UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

FACULTAD DE CIENCIAS SOCIALES Y JURÍDICAS DE ELCHE

GRADO EN ESTADÍSTICA EMPRESARIAL

ANÁLISIS DE DATOS EN CENTROS DE ATENCIÓN TEMPRANA

TRABAJO FIN DE GRADO

Diciembre - 22

AUTOR: Diego José García González
DIRECTOR/ES: Kristina Polotskaya
Alejando Rabasa Dolado

Índice de contenido

Índice de figuras

Índice de tablas

Capítulo 1 Introducción

1.1.- EMPRESA/ENTORNO DE APLICACIÓN

El presente estudio se lleva a cabo en el marco de un convenio que tiene suscrito la UMH con la Generalitat Valenciana, para el análisis y seguimiento de los recursos públicos en el ámbito de los servicios sociales públicos.

Los datos manejados provienen de una simulación a partir de datos reales y anonimizados para cumplir escrupulosamente con las cláusulas de protección de datos vigentes.

En este TFG se han recogido información **anonimizada** de encuestas realizadas a parientes de niños que están en centros de atención temprana de la Comunidad Valenciana(fig 1), estos son centros públicos multidisciplinarios donde se busca la atención a la población infantil entre 0 y 6 años que presentan trastornos en su desarrollo. Estas encuestas recogen información de centros de las 3 provincias de la Comunidad Valenciana (Alicante, Valencia y Castellón) y en ellas los familiares responden a preguntas referentes a la evolución del tratamiento y dan información respecto a su valoración del centro y de su personal. Con las 306 encuestas que se han recogido de los centros se desea ver en qué factores pueden mejorar los centros y que es lo que más valoran los familiares a la hora del tratamiento de sus niños. Para ello se usan técnicas de machine learning y análisis de datos.

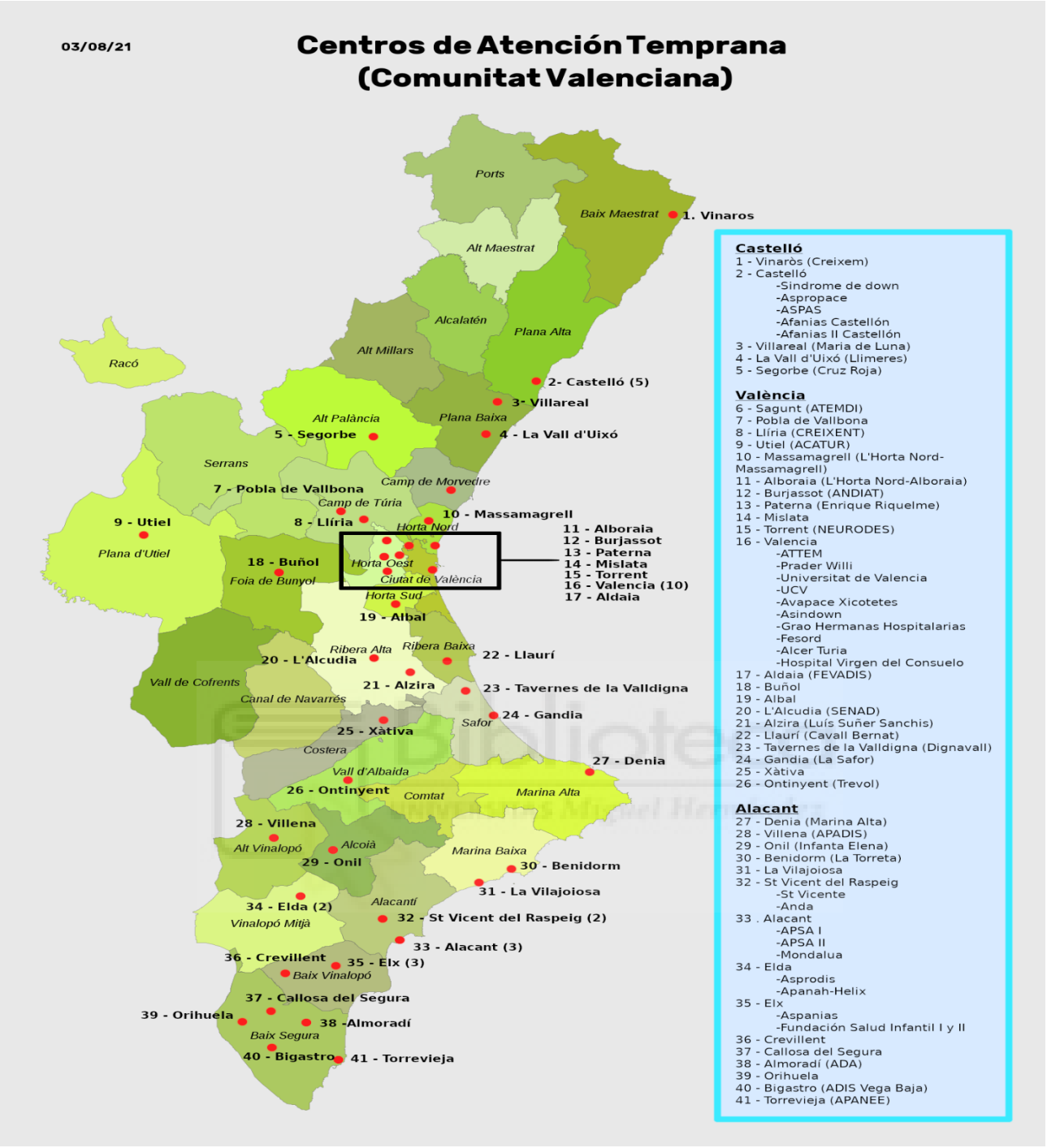

figura 1, centros de atención de temprana de la Comunidad Valenciana

1.2.- JUSTIFICACIÓN DEL PROYECTO

El personal de los centros de atención temprana busca mejorar sus sistemas y sus procedimientos a la hora de tratar a sus pacientes, para ello le piden a los familiares de estos, ya que los pacientes son muy pequeños, que rellenen una encuesta con diferentes aspectos relativos al centro y a su trato con los profesionales.

El objetivo y el motivo por el que se han facilitado estos datos es para llevar a cabo un análisis de los mismos y sacar conclusiones que ayuden a mejorar los centros, dándoles estas conclusiones de vuelta se espera que los centros puedan focalizarse en mejorar los aspectos que más preocupan a sus pacientes y mejoren la calidad de los mismos, también se busca comparar los distintos centros recogidos en la base de datos para comprobar cuáles tienen mejores resultados que otros.

Para ello se realizará un análisis descriptivo que permita tener un análisis preliminar de los datos y posteriormente se usarán técnicas de Machine Learning obteniendo un análisis más profundo de los datos y conclusiones más precisas.

Los datos disponibles se obtienen de encuestas realizadas por el centro a 306 familiares distintos donde las preguntas que contestan van desde datos personales del familiar hasta valoraciones subjetivas del centro en el que está el niño.

Otro objetivo por el que se quiere realizar este TFG es para comparar distintos métodos de Machine Learning que sirvan para el mismo objetivo y concluir qué método es el que mejor explica los datos en este caso, para ello se usará la herramienta Python donde existen diferentes librerías en las que se pueden aplicar estos algoritmos, para comparar los métodos se deberá explicar cómo funciona cada método y las diferencias que existen entre estos.

1.3.- OBJETIVOS

El objetivo general de este TFG es realizar un análisis formal de los datos de los Centros de Atención Temprana que culmine con la generación de modelos que aporten información estratégica de utilidad para los responsables del sistema de servicios sociales.

Los objetivos específicos, que desde el punto de vista analítico, se deben cubrir para alcanzar dicho objetivo general, son los siguientes:

-Seleccionar los atributos necesarios de la base de datos recibida.

-Hacer una primera limpieza de los datos que permita trabajar con estos.

- Hacer un análisis descriptivo que permita visualizar bien los datos.

- Conseguir realizar una reducción de la dimensionalidad de la base de datos.

-Comparar y clusterizar los centros de la base de datos

- Seleccionar qué variables son las que más influyen en el buen desempeño de los centros.

- Comparar el uso de distintos métodos para cada objetivo y ver cuál funciona mejor en cada caso.

- Ayudar a los centros a saber qué atributos pueden mejorar para mejorar su rendimiento.

1.4.- LÍMITES DEL PROYECTO

Debido a la falta de encuestas realizadas sería difícil predecir si las conclusiones que se realizan en este TFG serán aplicables a una gran cantidad de datos, los modelos propuestos para este TFG son aproximaciones realizadas con pocas encuestas, una mayor recolección de estos datos darían conclusiones que más se acercan a la realidad de estos centros. Otro punto en el que se podría mejorar el proyecto sería obtener datos sobre el personal del centro, incluso obtener datos donde se diferencie la función que desempeñan (figura 2).

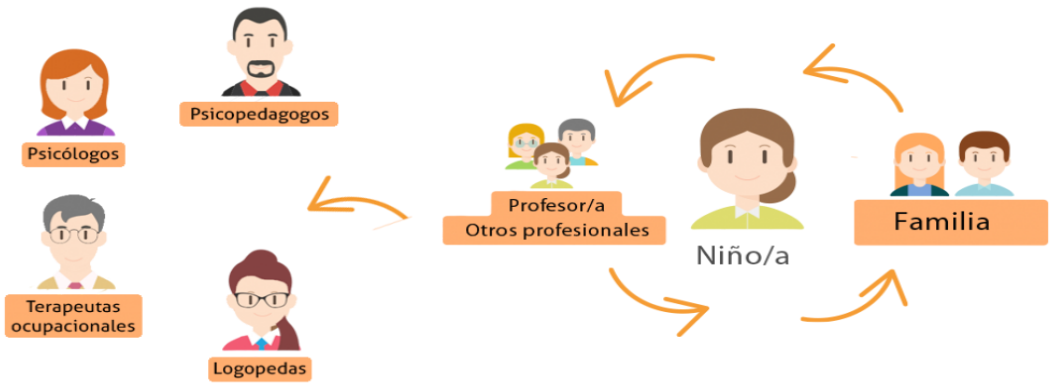

figura 2, profesionales en un centro de atención temprana

Capítulo 2 Antecedentes y estado de la cuestión

2.1.- SITUACIÓN ACTUAL

Los diferentes centros de atención temprana buscan mejorar la calidad de su servicio prestado a los niños y familiares, en este TFG se busca dar información acerca de qué centros tienen mejor calificación que otros pudiendo así identificar cuáles funcionan mejor.

Esta clasificación se hace a raíz de las encuestas donde los encuestados indican la calidad de las instalaciones y de los profesionales del centro. Estos centros buscan también poder mejorar su servicio de cara a sus pacientes, para ello se usan las encuestas para indicar que factores influyen más en la evolución positiva al tratamiento de los niños, con esta información los centros se podrán focalizar en mejorar los aspectos más importantes para los pacientes.

2.2.- HERRAMIENTAS DISPONIBLES EN EL MERCADO

Los datos recibidos por parte de los centros los recibimos en un fichero Excel en el que se recoge información de los familiares de los niños con tratamiento, para este TFG se usará el lenguaje Python en el que con diferentes librerías se pueden aplicar los diferentes métodos que se buscan y sacar los resultados esperados, también permitirá comparar entre métodos que se usan para lo mismo y saber qué método funciona mejor en este caso. Para representar las tablas del TFG también se ha usado un Excel (tabla1 pag 17, tabla2 pag 18, tabla 3 pag 19)

2.2.1.- RESUMEN

Se quiere realizar un análisis sobre distintos centros de atención temprana de la Comunidad Valenciana, para ello se disponen de datos relativos a encuestas realizadas a familiares de niños ingresados en estos centros. Estos datos dan distinta información sobre los profesionales del centro y la calidad del mismo, además los familiares también dan información sobre la relación del niño con estos profesionales.

Teniendo en cuenta estos datos recogidos se ha decidido hacer una reducción de la dimensión de las variables recogidas en las encuestas. Una vez realizada esta reducción dimensional se clasifican las nuevas variables utilizando un método de kmedias. Los resultados mostrados por centros permiten comparar las diferencias entre estos y ver en qué aspectos pueden mejorar.

A continuación se han realizado 4 modelos para predecir la variable dependiente, un modelo usando el método xgboost con las variables binarias y otro sin las variables binarias, la idea de realizar uno con las variables binarias y otro sin las binarias es observar si al usar las variables binarias el modelo sufriría un sobreajuste y por lo tanto predecirá peor. Los otros dos modelos que se han usado han sido con el método Random Forest Regressor, uno con variables binarias y otro sin estas. La idea al usar estos 4 modelos es comparar en este caso qué modelo funciona mejor para estos datos, además estos modelos permitiran ver qué pesos tiene cada variable en los modelos predictivos.

Capítulo 3 Hipótesis de trabajo

Este TFG se desarrolla usando el lenguaje de programación Python en el entorno de Spyder en la versión 5.1.5. Aparte es necesario instalarse el software de visualización "graphviz" para poder realizar los gráficos en los árboles de decisión. Las distintas librerías que permiten importar este lenguaje consiguen poder realizar una gran cantidad de gráficos y usar técnicas de machine learning con las que se puede desarrollar un análisis de las bases de datos, actualmente es uno de los lenguajes más usados para esta función.
Las distintas librerías usadas en este caso han sido:

-sklearn: permite usar distintas técnicas de machine learning, en este caso se ha usado para estandarizar los datos, para transformar las variables categóricas en numéricas, para realizar en análisis de Kmedias, el de Componentes Principales, Análisis Factorial y también para realizar los RandomForest, además también sirve para dividir los datos en la parte de entrenamiento y en la realización de los gráficos en forma de árbol.

-Pandas: Este paquete permite cargar los datos en Excel y crear todos los dataframes del código, aparte permite crear las variables binarias en variables dummies y crear tablas de contingencia.

-matplotlib: Se usa en la realización de los gráficos del TFG.

-numpy: La librería de numpy permite realizar distintas operaciones aritméticas que se realizan durante el TFG, como raíces y sumas.

Xgboost: Permite realizar el modelo predictivo de Xgboost.

AdjustText: En los gráficos de Kmedias permite separar el texto y crear líneas señalando que nombre es cada punto, es muy útil a la hora de ordenar el texto en gráficos con muchos nombres.

Capítulo 4 Metodología y resultados

4.1.- PLANIFICACIÓN DEL PROYECTO

4.1.1- DIAGRAMA DE GANTT

Se ha realizado un diagrama de Gantt (figura 3) con el ciclo de vida realizado durante este proyecto.

Tareas	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre
Observacion de las encuestas										
Preprocesamiento de los datos										
Realización de la estadistica descriptiva										
Reducción de la dimensionalidad y clusterización										
Modelos predictivos										
Conclusiones del trabajo										

Figura 3, Diagrama de Gantt con el ciclo de vida del proyecto.

Las distintas etapas del proyecto consisten en:

Observación de las encuestas: Observar que encuestas se iban a elegir para llevar a cabo el análisis posterior y ver qué variables pueden ser útiles para el mismo.

Preprocesamiento de los datos: Transformación de las variables que se usarán en el análisis y eliminación de las variables que no se usarán en el mismo.

Realización de la estadística descriptiva: Observar la correlación entre variables y ver la variabilidad de estas.

Reducción de la dimensionalidad y clusterización: Aplicar técnicas que permitan reducir la dimensión de los datos y clusterizar estas nuevas dimensiones.

Modelos predictivos: Elegir qué variable se desea predecir y crear modelos que permitan su predicción.

Conclusiones del trabajo: Analizar los resultados obtenidos de los modelos anteriores y escribir las conclusiones del proyecto.

4.1.2- DATOS DEL PROYECTO

Los datos en los que se ha basado este TFG han sido extraídos mediante procesos de simulación y anonimización a partir de los datos facilitados por los distintos centros de atención temprana. Estos centros hacen una encuesta a diferentes miembros de los centros y responden distintas preguntas relativas al centro y a su experiencia en el centro. De las distintas encuestas que se han facilitado se ha decidido escoger con la encuesta a los familiares de los niños ya que es la que más filas tiene.

Esta base de datos consta de 359 filas y 51 columnas de las cuales no se usarán todas las columnas ya que muchas son irrelevantes para el análisis que se va a realizar.

De esta base de datos se eliminan los centros que solo tengan una muestra y las filas que tengan valores vacíos con lo que se queda al final con 247 filas.

De las diferentes variables se usan las que consideramos relevantes que son:

- El centro del que viene la encuesta se trata de una variable categórica.

-21 variables ordinales cualitativas en una escala de Likert y que responden a distintas preguntas con respecto al centro y la calidad del personal y de las instalaciones.

-10 variables binarias que se transformarán posteriormente en variables dummies que son relativas a la situación actual de las personas del centro.

Para poder realizar los análisis posteriores se han transformado las variables ordinales cualitativas en cuantitativas ya que facilita mucho a la hora de poder realizar el análisis posterior.

4.2.- ANÁLISIS DESCRIPTIVO DE LOS DATOS

Antes de realizar los análisis pertinentes se realiza una estadística descriptiva de los datos para describir comprender mejor los datos que se están tratando,

En primer lugar, se observan las correlaciones entre las variables que se pueden usar para la reducción de la dimensionalidad.

Como se ve en la figura 4 hay una gran variedad de correlaciones entre las variables variando entre 0.24 y 0.87, por lo que a la hora de realizar la reducción de dimensiones [2] habrá que usar las variables más correladas entre sí.

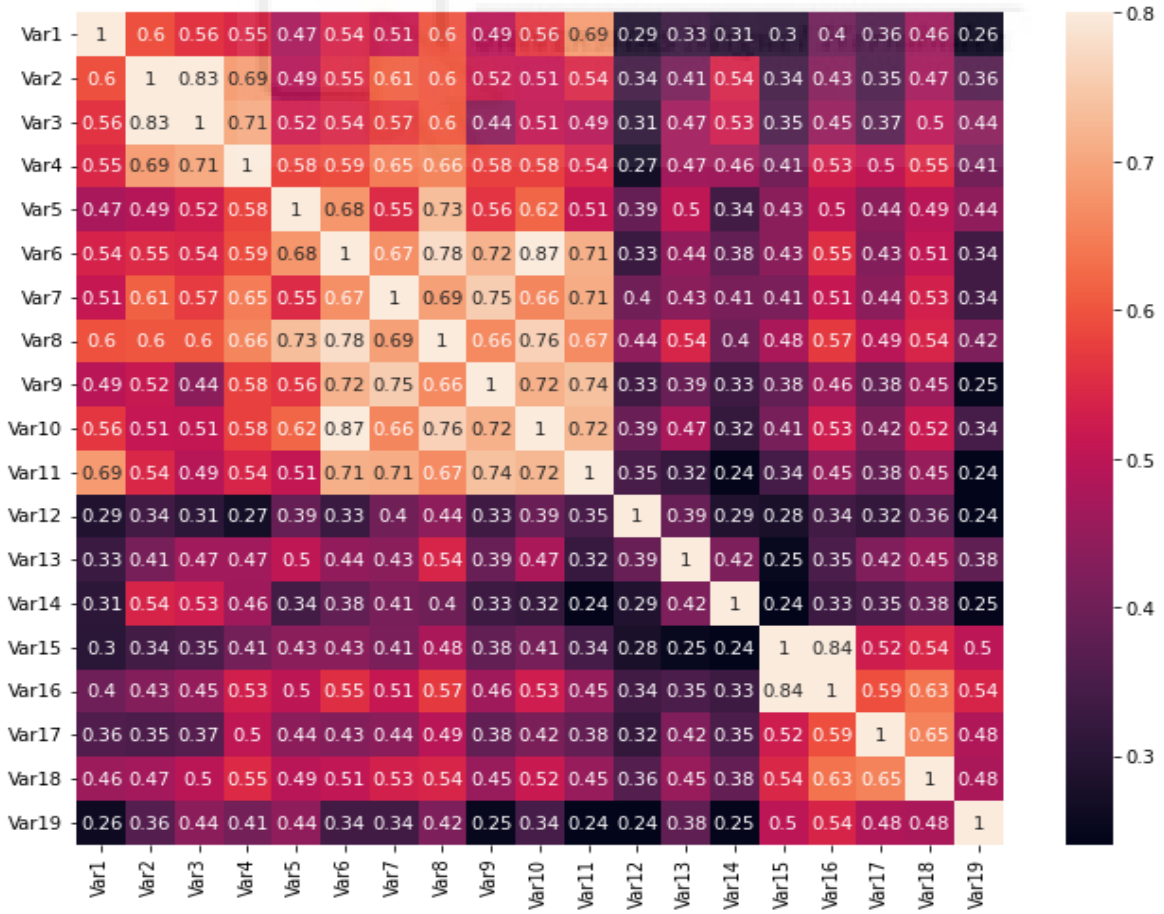

Figura 4, matriz de correlaciones

Se observa como los valores más oscuros representan que esas variables tienen menos correlación mientras que cuando el color es más claro esas variables están más correladas. En el gráfico también se observa que la diagonal vale 1, esto es debido a que la correlación de una variable con ella misma siempre vale 1.

Otro gráfico que se ha realizado para el análisis descriptivo son unos gráficos boxplot sobre las variables del análisis. Este tipo de gráfico permite ver la mediana y los cuartiles de las variables lo que da una idea de la varianza de estas.

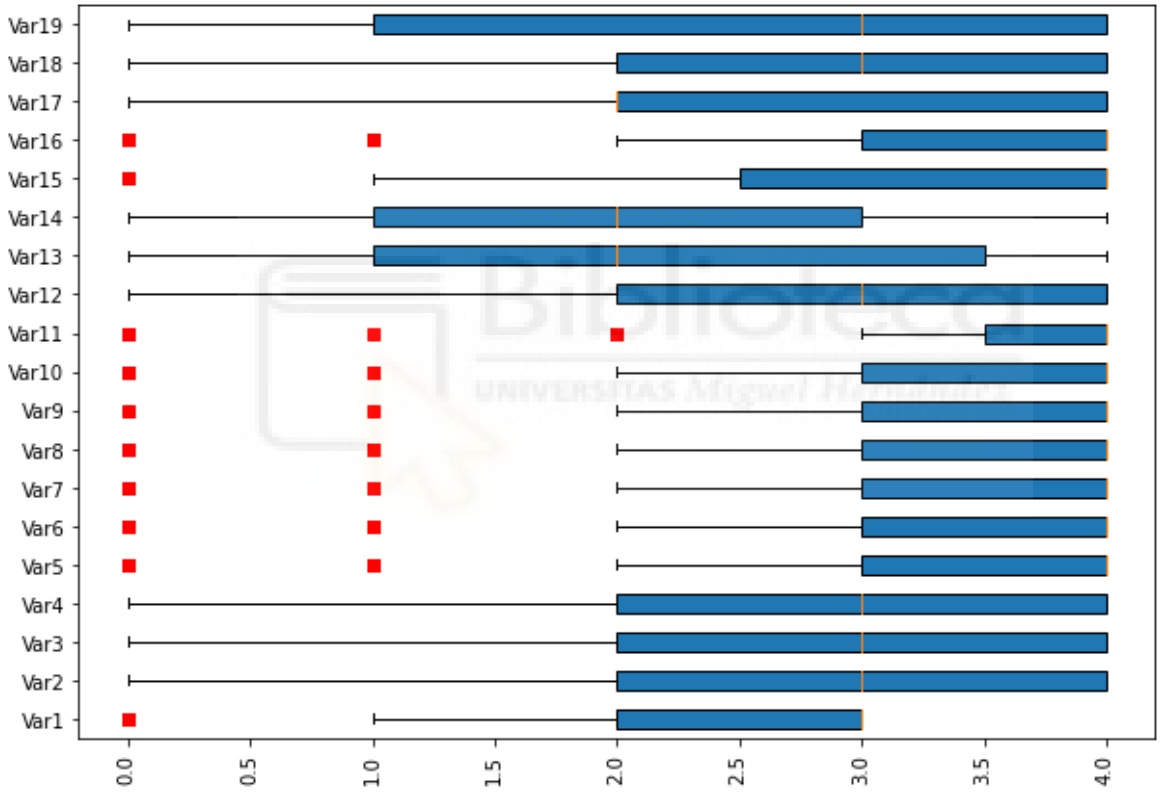

Figura 5, gráficos boxplot

Se observa que la mayoría de las variables tienen muy buenas valoraciones por parte de los familiares estando todas salvo dos la mediana entre 3 o 4. Si se observa el gráfico en la variable número 2 que es la que se usará para el modelo predictivo se ve que la mediana está en 3 y los cuartiles en 2 y 4, el extremo inferior de la variable está en el 0.

Por último ,para ver las diferencias entre centros, en el anexo se observa un gráfico en el que se ha obtenido agrupando por centros el valor de una variable que mide la calidad que le dan al centro los encuestados (figura 1 anexo).

4.3.- REDUCCIÓN DE LA DIMENSIONALIDAD Y CLASIFICACIÓN K MEDIAS

En este apartado se usan 3 técnicas distintas para reducir la dimensión de la base de datos, luego para la clusterización posterior, se usan las variables obtenidas con el método que mejores resultados haya dado.

El primer análisis que se usa para reducir la dimensionalidad es el de componentes principales [4]. Esta técnica permite agrupar la información de las variables en una nueva variable, es decir, se crea una nueva variable que contiene la información de las variables anteriores, para crear estas nuevas variables este método realiza una combinación lineal de las variables antiguas buscando maximizar la varianza, así se crean unas nuevas variables que mantienen gran parte de la información de las variables antiguas.

Este método es muy útil a la hora de poder reducir la dimensionalidad de la base de datos al permitir reducir el número de variables de la base de datos sin perder una gran cantidad de información, además al reducir la dimensionalidad de la base de datos se podrán realizar posteriormente otras técnicas como la clusterización con mejores resultados.

En primer lugar, como posteriormente se quiere comparar los centros de atención temprana, eliminamos de nuestro análisis todas las variables que no hacen referencia al centro de atención temprana. Posteriormente se estandarizan nuestros datos para poder realizar la técnica de componentes principales. En este caso se pedirá al modelo que extraiga 2 componentes nuevas, la varianza explicada por estas nuevas componentes son de 0.617 por la primera y de 0.097 por la segunda, los pesos de las componentes son:

	Primera componente	Segunda componente
Var 4	-0.286163	0.101944
Var 6	-0.328508	0.270392
Var 7	-0.325882	0.108016
Var 8	-0.290039	-0.376273
Var 9	-0.340296	0.0101366
Var 10	-0.31777	0.263743
Var 11	-0.313858	-0.264107
Var 12	-0.328021	0.232161
Var 13	-0.310635	0.286609
Var 14	-0.209909	-0.583788
Var 15	-0.237556	-0.383379

tabla 1, pesos de las componentes en el análisis de componentes principales

La segunda técnica que se usa para reducir la dimensionalidad es la técnica de análisis factorial, funciona muy similar a la técnica de componentes principales pero en este caso, las variables originales se definen como combinaciones lineales de los factores.

Al usar un método distinto a la hora de reducir las dimensiones lo que se busca es poder comparar los métodos [5] y ver cuál funciona mejor en esta base de datos.

Se extraen en este metodo tambien dos componentes, en este caso la varianza explicada por estos es de 0.583 por el primer componente y 0.063 por el segundo y la matriz de pesos en este caso sería:

	Primera componente	Segunda componente
Var 4	-0.69936	-0.0750484
Var 6	-0.891389	0.247239
Var 7	-0.808199	-0.0997402
Var 8	-0.701311	-0.46993
Var 9	-0.876832	-0.0668136
Var 10	-0.80843	0.0677745
Var 11	-0.77662	-0.384788
Var 12	-0.884822	0.257298
Var 13	-0.796537	0.0983135
Var 14	-0.483733	-0.341556
Var 15	-0.561634	-0.227258

tabla 2, pesos de las componentes en el análisis factorial sin rotación

Por último, se ha realizado otro análisis factorial pero esta vez con 3 componentes, además se ha usado en este caso la rotación varimax para añadir otra variación con respecto a los otros métodos.

En esta nueva reducción la primera componente explica el 31.934 la segunda un 22.035 y la tercera un 14.299, y la tabla de pesos sería:

	Primera componente	Segunda componente	Tercera Componente
Var 4	-0.467749	-0.38502	0.370281
Var 6	-0.856232	-0.258523	0.25446
Var 7	-0.47874	-0.393037	0.629049
Var 8	-0.286367	-0.723642	0.308304
Var 9	-0.667461	-0.522344	0.261677
Var 10	-0.588846	-0.24006	0.601046
Var 11	-0.395767	-0.70856	0.306731
Var 12	-0.866461	-0.248547	0.237022
Var 13	-0.61319	-0.228184	0.52837
Var 14	-0.213736	-0.545842	0.131561
Var 15	-0.358554	-0.535227	0.0627913

tabla 3, pesos de las componentes en el análisis factorial con rotación varimax

Una vez realizada la reducción de la dimensionalidad se quiere clasificar los resultados. Para ello se decide aplicar el método de Kmedias [3] que permite esta clasificación a través de medir la distancia entre las distintas variables.

Este método permitirá ver las diferencias entre centros ya que los clasifica en distintos clusters y es muy útil a la hora de ver las diferencias entre estos, además permitirá ver los puntos positivos y negativos de cada centro.

Para realizar el análisis posterior de Kmedias se usa el análisis factorial con la rotación, ya que las distintas variables están explicadas más fuertemente por alguna componente, no se usa el análisis factorial con dos componentes al tener todas las variables un peso mayor en la primera componente y esto no permitir diferenciar correctamente entre las dos componentes.

Una vez elegido el método de reducción que se escoge se ha decidido darle un nombre a cada componente nueva. A la primera se le ha llamado Comunicación y atención de los profesionales. La segunda variable se le ha llamado Infraestructura y ubicación. Estas dos componentes tienen pesos negativos en todas las variables por lo que esto significa que existe una correlación inversa. La tercera componente se ha llamado actuación de los profesionales con su familiar.

Una vez aclarado qué variables explican cada factor se procede a realizar un método de clasificación Kmedias con los centros agrupados.

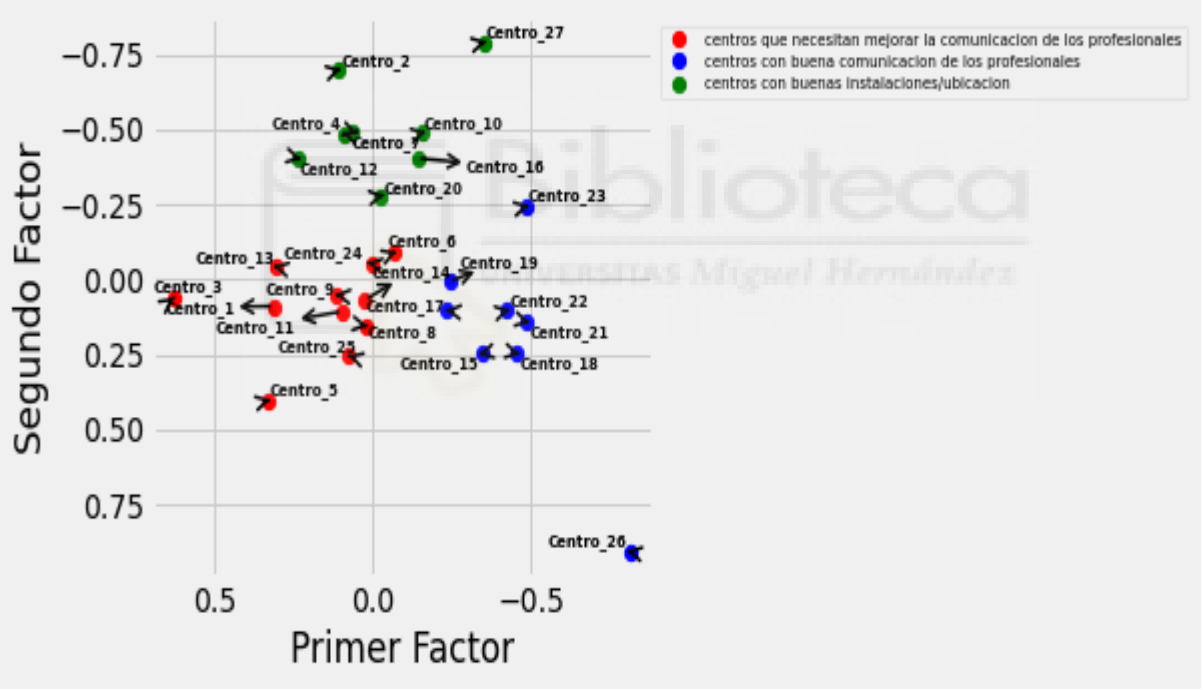

Figura 6, gráfico de primer y segundo factor

En este gráfico se observa la diferencia entre centros, se ha decidido usar 3 clusters por la distribución de los datos, también se ha decidido en el caso de el primer y segundo factor invertir el eje en el que se representan al tener pesos negativos en ella ya que de esta forma es más fácil de interpretar.

Se realizan los mismos gráficos para las otras combinaciones de factores.

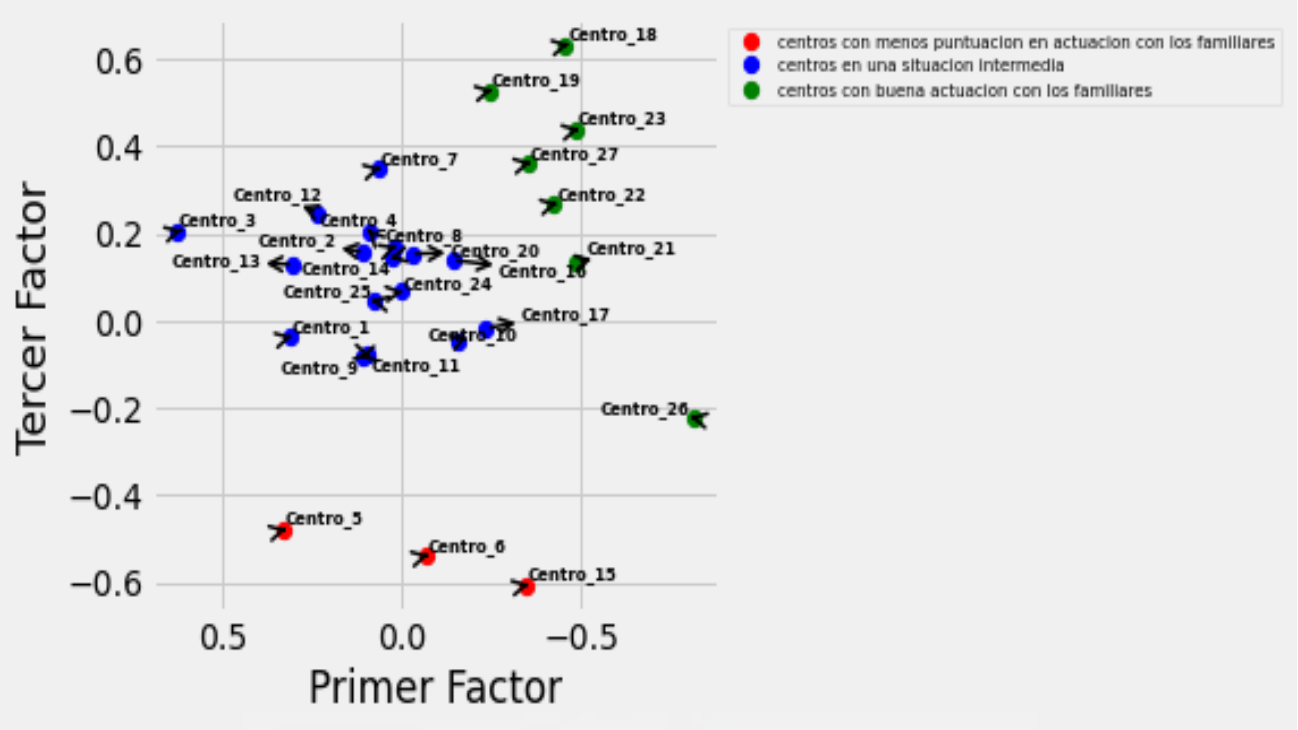

Figura 7, gráfico de primer y tercer factor

En este caso también se ha decidido dividir en 3 clusters ya que se observan tres nubes de puntos diferenciadas. Se observa que el cluster rojo solo contiene tres puntos, pero estos están muy alejados del resto de la nube de puntos.

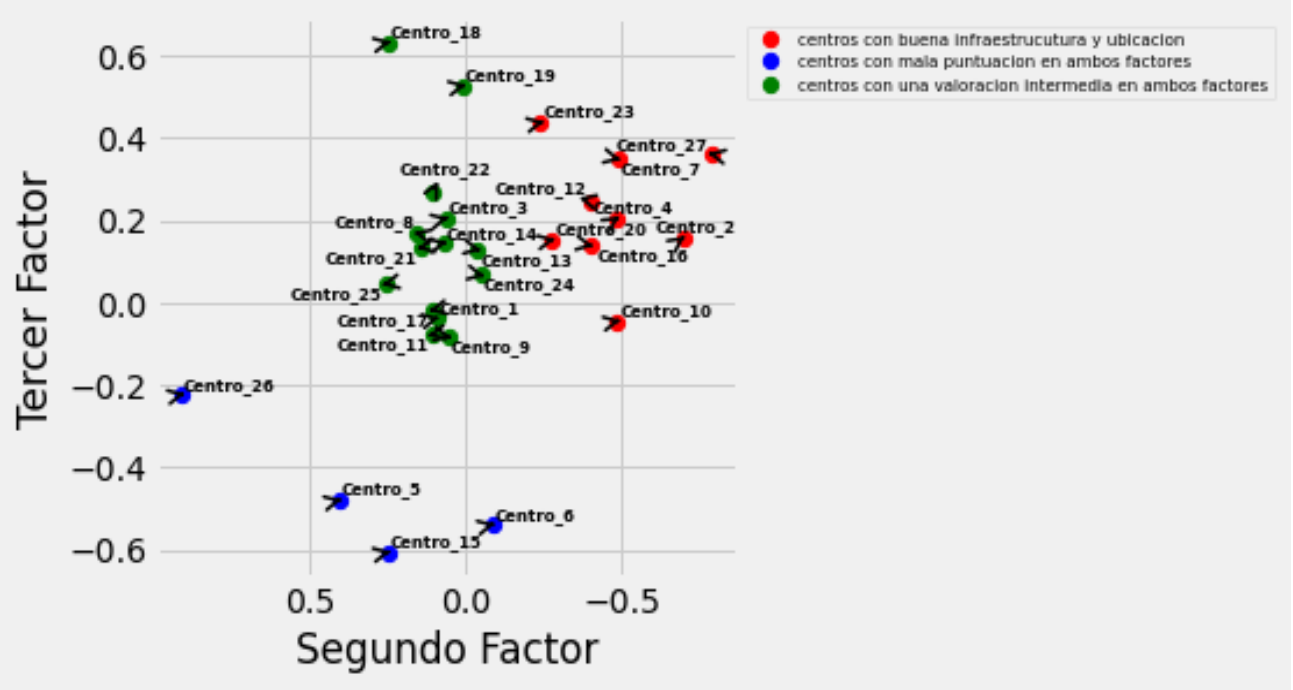

Figura 8, gráfico de tercer y segundo factor

En el caso de los factores 2 y 3, también se ha decidido elegir tres clusters a pesar de que a simple vista puede parecer que los datos son más homogéneos que en los casos anteriores.

En el anexo podemos ver los mismos tres gráficos anteriores pero sin agrupar los valores por centro (figura 2 anexo, figura 3 anexo, figura 4 anexo) estos permiten ver todos los puntos de la base de datos pero no permiten poder diferenciar claramente entre los centros

4.4.- PREDICCIÓN E IMPORTANCIA DE CADA VARIABLE.

En este apartado se han creado 4 modelos predictivos diferentes y vamos a comparar sus resultados para ver cuál responde mejor a estos datos, se usarán dos modelos xgboost uno incluyendo las variables binarias de la base de datos y otros sin estas y otros dos modelos usando la técnica RandomForest uno con las binarias y otro sin las binarias.

El modelo xg boost se trata de un modelo predictivo supervisado el cual usa árboles decisión junto con potenciación del gradiente, de esta forma mezclando estas dos técnicas se pueden obtener mejores resultados que usando una sola técnica.

En este caso será útil para poder predecir la variable 2 que es la variable dependiente de esta base de datos y la cual interesa predecir su valor y observar qué variables son las que más influyen a esta.

El primer modelo con el algoritmo xgboost da estos resultados.

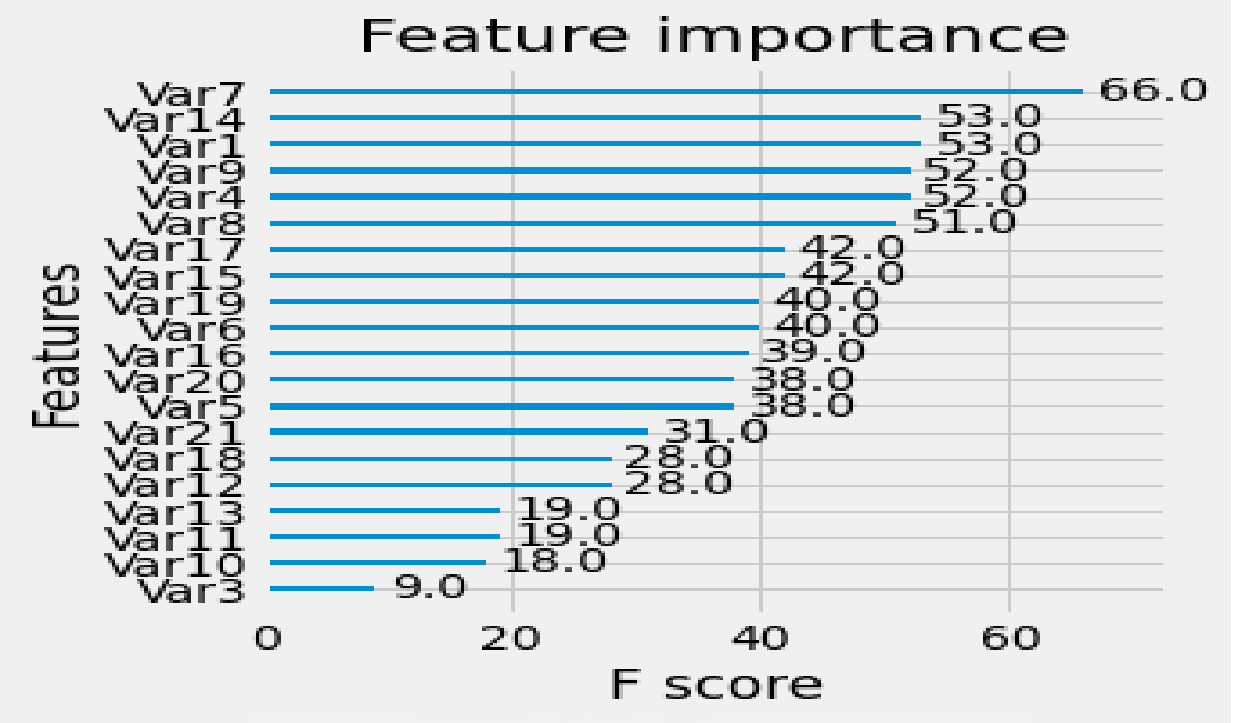

Figura 9, gráfico importancia de las variables según el F score (modelo xgboost sin binarias)

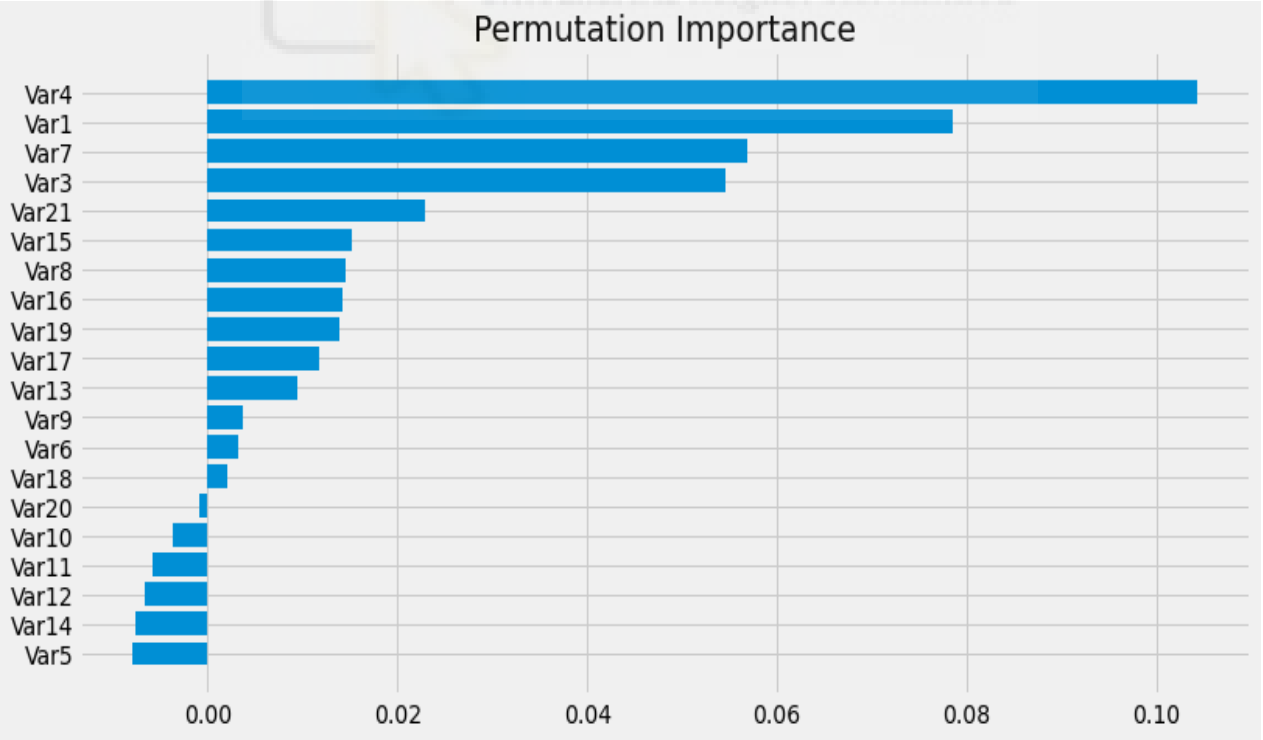

Figura 10, gráfico importancia de las variables según la permutación (modelo xgboost sin binarias)

Y el primer árbol que usaría sería:

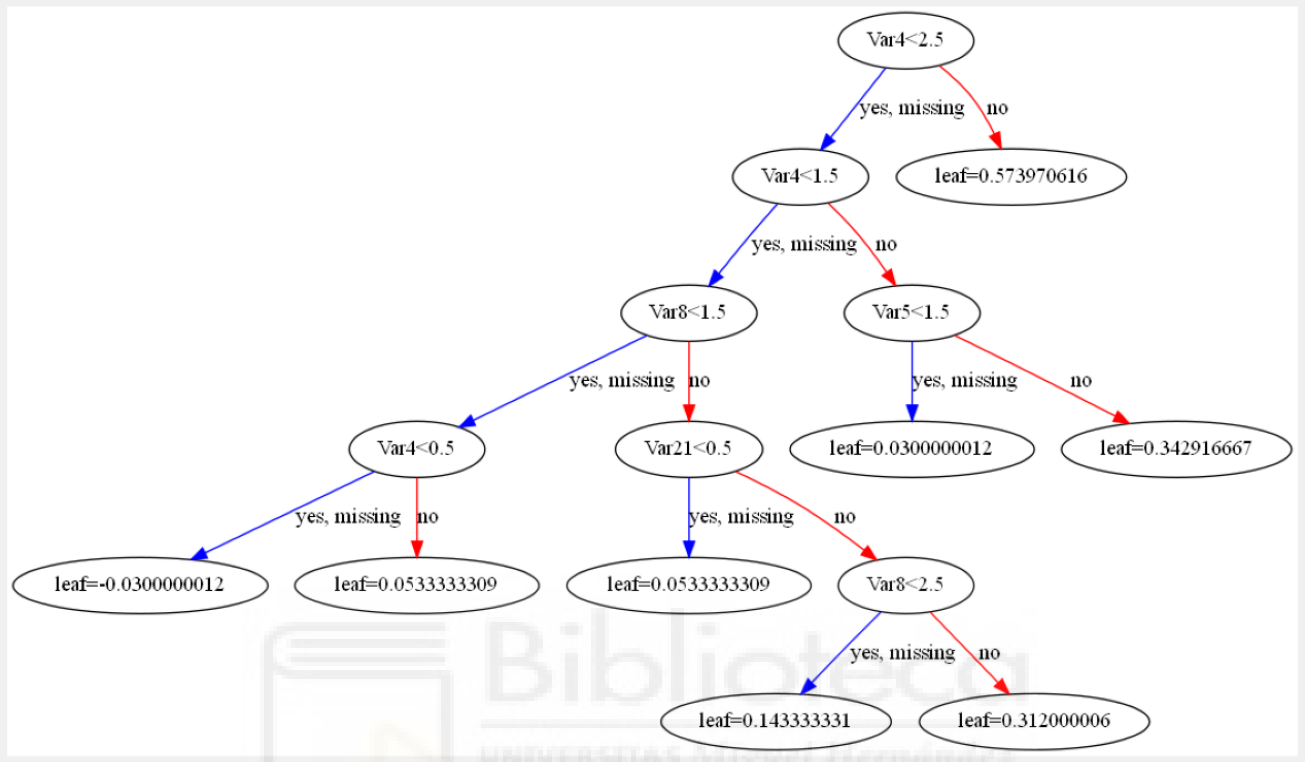

Figura 11, ejemplo árbol modelo xgboost sin binarias

El primer gráfico indica la importancia de cada variable según las veces que aparecen en los árboles de decisión, el segundo gráfico permuta las variables y crea modelos sin esa variable y compara ese nuevo modelo con el que tiene todas las variables, esto permite saber también qué importancia tiene cada variable en nuestro modelo.

Si comparamos ambos gráficos se ve como hay una diferencia de pesos de las variables según el sistema que se use, la puntuación f que mide la aparición en árboles de decisión le da más pesos a la variable 7, sin embargo, la mayor puntuación en el método de permutaciones es en la variable 4.

El método con binarias da estos resultados:

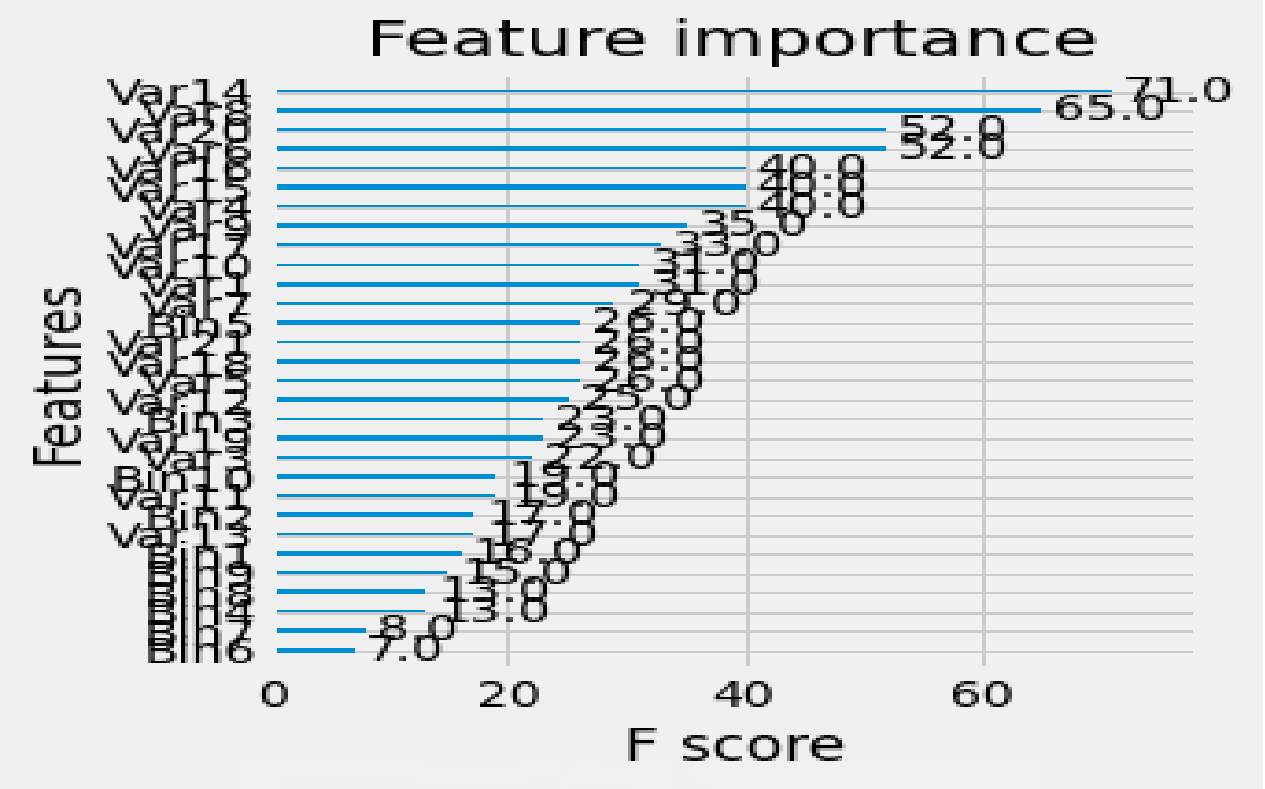

Figura 12, gráfico importancia de las variables según el F score (modelo xgboost con binarias)

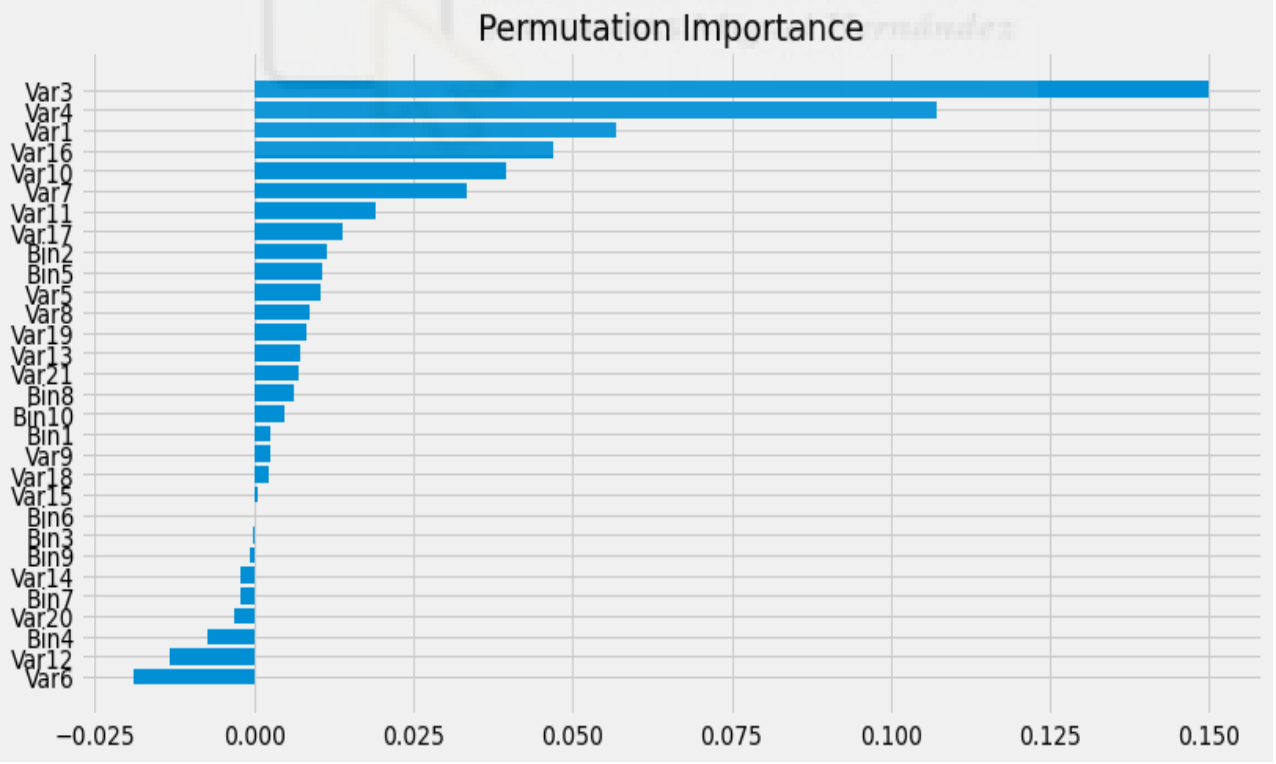

Figura 13, gráfico importancia de las variables según la permutación (modelo xgboost con binarias)

Y el primer árbol de decisión sería:

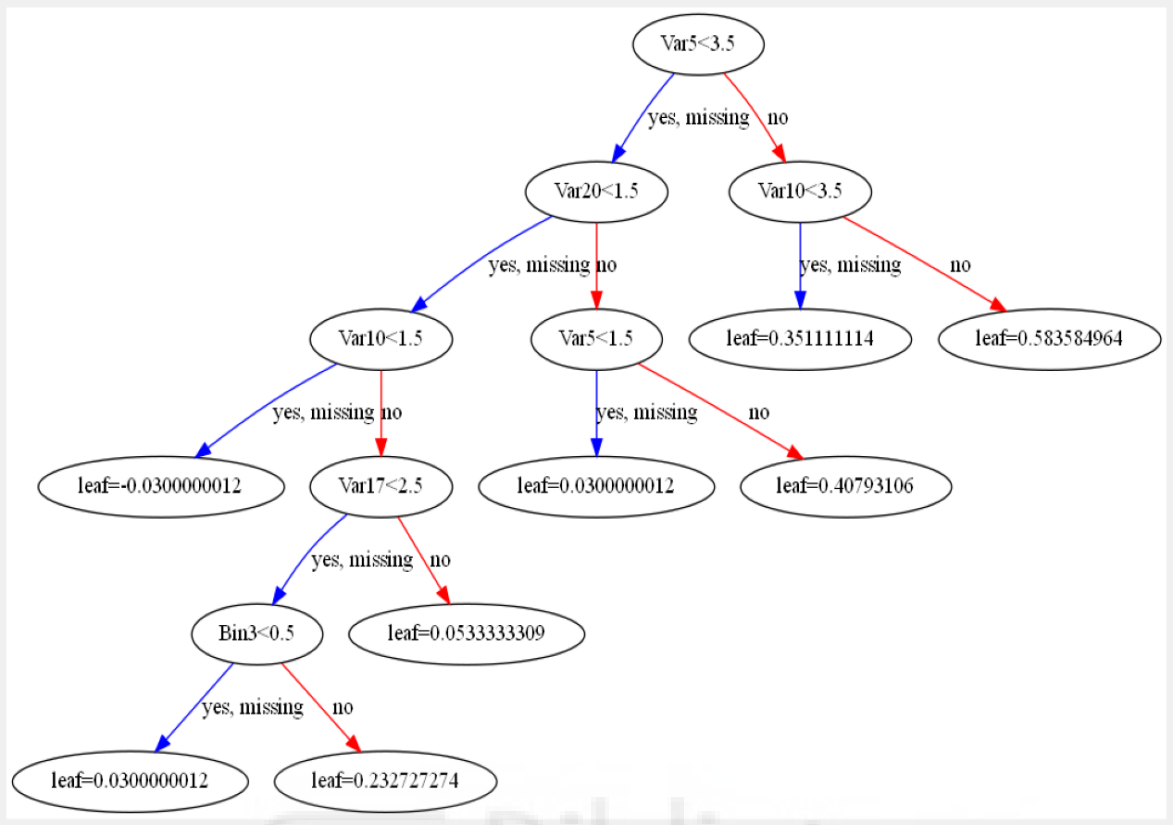

Figura 14, ejemplo árbol modelo xgboost con binarias

En este caso las variables que tienen más importancia son la variable 14 para la puntuación f y la variable 3 para el método de permutaciones.

El segundo método de predicción que se usa es el de Random Forest[1], en este algoritmo se usan varios árboles de decisión y se combinan el resultado final de estos para obtener un mejor resultado que el que se obtiene si se usa un solo árbol, cada árbol se entrena con un subconjunto de la base de datos distinto para obtener distintos resultados.

En este caso se comparará la importancia entre variables entre el método de permutación [6] y el coeficiente de Gini [7] que mide qué variables reducen la impureza del modelo. El modelo sin binarias obtendría los siguientes gráficos:

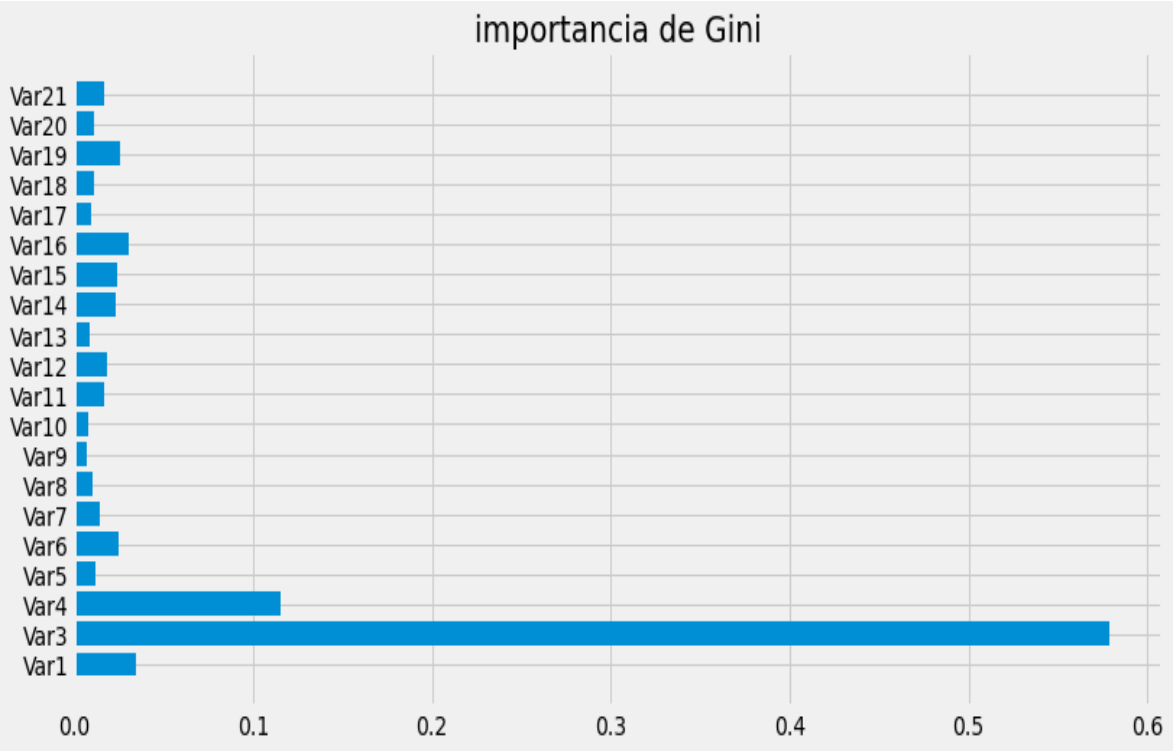

Figura 15, gráfico importancia de las variables según la importancia de Gini (modelo Random Forest sin binarias)

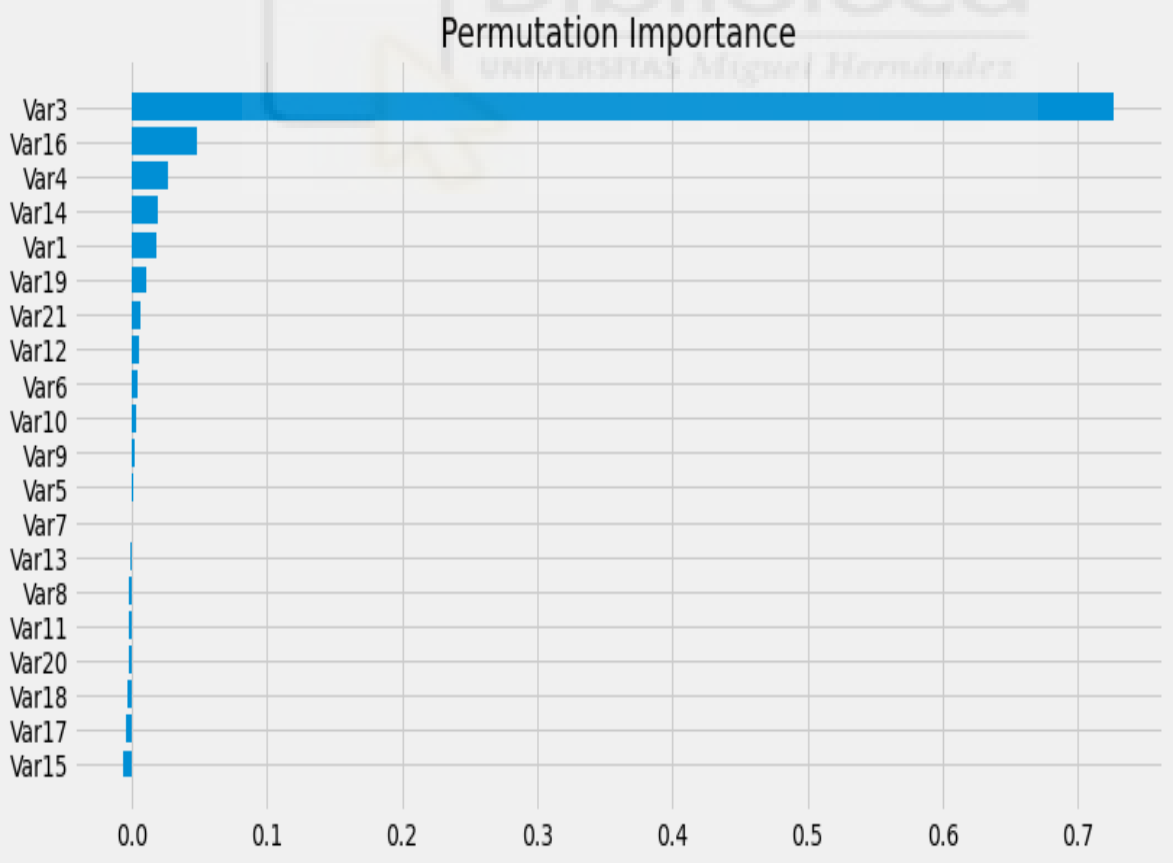

Figura 16, gráfico importancia de las variables según la importancia de la permutación (modelo Random Forest sin binarias)

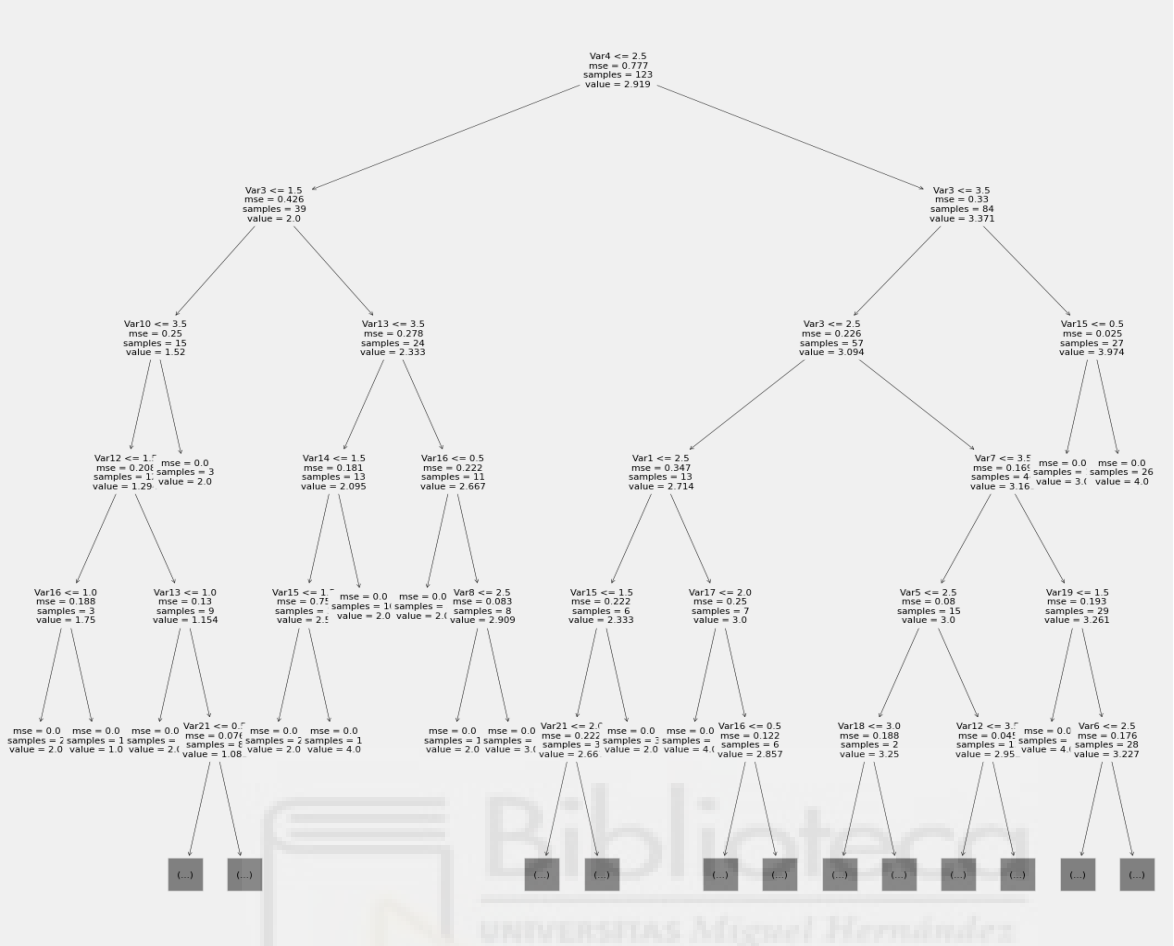

Figura 17, ejemplo árbol modelo Random Forest sin binarias

En este caso ambos métodos le dan mucha importancia a la variable 3 esto implica que en este modelo, esta variable tendría mucho más peso que el resto a la hora de predecir nuestra variable dependiente y habría que vigilar especialmente.

Por último el modelo Random Forest con variables binarias mostraría los siguientes gráficos:

Figura 18, gráfico importancia de las variables según la importancia de Gini (modelo Random Forest con binarias)

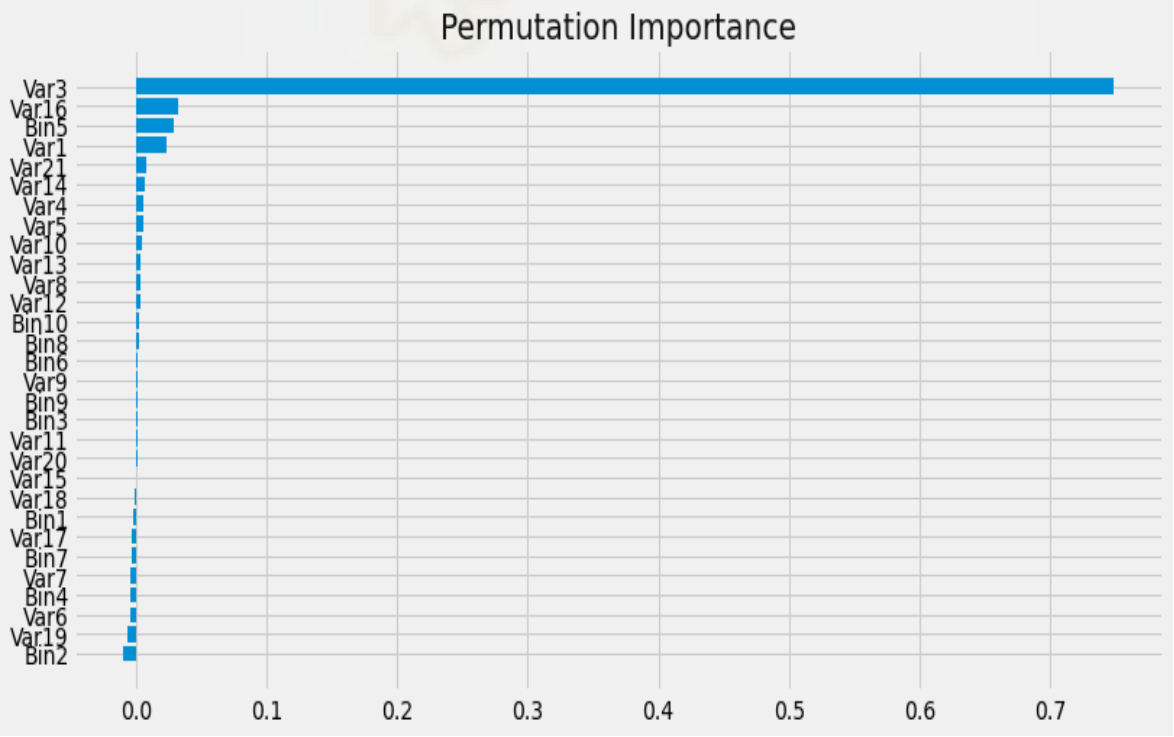

Figura 19, gráfico importancia de las variables según la importancia de la permutación (modelo Random Forest con binarias)

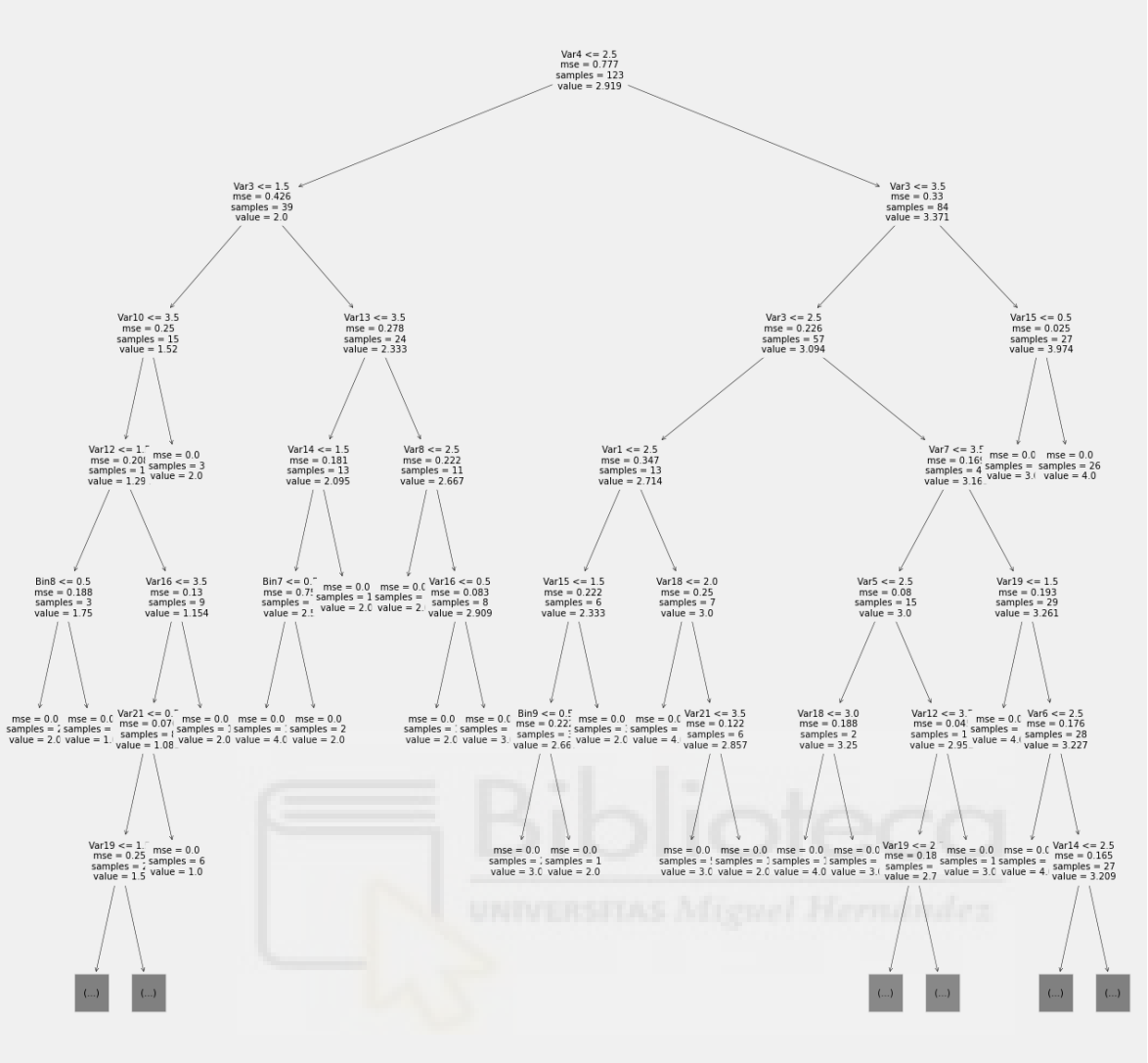

Figura 20, ejemplo árbol modelo Random Forest con binarias

En este caso ambos criterios de importancia siguen dando más importancia a la misma variable.

Aparte de observar la importancia que le da cada modelo a cada variable también se quiere saber qué modelo predice mejor los datos, para ello se ha pedido el error cuadrático medio y los resultados han sido los siguientes:

RandomForest con binarias: 0.582
RandomForest sin binarias: 0.563
Xgboost con binarias: 0.685
Xgboost sin binarias 0.644

Se observa que generalmente para estos datos funciona mejor el modelo de RandomForest y que si toca elegir el que tiene o no tiene variables binarias es mejor con el que no tiene binarias al tener el error cuadrático medio más bajo de los 4, las variables binarias suponen un sobreajuste del modelo.

Capítulo 5 Conclusiones y trabajo futuro

5.1.- CONCLUSIONES

Como conclusiones se observa que los objetivos del TFG se han cumplido, se han comparado los distintos centros encuestados que a pesar de que todos tenían muy buenas notas, se puede observar en qué aspectos deben mejorar según su puntuación en las dimensiones, se han conseguido comparar distintos métodos de machine learning para observar cuáles funcionan mejor en este tipo de datos, se ha podido crear distintos clusters para diferenciar entre los centros encuestados, se ha podido ver el peso de las variables en los distintos modelos predictivos y se han podido comparar entre ellos usando los errores cuadráticos medios de las predicciones.

En conclusión, se puede decir que los métodos usados serán útiles para los centros para ver en qué aspectos cambiar, de cara a aplicar futuras mejoras en su funcionamiento.

Aunque es cierto que se requieren recoger muchas más encuestas para tener resultados más fiables este TFG analiza qué métodos pueden ser más eficaces en el caso de que posteriormente con una mayor recogida de datos se apliquen las técnicas usadas en este TFG.

5.2.- POSIBLES DESARROLLOS FUTUROS

Como he comentado anteriormente sería muy útil recoger más encuestas de los centros especialmente en los centros que solo tienen una o dos encuestas registradas, en estos casos el análisis no dispone de las muestras suficientes para llegar a conclusiones fiables.

También sería interesante recoger otras variables de los profesionales que trabajan en el centro, variables como qué metodología usan en cada caso o que aspectos observan ellos más beneficiosos para los niños.

Otro aspecto que se debe mejorar es la escala de Likert de la encuesta que se realiza ya que solo le permite al encuestado responder en la mayoría de casos entre 5 respuestas, una escala del 0 al 10 daría más libertad al encuestado a la hora de plasmar en la encuesta su opinión sobre los aspectos preguntados Con una mayor recogida de datos, mejorando la calidad de las encuestas realizadas y con las variables adecuadas el modelo predictivo y de clasificación darían un salto cualitativo significante que permitiría llegar a mejores conclusiones.

BIBLIOGRAFÍA

- [1] Leo Breiman. Random forests. Machine learning, 2001.

- [2] Introduction to Machine Learning with Python: A Guide for Data Scientists, Andreas C. Müller y Sarah Guido, 2016

- [3] Application of ant K-means on clustering analysis, R. J. Kuo, H. S. WANG, TUNG-LAI HU AND S. H. CHOU ,2005

- [4] Principal Component Analysis,Ian Jolliffe , 2005

- [5] Issues and recommendations for exploratory factor analysis and principal component analysis, James B. Schreiber, 2021

- [6] Correlation and variable importance in random forests, Baptiste Gregorutti, Bertrand Michel y Philippe Saint-Pierre ,2017

- [7] Empirical characterization of random forest variable importance measures, Kellie J.Archer, Ryan V.Kimes, 2008

ANEXO: tablas, gráficos

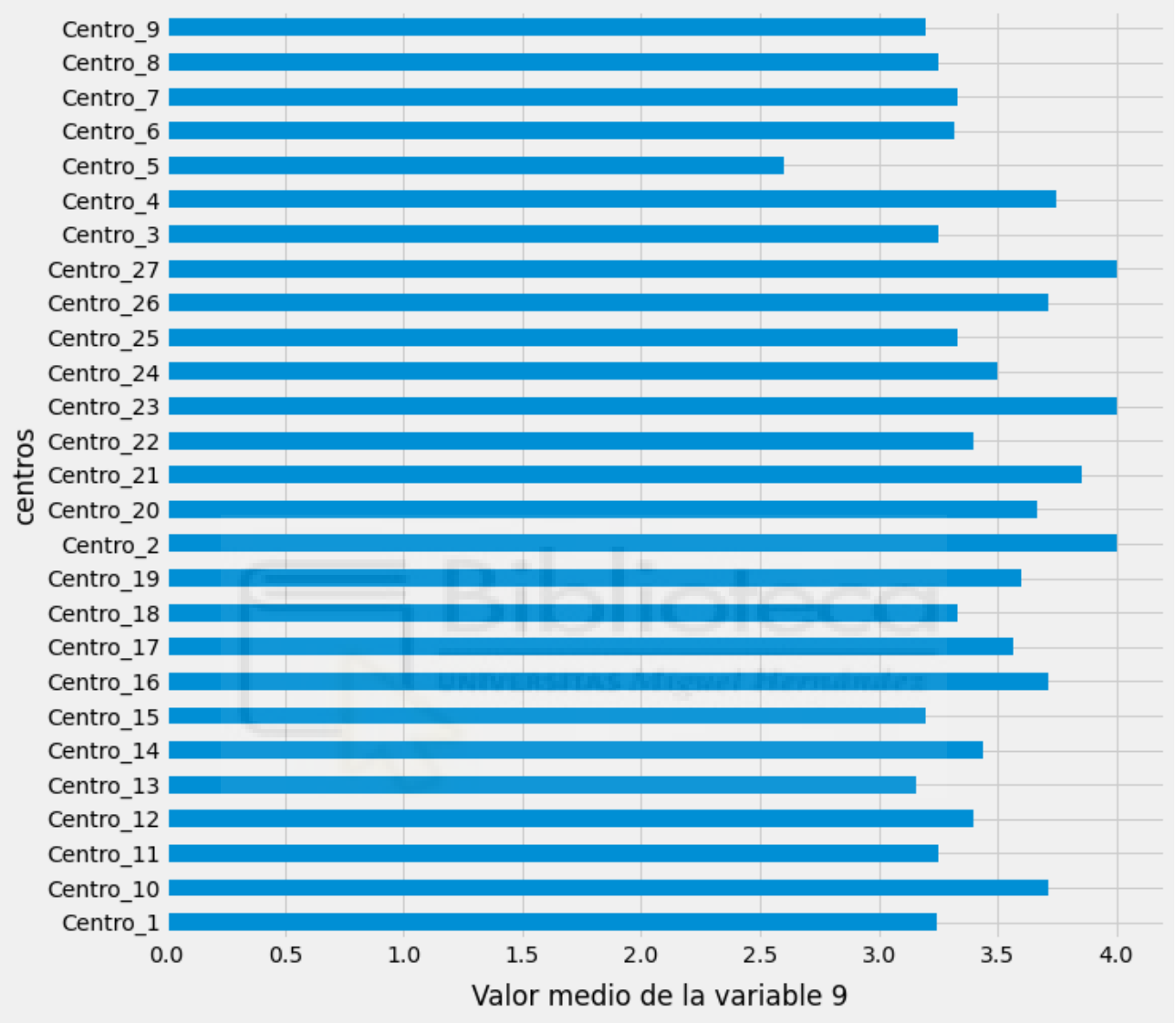

Figura 1 anexo, valor medio de la variable 9 agrupada por centros

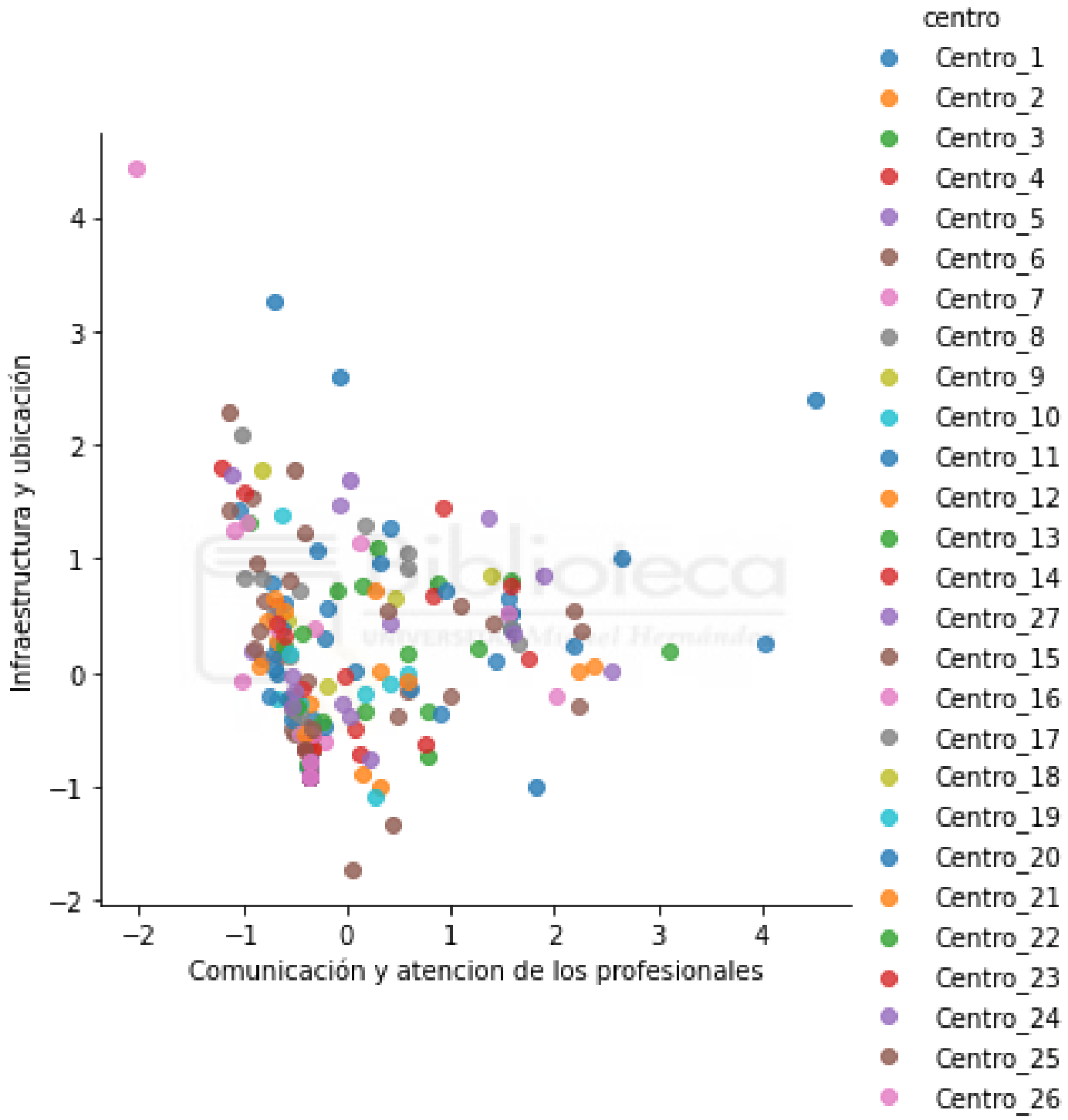

Figura 2 anexo, dimensiones 1 y 2 sin agrupar por centro

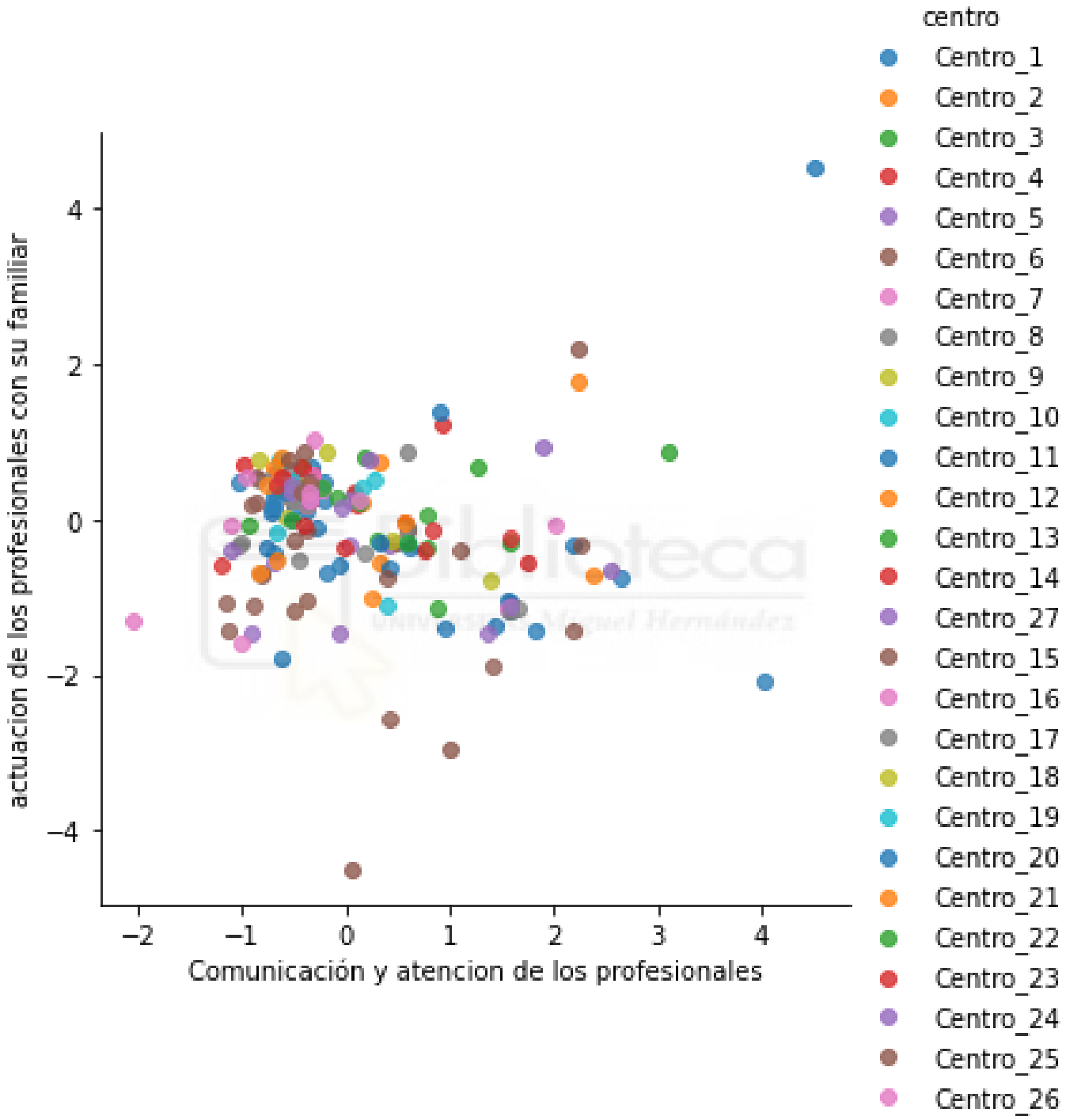

Figura 3 anexo, dimensiones 3 y 1 sin agrupar por centro

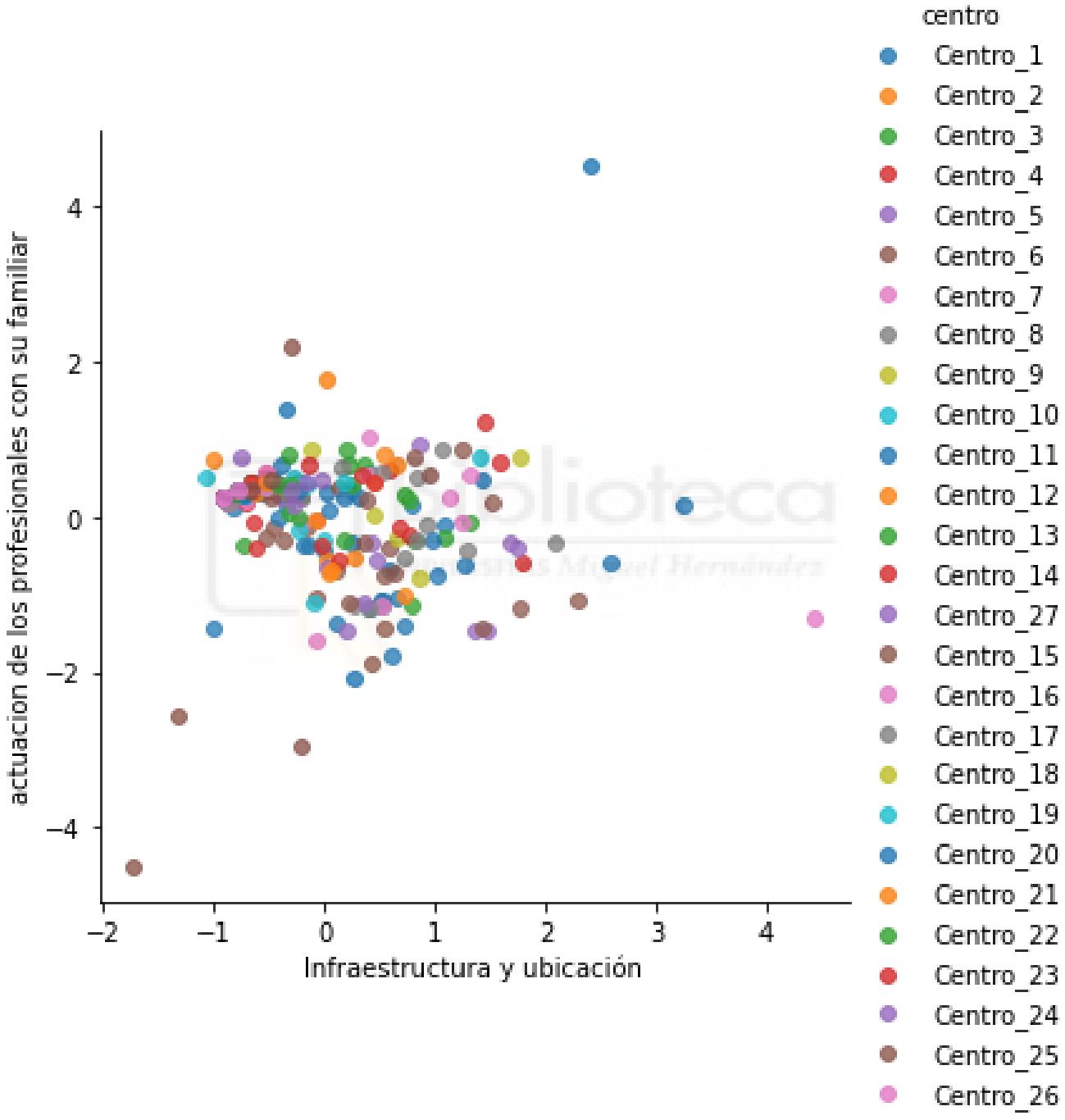

Figura 4 anexo, dimensiones 3 y 2 sin agrupar por centro