importancia de todas las variables suma 100 %. De otro lado, es necesario manifestar que este gráfico sólo está relacionado con la importancia de cada una de las variables para realizar un pronóstico. Es decir, es independiente de que si éste pronóstico es preciso o no.

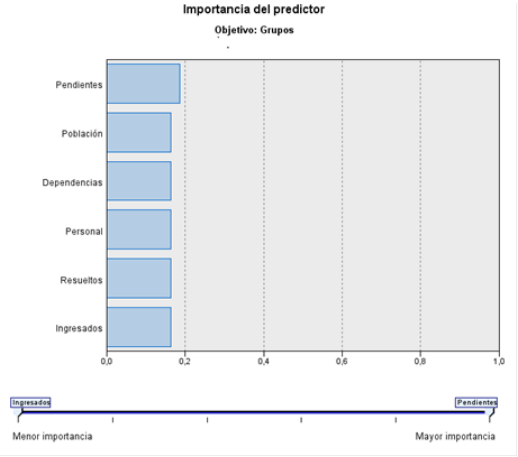


Figura 7. Gráfico de importancia del predictor.

### 3.3.7. Tabla de vecinos y distancias

El cuadro 3, presenta a Junín (focal) con sus tres vecinos más próximos, con sus respectivas distancias (distancia euclidiana) que se mide desde Junín hasta cada uno de sus tres vecinos. Por

ejemplo, el primer vecino más próximo es Lambayeque cuya distancia tiene un valor de 0.373, el segundo vecino es Callao con una distancia de 0.504 y el tercer vecino es Ica con una distancia de 0.540. Se debe precisar que se puede considerar otra Corte Superior como caso focal y realizar un análisis similar a lo descrito anteriormente cuando Junín era caso focal.

Cuadro 3. Vecinos más próximos y distancias para k=3.

**Vecinos más próximos k y distancias**

**Mostrado para los registros focales iniciales**

| Registros focal | Vecinos más próximos | | | Distancias más próximas | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| Junín | Lambayeq | Callao | Ica | 0,373 | 0,504 | 0,540 |

### 3.3.8. Mapa de cuadrantes

La figura 8, presenta los valores de objetivo por predictores (variables) para los registros focales y vecinos más próximos iniciales, se observa a Junín (focal) y sus tres vecinos más pr óximos representados en un diagrama de dispersión con los grupos en el eje vertical (y) y las variables en el eje horizontal (x). Del mismo modo, si nos centramos en el mapa de cuadrantes respecto de la variable pendiente se evidencia que Ica y Callao se encuentran situados debajo de la media (líneas punteadas) mientras que Junín y Lambayeque se encuentran en la parte superior de la media, lo mismo sucede con las variables ingresados y resueltos, mientras que para las variables personal y dependencias se muestra que Ica, Callao y Junín se encuentran debajo de la media, mientras que Lambayeque se sitúa en parte superior de la media.

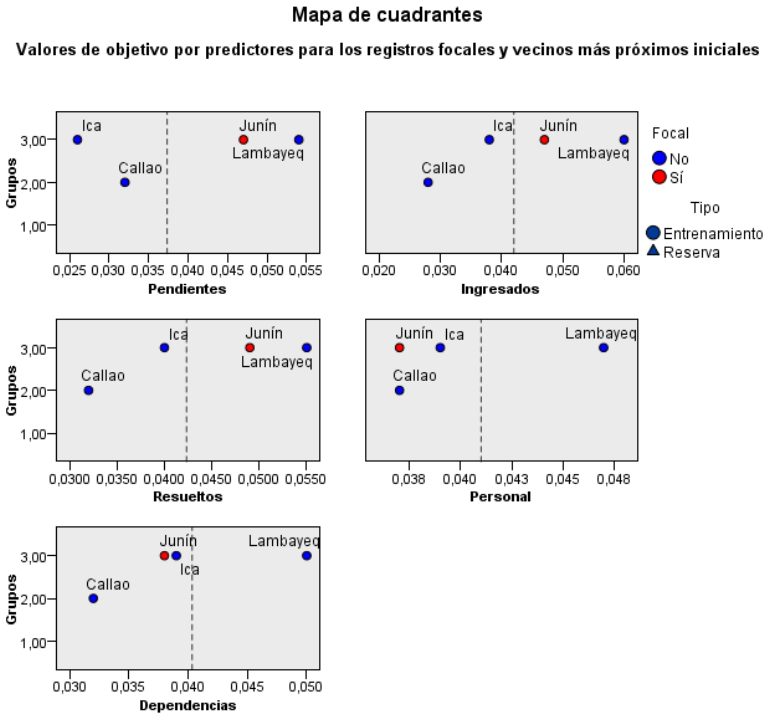


Figura 8. Mapas de cuadrantes para k=3.

### 3.3.9. Tabla de clasificación tres vecinos

El cuadro 4, muestra los grupos iniciales mediante conglomerados jerárquicos, los grupos pronosticados mediante tres vecinos más próximos, las particiones, los pliegues y las probabilidades de clasificación para cada grupo. Si observamos la columna grupo pronosticado mediante tres vecinos más próximos, se evidencia que algunas Cortes Superiores cambian de grupo respecto a la columna Grupo (conglomerados jerárquicos).

Cuadro 4. Grupos y probabilidades para k=3.

| Corte Superior | Grupos | Grupos Pronosticado VMP | Particiones VMP | Pliegues VMP | Probabilidad 1 - VMP | Probabilidad 2 - VMP | Probabilidad 3 - VMP |
|---|---|---|---|---|---|---|---|
| Amazonas | 1 | 1 | 1 | 2 | 0.67 | 0.17 | 0.17 |
| Ancash | 2 | 2 | 1 | 2 | 0.17 | 0.67 | 0.17 |
| Apurimac | 1 | 1 | 1 | 2 | 0.67 | 0.17 | 0.17 |
| Cañete | 1 | 1 | 1 | 3 | 0.67 | 0.17 | 0.17 |
| Huancavelica | 1 | 1 | 1 | 1 | 0.67 | 0.17 | 0.17 |
| Huaura | 2 | 2 | 1 | 1 | 0.17 | 0.67 | 0.17 |
| Madre Dios | 1 | 1 | 1 | 2 | 0.67 | 0.17 | 0.17 |
| Moquegua | 1 | 1 | 1 | 2 | 0.67 | 0.17 | 0.17 |
| Pasco | 1 | 1 | 1 | 2 | 0.67 | 0.17 | 0.17 |
| Santa | 2 | 2 | 1 | 1 | 0.17 | 0.67 | 0.17 |
| Sullana | 1 | 1 | 1 | 3 | 0.67 | 0.17 | 0.17 |
| Tacna | 1 | 1 | 1 | 1 | 0.67 | 0.17 | 0.17 |
| Tumbes | 1 | 1 | 1 | 1 | 0.67 | 0.17 | 0.17 |
| Ucayali | 1 | 1 | 1 | 1 | 0.67 | 0.17 | 0.17 |
| Arequipa | 3 | 3 | 0 | 0 | 0.17 | 0.17 | 0.67 |
| Cusco | 3 | 3 | 0 | 0 | 0.17 | 0.33 | 0.50 |
| Junin | 3 | 3 | 1 | 3 | 0.17 | 0.33 | 0.50 |
| Lima Sur | 3 | 2 | 1 | 3 | 0.17 | 0.67 | 0.17 |
| Piura | 3 | 3 | 0 | 0 | 0.17 | 0.33 | 0.50 |
| Puno | 3 | 2 | 1 | 3 | 0.17 | 0.67 | 0.17 |
| Ayacucho | 2 | 2 | 1 | 3 | 0.17 | 0.67 | 0.17 |
| Cajamarca | 2 | 2 | 0 | 0 | 0.17 | 0.50 | 0.33 |
| Callao | 2 | 2 | 1 | 1 | 0.17 | 0.50 | 0.33 |
| Huánuco | 2 | 2 | 1 | 2 | 0.17 | 0.67 | 0.17 |
| Ica | 3 | 2 | 1 | 1 | 0.17 | 0.67 | 0.17 |
| Loreto | 2 | 2 | 0 | 0 | 0.33 | 0.50 | 0.17 |
| San Martín | 2 | 2 | 1 | 3 | 0.17 | 0.67 | 0.17 |
| La Libertad | 3 | 3 | 1 | 3 | 0.17 | 0.17 | 0.67 |
| Lambayeque | 3 | 3 | 1 | 1 | 0.17 | 0.17 | 0.67 |
| Lima Norte | 3 | 3 | 0 | 0 | 0.17 | 0.17 | 0.67 |

### 3.3.10. Dispersión para 3 vecinos más próximos para los grupos

La figura 9, presenta la gráfica de la variable resueltos versus la variable ingresados en un espacio bidimensional de los valores pronosticado para la agrupación realizada por tres vecinos m ás próximo, donde los puntos representan las Cortes Superiores de Justicia en sus respectivos grupos. Se observa que los grupos formados son diferentes, es decir, no existen solapamientos entre los grupos (pequeño, mediano y grande). En consecuencia la formación de los grupos por tres vecinos más próximos es significativa.

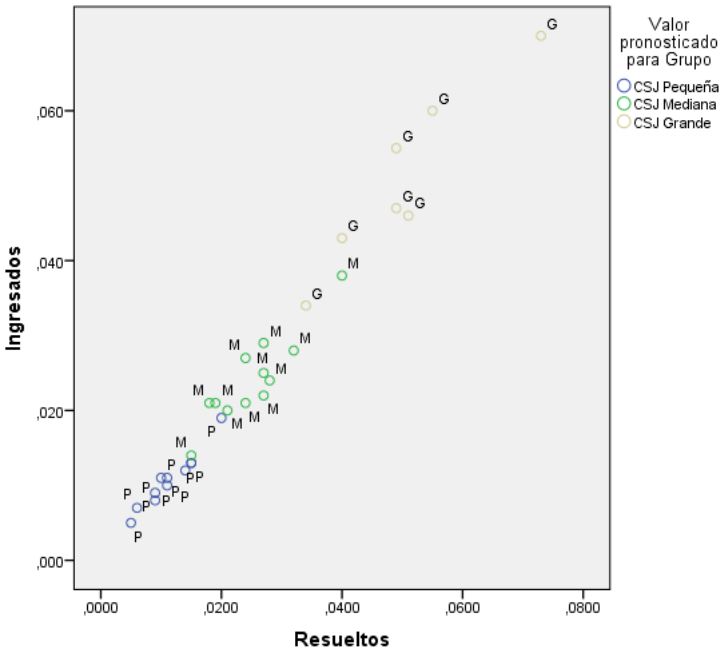


Figura 9. Dispersión del Modelo 3 vecinos más próximos.

### 3.4. Modelo de Predicción mediante tres vecinos más próximos

Para aplicar el modelo de predicción se utiliza como ejemplo los datos de la Corte Superior de Lima, cuyo vector de variables (Y) es de seis dimensiones como se muestra. Y=(pendientes, ingresos, resueltos, personal, dependencias, población).
Desarrollando el modelo:
knn(X, Y, C, k=3, prob=TRUE)
Para valores de:
$Y < -matrix(c(0{,}30, 0{,}24, 0{,}22, 0{,}25, 0{,}20, 0{,}16), nrow = 1)$
$X < -Tesis[2:7]$
$C < -Tesis[, 8]$ Se obtiene.

Cuadro 5. Modelo de predicción para k=3.

| Modelo | Predicción |
|---|---|
| knn(X, Y, C, k = 3, prob=TRUE) | Grupo Grande |

## 4. Conclusiones

- La Corte Superior de Lima es excluida para encontrar los modelos de clasificación y predicción del presente estudio, dado que, las características de información que despliega, son muy diferentes en cuanto a la magnitud de las variables (Figura 1) y numéricamente distante del resto de las Cortes Superiores, en cada una de las variables (Figura 2).

- Se estableció el modelo de tres conglomerados (Figura 3); Pequeño, Mediano y Grande, que se sustenta en el método de conglomerados jerárquico con vecinos más próximos, que levanta el árbol jerárquico basado en disimilaridad de las Cortes Superiores de Justicia.

- Se estableció el modelo óptimo de clasificación y predicción para tres vecinos más próximos, debido a que el error cuadrático (registro de errores) es 12 % mientras que para 4 y 5 vecinos es mayor al 20 %, (Figura 5).
- El modelo encontrado de tres vecinos más próximos, es preciso y exacto, esto se evidencia mediante, el índice de precisión del modelo (Cuadro 2) para la muestra de reserva es 100 % y para la muestra de entrenamiento es 87.5 % y el error del modelo de clasificación (Cuadro 1) para reserva es de 0 % y para entrenamiento es de 12.5 %. De otro lado, El modelo de tres vecinos más próximos permite predecir la clasificación de futuras Cortes Superiores de Justicia (Cuadro 5).
- El modelo verifica que la variable más importante es pendientes (Figura 7) seguidos de las demás variables. La graficas de homólogos, tablas de vecinos y distancias, mapas de cuadrantes, tabla de clasificación tres vecinos y la gráfica de dispersión permite observar las relaciones existentes entre las Cortes Superiores y sus respectivas variables.

## Referencias Bibliográficas

[1] Abellanas M. (1993). *Sobre la vecindad geométrica*. España: Universidad Politécnica de Madrid.

[2] Anderberg, G. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

[3] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regresion Trees*. California: Chapman & Hall.

[4] Cortijo, F.J. (2001). *Aproximación no paramétrica*. Uruguay: Universidad de la República Uruguay.

[5] Jhonson, R. y Wichern, D. (1992). *Applied Multivarte Statistical Analysis*. London: The Prentice Hall International Editions.

[6] Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate Analysis*. New York: Academic Press.

[7] Micó L. & Oncina J. (1998). *Comparison of fast nearest neighbour classiers for handwritten character recognition. Pattern Recognition Letters (pp. 351−356)*. Spain: Universidad de Alicante.

[8] Peña, D. (2002). *Análisis de datos Multivariante*. España: Mc Graw-Hill interamericana de España.

[9] Quezada, L. (2017). *K-vecino más próximos en una aplicación de clasificación y predicción en el Poder Judicial del Perú*. Tesis de Maestría en Estadística Matemática. Universidad Nacional Mayor de San Marcos, Lima, Perú.