



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



DETECCIÓN DE CÁNCER DE VEJIGA MEDIANTE LA MEDICIÓN DE ORINA CON LENGUA ELECTRÓNICA

MÁSTER UNIVERSITARIO EN INGENIERÍA DE ANÁLISIS DE
DATOS, MEJORA DE PROCESOS Y TOMA DE DECISIONES

Septiembre, 2019

DIRECTOR: FERRER RIQUELME, ALBERTO J.
COTUTORA: MARTÍNEZ BISBAL, M. CARMEN
AUTOR: MONREAL TRIGO, JAVIER

ÍNDICE

Resumen.....	3
1. Introducción	3
2. Objetivo	4
3. Descripción de la base de datos	4
4. Comparación con detección con cáncer de próstata	6
5. Materiales y métodos.....	8
5.1. Herramientas <i>software</i>	8
5.2. Centrado y normalización	8
5.3. Estructura N-dimensional de los datos	9
5.4. Pretratamientos	9
5.4.1. Autoescalado [P1].....	10
5.4.2. Centrado en las observaciones	11
5.4.3. Escalado en los sensores (electrodos) [P2]	11
5.4.4. Escalado en los parámetros (puntos de corriente) [P3]	11
5.4.5. Escalado en sensores y parámetros [P4]	11
5.5. Obtención de variables adicionales	11
5.6. Metodología.....	13
5.6.1. Área bajo la curva ROC	13
5.6.2. Sensibilidad y especificidad	14
5.6.3. Escalado pareto por bloques	15
5.6.4. Clasificador binomial	16
5.6.5. ANOVA e Intervalos de <i>Tukey</i>	24
6. Comparación de modelos.....	24
6.1. Clasificación binomial sin variables adicionales	25
6.1.0. Modelo base: PLSDA.....	25
6.1.1. Modelos basados en PCR y PFR	26
6.1.2. Modelos basados en SVM.....	26
6.1.3. Modelos basados en sNPLSDA.....	27
6.1.4. Modelos basados en ANN.....	27
6.1.5. Modelos ensemble	28
6.1.6. Modelo final.....	32
6.2. Clasificación binomial con variables adicionales.....	33
6.2.0. Modelo base: PLSDA.....	33
6.2.1. Modelos basados en PCR y PFR	34

6.2.2. Modelos basados en SVM.....	34
6.2.3. Modelos basados en sNPLSDA.....	35
6.2.4. Modelos basados en ANN.....	35
6.2.5. Modelos ensemble	36
6.2.6. Modelo final.....	40
7. Conclusiones.....	40
8. Trabajos futuros	41
9. Recursos informáticos utilizados	42
10. Bibliografía.....	43
Anexo: Sin variables adicionales.....	45
Anexo: con variables adicionales.....	89

RESUMEN

Los métodos actuales de detección temprana de cáncer de vejiga presentan baja sensibilidad para la detección de tumores de bajo grado, son invasivos, de coste elevado y operador dependientes. En este proyecto se desarrollan distintos modelos clasificadores basados en el análisis de orina mediante lengua electrónica voltamétrica de cuatro electrodos nobles y cuatro electrodos no-nobles como una potencial herramienta de apoyo al diagnóstico no invasivo de cáncer de vejiga. Para la clasificación los resultados arrojados por un PLSDA son insuficientes, por lo que se exploran técnicas de clasificación más avanzadas. Para ello, de los 966 puntos de corriente medidos por muestra y electrodo, se obtienen cuatro estadísticos adicionales: derivadas locales de las formas de onda, integración de los pulsos de la misma, componentes frecuenciales discretas, y corrientes iniciales y finales que corresponden a la corriente mínima generada respecto al potencial de reposo del sistema electroquímico. Adicionalmente, se prueban distintos pretratamientos. Los modelos clasificadores desarrollados se basan en regresiones lineales, máquinas de soporte vectorial (SVM) y redes neuronales artificiales (ANN) con y sin reducción de la dimensión (mediante PCA y PFA), además de sNPLSDA. Finalmente, se desarrollan modelos conjuntos combinando las técnicas mencionadas, propias de distintas disciplinas (minería de datos, análisis multivariante, análisis N-dimensional y redes neuronales artificiales), obteniéndose una tasa de acierto binomial sin variables adicionales del $77.3\% \pm 5.8\%$ y del $74.4\% \pm 8.0\%$ con ellas.

1. INTRODUCCIÓN

La cistoscopia y la citología urinarias son las pruebas realizadas en la rutina clínica para el diagnóstico y el seguimiento del cáncer de vejiga [1]. Sin embargo, la citología urinaria tiene baja sensibilidad en la detección de tumores de grado bajo, y la cistoscopia es invasiva, de coste elevado y su resultado es operador-dependiente [1] y presenta limitaciones en la diferenciación entre inflamación y lesiones malignas, y en el diagnóstico del *carcinoma in situ*, que es un tumor considerado de alto riesgo de progresión [2]. En este contexto, resulta de gran interés el desarrollo de estrategias no cruentas para la mejora del diagnóstico precoz y el seguimiento. En la búsqueda de biomarcadores para cáncer de vejiga, la orina resulta un candidato ideal ya que recoge células y metabolitos liberados de la vejiga y las lesiones con las que está en contacto [3]. Existen varios ensayos que han recibido la aprobación de la FDA para su uso en el diagnóstico y seguimiento del cáncer de vejiga, pero ninguno de ellos mejora los resultados de la cistoscopia [4]. En el presente trabajo se propone el diseño de un modelo clasificador para un sistema de detección temprana y no invasiva basado en el análisis de la orina mediante lengua electrónica voltamétrica. El procesamiento de las señales puede realizarse inmediatamente mediante un ordenador personal, obteniendo resultados instantáneos, reduciendo el tiempo de espera.

La lengua electrónica es un sensor avanzado que consta de un conjunto de electrodos de distintos materiales, a los cuales se les aplica un tren de pulsos de voltaje (existen otras técnicas de estimulación) y mide las corrientes generadas en una disolución (muestra). Las corrientes generadas varían dependiendo del material que compone cada electrodo, del potencial aplicado en el instante del pulso y de las concentraciones de los distintos compuestos, teniendo sensibilidad cruzada entre electrodos. Principalmente, cada corriente generada se puede explicar según la contribución de tres componentes: la difusión de cargas por atracción/repulsión electrostática, las componentes lineales por las propiedades físicas del sistema eléctrico (i.e. su capacitancia, inductancia y resistividad equivalentes) y las reacciones oxidación-reducción que puede que tengan lugar en la interface entre el electrodo y la disolución. En el proyecto de investigación del que se han obtenido los datos se han realizado medidas de orina obtenida antes de la cirugía a pacientes que padecían de cáncer de vejiga en distinto grado mediante una lengua electrónica (FRAPlus, desarrollada en el Instituto Interuniversitario de Investigación de Reconocimiento

Molecular y Desarrollo Electrónico, IDM) con cuatro electrodos de metales nobles y cuatro de metales no-nobles.

A un conjunto de estos pacientes se les tomó muestra de orina después de la cirugía, asumiendo estas muestras como orinas control, de pacientes libres de cáncer, y en aquellos casos en los que el paciente presentó recurrencia (aparición de lesiones tras la cirugía) de recaer se tomó muestras de orina antes de la operación, y tras la recuperación del paciente de la misma.

En primera instancia, es lógico pensar que las células tumorales pueden modificar el perfil metabólico de la orina, puesto que se encuentran cerca de la misma, y debido a la alta sensibilidad cruzada de la lengua electrónica, es posible que alguna de las sustancias químicas, sea los metabolitos en los que se focaliza el estudio citológico o no, varíen en su concentración y modifique el comportamiento de las corrientes generadas. Los trabajos publicados recientemente del mismo grupo de investigación con muestras similares a las estudiadas en este proyecto, en los que se observan cambios en el perfil metabólico de las orinas mediante técnicas de análisis masivo como son la espectrometría de masas y la espectroscopía de resonancia magnética nuclear confirman estas hipótesis [5-7].

A continuación, se enuncian los principales objetivos del proyecto, se describe la base de datos y se compara con el antecedente de *detección de cáncer de próstata*, utilizando la misma tecnología [8]. Seguidamente, se describen los estadísticos adicionales obtenidos, los pretratamientos utilizados y la metodología en la construcción de modelos. El establecimiento de la base de referencia a batir viene dado por un PLSDA (*Partial Least-Squares-Discriminant Analysis*). El siguiente paso es la construcción de modelos, que para la clasificación binomial se realiza en primer lugar sin variables adicionales, y luego con ellas. Finalmente se apuesta por la integración de modelos, mejorándose los resultados de los modelos clasificadores por separado, siendo muchos de ellos superiores al PLSDA establecido como base de referencia.

2. OBJETIVO

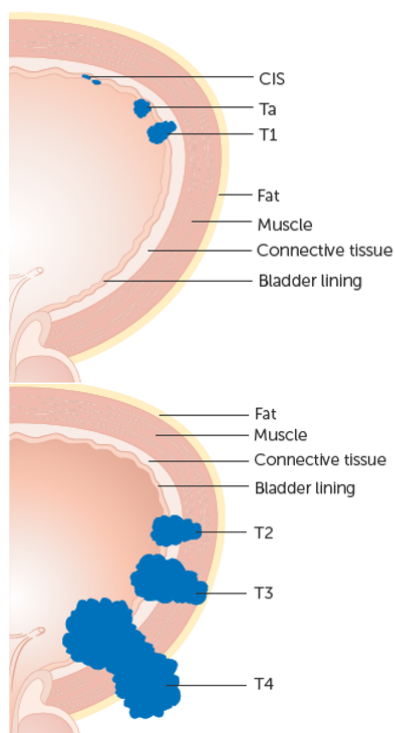
El objetivo principal es el desarrollo de un algoritmo clasificador que sea capaz de, dado un conjunto de datos correspondiente a la medida de la lengua electrónica, determinar si el paciente del que se ha tomado la muestra de orina tiene cáncer de vejiga.

3. DESCRIPCIÓN DE LA BASE DE DATOS

Se tienen datos de 154 pacientes. De estos se recogen las orinas antes de la cirugía, y de una buena parte de ellos se recoge orina un mes después de la cirugía, sumando un total de 218 muestras de orina. De los mismos, 207 son de los que se tiene información clínica completa. Los 11 restantes presentaban un diagnóstico anatomopatológico incierto, por lo cual se excluyeron del estudio estadístico. En el cáncer de vejiga se distinguen diferentes tipos, atendiendo por una parte a los diferentes grados de diferenciación celular que presenta el tejido tumoral, y por otra parte a los diferentes grados de invasión del tumor hacia las capas de la vejiga, tejido muscular e invasión de órganos cercanos o metástasis. Esta clasificación se encuentra detallada en la *Figura 1*. Atendiendo a las 207 muestras: éstas se pueden clasificar según provengan del paciente antes de la cirugía (paciente con cáncer) o paciente tras la cirugía (paciente sin cáncer).

En el conjunto de datos se encuentra información del tumor primario, y no del número de ellos (de haber varios). A continuación, en la *Figura 1*, se muestra cómo se clasifica el desarrollo tumoral de acuerdo al estadio del tumor primario. Se tienen muestras de orina de pacientes con Tx, Ta, Tis, T1, T2 y T3, siendo los más abundantes Ta (21.74% del total de muestras) y T2 (29.95%), seguidas de T3 (5.32%). Se dispone

de únicamente una muestra de Tx y T1 (0.48% cada una) y dos muestras de Tis (0.97%). Todas estas muestras componen un 58.94% del total, correspondiendo el resto (41.06%) a las muestras de orina de pacientes sin cáncer. Esta información se encuentra resumida en la *Tabla 1*.



T hace referencia al estadio del tumor primario.

- **Tx**: No puede valorarse.
- **T0**: No invasivo.
 - **Tis**: Carcinoma preinvasivo, *in situ*, plano.
 - **Ta**: Carcinoma papilar preinvasivo.
- **T1**: Masa móvil inapreciable tras la resección transuretral completa.
- **T2**: Induración de la pared vesical móvil. Invasión microscópica de la capa muscular superficial. No existe induración residual tras la resección transuretral.
- **T3**: Persiste la induración tras la resección.
- **T4**: Tumor fijo e invasor de las estructuras vecinas.

Figura 1. Estadío del cáncer de vejiga de acuerdo al desarrollo del tumor primario. Fuente: *Cancer Research UK (2019)*.

Tabla 1. Número de pacientes (con tipos) y controles.

TIPO DE MUESTRA	NÚMERO
Sanas	85
Cáncer	122
Tx	1
Tis	2
Ta	45
T1	1
T2	62
T3	11

Las muestras corresponden a pacientes con tumores de vejiga antes de la operación (grupo con cáncer), y tras la misma (pacientes libres de cáncer, grupo control).

Cada muestra consta de las medidas de ocho electrodos, que se agrupan de acuerdo a sus propiedades en nobles (N): oro (Au), plata (Ag), platino (Pt) y cobalto (Co); y en no nobles (NN): cobre (Cu), níquel (Ni), rodio (Rh) e iridio (Ir). A los ocho se les aplicó un tren de voltajes y se midieron 966 puntos de corriente por electrodo. El tren de voltajes consta de 48 pulsos, de distinta amplitud, que siguen un orden ascendente hasta su cénit en el primer cuarto de la forma de onda, desciende hasta el punto mínimo de

la forma de onda en el tercer cuarto, y volviendo a reposo en el último tramo. Cada pulso se aplica durante la misma duración, seguido del mismo período con potencial cero. Por cada aplicación de pulso o vuelta al potencial cero, se induce la corriente medida, por lo que se trata de una función a 96 tramos.

Cada corriente generada se puede explicar por tres componentes: la difusión de cargas por atracción/repulsión electrostática, las componentes lineales por las propiedades físicas del sistema eléctrico (i.e. su capacitancia, inductancia y resistividad equivalentes) y las reacciones oxidación-reducción que puede que tengan lugar en la interface entre el electrodo y la disolución, que por tanto están relacionadas con los electrodos y metabolitos presentes en el medio. Como la aplicación de voltaje y medición de corriente en los electrodos se realiza de forma simultánea, en tiempo real se puede describir el funcionamiento de la lengua electrónica como ocho funciones univariantes, una por cada uno de los electrodos, con 966 puntos de corriente cada uno. De esta forma, los datos pueden estructurarse tridimensionalmente: en muestras, puntos de corriente, y electrodos.

4. COMPARACIÓN CON DETECCIÓN CON CÁNCER DE PRÓSTATA

El principal antecedente a este trabajo con el uso de lengua electrónica en la orina es la detección de cáncer de próstata mediante la medición de muestras de orina con una lengua electrónica voltamétrica

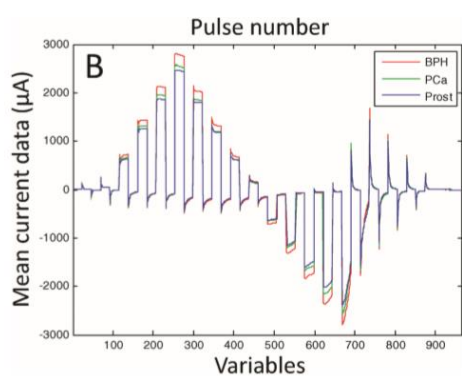


Figura 2. Forma de onda de corriente promedio por clase en el electrodo de Ag para la detección de cáncer de vejiga. Fuente: Pascual, L. et al. (2016).

[8], en el que se obtienen resultados con un 91% de sensibilidad y 73% de especificidad. Si bien la metodología de análisis de las muestras es similar, el contexto clínico es diferente en una y otra patología. Los cambios metabólicos en uno y otro caso pueden ser bien diferentes, ya que en el caso del cáncer de próstata se trata del desarrollo de lesiones tumorales en una glándula que secreta su contenido a la orina, y en el caso del cáncer de vejiga son lesiones presentes en la pared de la vejiga, que desprenden sustancias y células, y que entran en contacto con la orina por un tiempo indeterminado. Igualmente, existen diferencias obvias entre la población en estudio en uno y otro caso, ya que el cáncer de próstata es una enfermedad que afecta a varones, mientras que el cáncer de vejiga afecta a hombres y mujeres.

Las formas de onda de corriente del electrodo de plata (Ag) promedias para cada grupo de pacientes (PCa: cáncer de próstata, verde; BPH: hiperplasia benigna de próstata, rojo; Prost: orina obtenida tras eliminación de la próstata mediante cirugía, azul) en el estudio de cáncer de próstata [8] se muestran en la Figura 2.

Promediando las formas de onda de las muestras de orina del presente estudio (muestras con cáncer de vejiga, frente las que no presentan células tumorales obtenidas tras la cirugía), se obtienen unos resultados similares (véase la Figura 3). Destaca en la misma la diferencia presente del cobalto (en la Figura 3 cada agrupación de pulsos positivos seguidos de una agrupación de pulsos negativos conforma la forma de onda registrada de un electrodo, siendo la sexta mostrada la del cobalto), en el que destacan mayores picos de corriente en pacientes sanos que en pacientes enfermos. Esto se ve con mayor claridad en la Figura 4, en detalle el electrodo de Co.

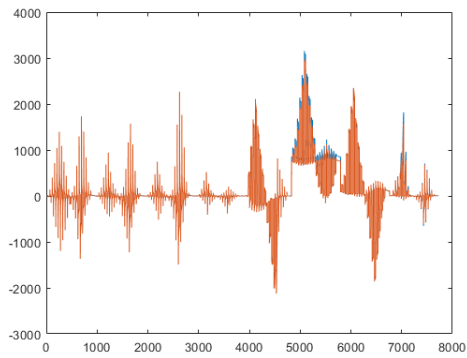


Figura 3. Promediado de las corrientes por enfermos (rojo) y sanos (azul) de los ocho electrodos.

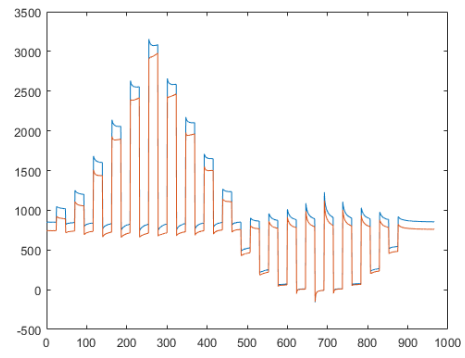


Figura 4. Promediado de formas de onda de corriente por enfermos (rojo) y sanos (azul) del electrodo de Co.

Sin embargo, esto no es algo aplicable como criterio clasificador, las formas de onda de corriente no tienen mayor amplitud en el caso de pacientes sanos. En las *Figuras 5, 6 y 7* se muestran distintos ejemplos en los que se comparan muestras sanas con enfermas, donde puede comprobarse que esta regla observada con el promedio no es generalizable.

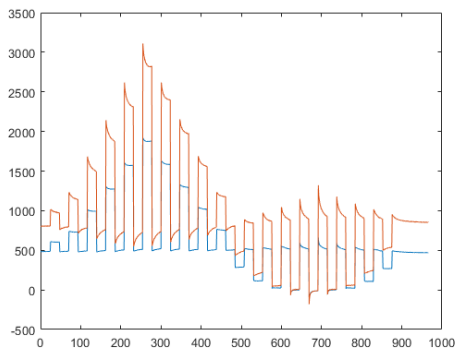


Figura 5. Formas de onda de corriente por enfermos (rojo) y sanos (azul) del electrodo de Co [Ejemplo 1].

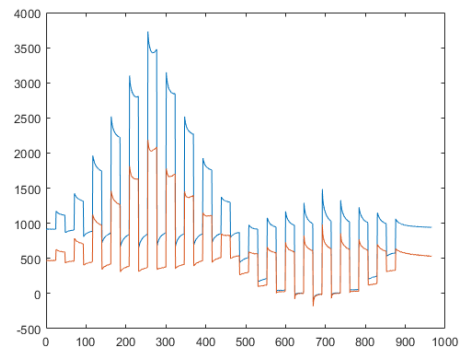


Figura 6. Formas de onda de corriente por enfermos (rojo) y sanos (azul) del electrodo de Co [Ejemplo 2].

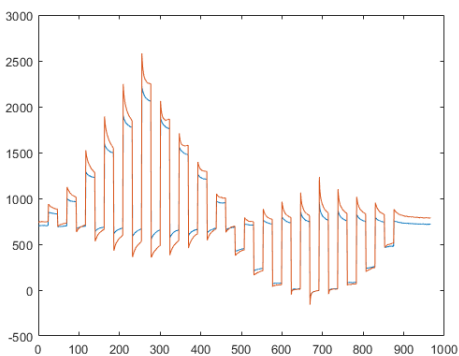


Figura 7. Formas de onda de corriente por enfermos (rojo) y sanos (azul) del electrodo de Co [Ejemplo 3].

Se hace evidente la necesidad de identificar otros patrones que permitan la clasificación, más allá del valor promedio de cada uno de los puntos de corriente por grupo. A continuación, en la *Figura 8*, se muestran superpuestas todas las formas de onda, correspondientes cada una a un paciente. La variabilidad es especialmente notoria en los cuatro electrodos no-nobles, pero no se observan agrupaciones de variabilidad (de acuerdo a los grupos sanos y con cáncer), lo que indica que probablemente será necesario la aplicación de técnicas avanzadas.

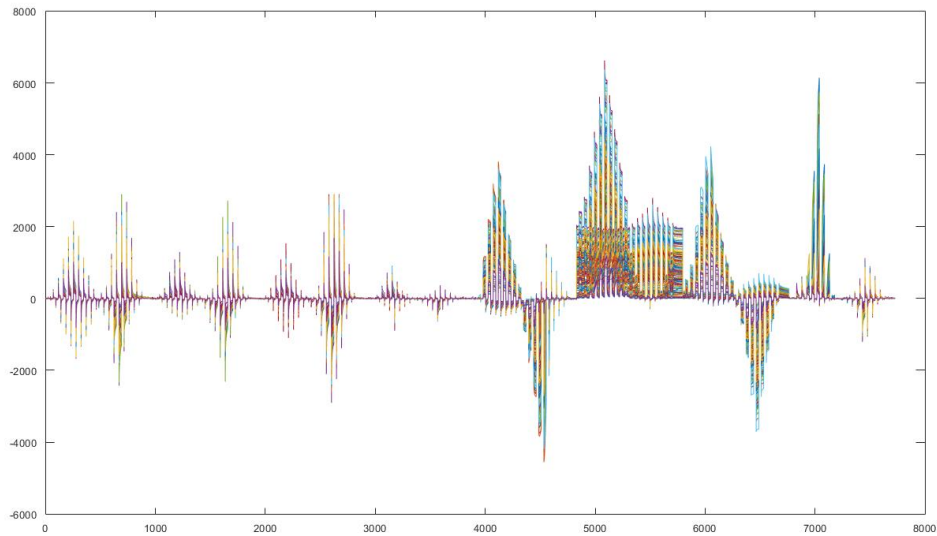


Figura 8. Variabilidad de las formas de onda obtenidas.

5. MATERIALES Y MÉTODOS

En el presente apartado se detallan las herramientas *software* de las que se hace uso, los preprocesamientos utilizados, las variables adicionales obtenidas y la metodología empleada para la construcción de modelos.

5.1. HERRAMIENTAS SOFTWARE

En el presente proyecto se utiliza *MATLAB* para la obtención de estadísticos o variables adicionales. Para el centrado y normalización de las muestras, la aplicación de pretratamientos, el desarrollo de los modelos, atendiendo a la metodología, se utiliza *R*, con los paquetes que se citan en el apartado 9, *Recursos Informáticos Utilizados*.

5.2. CENTRADO Y NORMALIZACIÓN

El centrado y la normalización por muestra son tratamientos básicos previos que permiten dotar a cada una de las formas de onda un peso inicial equivalente, sin modificación de la forma de las mismas, principal portadora potencial de la información relevante. Las ecuaciones aplicadas a cada forma de onda de cada electrodo M , para centrar y normalizar, son las siguientes:

$$M_{center} = M - \bar{M}$$

$$M_{norm\¢er} = \frac{M_{center}}{\max(M_{center}) - \min(M_{center})}$$

A modo de ejemplo se muestra el preprocesamiento en la *Figura 9*:

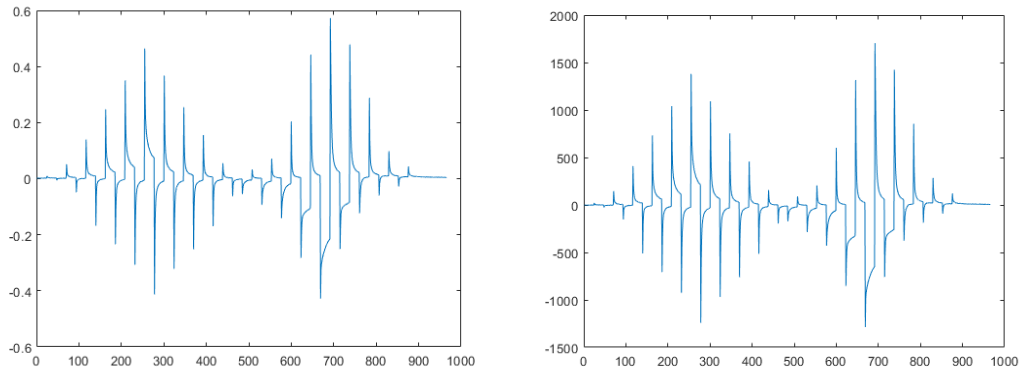


Figura 9. Formas de onda con (izquierda) y sin pretratamiento (derecha).

Este preprocesamiento básico (centrado y normalización) se ha aplicado a todas las formas de onda, independientemente de que se apliquen otros, más avanzados, que se detallan a continuación.

5.3. ESTRUCTURA N-DIMENSIONAL DE LOS DATOS

De acuerdo con lo comentado en el apartado de la *Descripción de la base de datos*, estos pueden estructurarse tridimensionalmente, siendo la primera dimensión o modo las observaciones, las cuales tienen asociadas matrices 8×966 , correspondientes a las ocho formas de onda de corriente, de 966 puntos (mediciones) de corriente cada una, de cada uno de los ocho electrodos que conforman la lengua electrónica. A continuación, en la *Figura 10*, se muestra una representación de la estructura tridimensional de los datos.

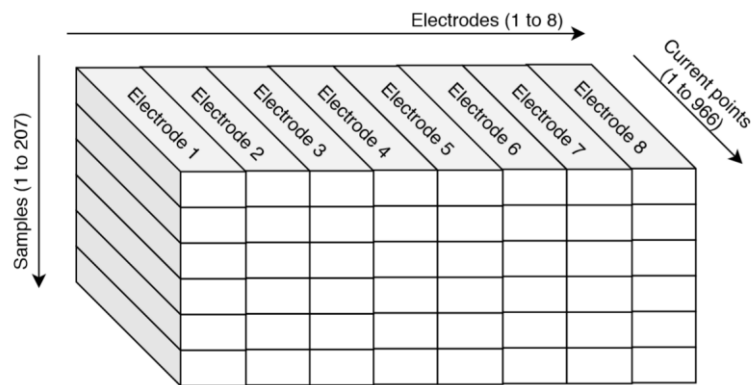


Figura 10. Representación de la estructura tridimensional de los datos.

5.4. PRETRATAMIENTOS

Los pretratamientos, o preprocesados, son funciones aplicadas al conjunto de datos, que adecuan los mismos de acuerdo a la necesidad posterior. La pretensión con los aplicados en este trabajo es poner en relieve la variabilidad asociada a la clasificación entre muestras de orina de pacientes con cáncer y sin él.

Una primera opción es utilizar MSC (*Multiple Scatter Correction*), ampliamente utilizado en química (véase la *Figura 11* como ejemplo de su aplicación en espectroscopia infrarroja en el ámbito alimentario [9]). Sin embargo, este preprocesamiento ofrece resultados nefastos aplicado a nuestro conjunto de datos, tal

como puede observarse en la *Figura 12*, ya que tras el preprocesamiento desaparecen la mayoría de variables.

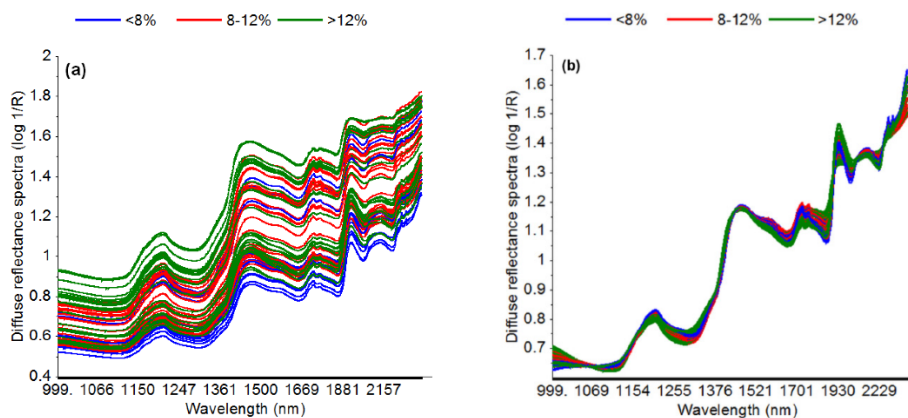


Figura 11. Formas de onda de ejemplo antes (izquierda) y tras (derecha) de MSC. Fuente: A. Adnan et al. (2017).

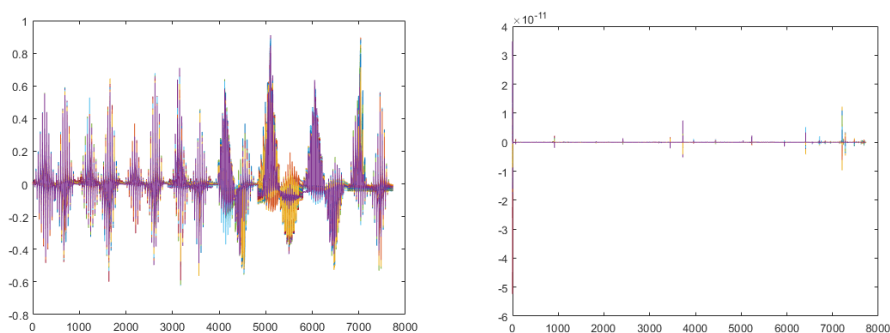


Figura 12. Formas de onda antes (izquierda) y tras (derecha) de MSC.

Como puede observarse, este pretratamiento no es adecuado para este conjunto de datos. Esto es debido a que el MSC está diseñado especialmente para la corrección de espectros, los cuales no presentan valores negativos ni funciones tan marcadamente discontinuas como las formas de onda de corriente presentes en las medidas de orina mediante lengua electrónica voltamétrica de este trabajo.

En la exploración del conjunto de datos y obtención de modelos discriminantes se evaluarán modelos:

- Con únicamente centrado y normalización por muestra [P0]

Y adicionalmente al centrado y normalización:

- Autoescalado (con varianza unitaria) [P1]
- Centrado en las observaciones y escalado en los sensores (electrodos) [P2]
- Centrado en las observaciones y escalado en los parámetros (puntos de corriente) [P3]
- Centrado en las observaciones y escalado en sensores y parámetros [P4]

5.4.1. AUTOESCALADO [P1]

Consiste en el centrado de todas las variables y su especificación con varianza unitaria, para ello:

1. Se obtiene la media de cada una de las variables.
2. Para cada una de las muestras, se sustrae el valor de la media de cada una de las variables respectivamente.
3. Se obtiene la desviación típica de cada una de las variables.
4. Para cada muestra, se divide el valor de cada variable por su desviación típica.

5.4.2. CENTRADO EN LAS OBSERVACIONES

En la estructura tridimensional de los datos, cada una de las dimensiones se conocen como *modos*. En los preprocesamientos P2, P3, y P4, únicamente aplicables ante una estructura multidimensional, se va a realizar un centrado en las observaciones (las muestras, modo uno, siendo los electrodos y los puntos de corriente los modos dos y tres, respectivamente).

1. Se obtiene la matriz de medias, siendo la media de cada punto de corriente por cada electrodo.
2. A cada muestra (que es una matriz de 8x966) se le sustrae la matriz de medias anteriormente calculada.

5.4.3. ESCALADO EN LOS SENSORES (ELECTRODOS) [P2]

Se aplica el centrado en las observaciones y se obtiene la desviación típica de cada uno de los electrodos. Es decir, de las ocho matrices de las 207 observaciones por los 966 puntos de corriente, se obtienen ocho escalares, correspondientes a las desviaciones típicas de cada uno de los electrodos.

A continuación, se dividen los elementos de cada una de las matrices obtenidas por el despliegue en electrodos por su correspondiente desviación típica, igualando la contribución en varianza de cada uno de los electrodos.

5.4.4. ESCALADO EN LOS PARÁMETROS (PUNTOS DE CORRIENTE) [P3]

Se aplica el centrado en las observaciones y se obtiene la desviación típica de cada uno de los puntos de corriente. Es decir, de las 966 matrices de las 207 observaciones por los 8 electrodos, se obtienen 966 escalares, correspondientes a las desviaciones típicas de cada uno de los puntos de corriente.

A continuación, se dividen los elementos de cada una de las matrices obtenidas por el despliegue en puntos de corriente por su correspondiente desviación típica, igualando la contribución en varianza de cada uno de los puntos de corriente.

5.4.5. ESCALADO EN SENSORES Y PARÁMETROS [P4]

Se aplica el centrado en las observaciones y se aplican los dos escalados consecutivamente, de forma que la contribución a la varianza del conjunto de datos es la misma para los ocho electrodos, entre sí, y para los 966 puntos de corriente, entre sí. Aunque se asemeja, no es equivalente al autoescalado.

5.5. OBTENCIÓN DE VARIABLES ADICIONALES

Además de los puntos de corriente, de los mismos pueden extraerse variables adicionales que no son combinación lineal del resto, o en caso de serlo, pueden presentar variabilidad ligada a la presencia o estadio del cáncer de vejiga que, si no se añaden al modelo, pueden encontrarse tapadas por el ruido inherente a cada punto de corriente.

Estas son:

- La pendiente de la curva de corriente. Esta se encuentra asociada al sumatorio de dos o más exponenciales. Por simplicidad, cada curva se ajusta mediante tres rectas locales, como se muestra en la *Figura 13*.
- Las componentes frecuenciales de cada muestra. Se obtienen las primeras 433 frecuencias mediante *Fast Fourier Transform (FFT)*. Véase la *Figura 14*.
- Área bajo las curvas de los pulsos.
- Corriente inicial y final asociada al OCP. El OCP es el potencial inicial del sistema electroquímico cuando no hay circulación de corriente, es el potencial de reposo. Al aplicarse un potencial distinto, se genera una corriente, que es distinta de cero si el potencial no coincide con el OCP. Esta corriente es proporcional a la conductividad de la disolución y a la diferencia entre el potencial aplicado y el OCP. En una frase, se toman el primer y último valor de la muestra sin normalizar ni centrar.

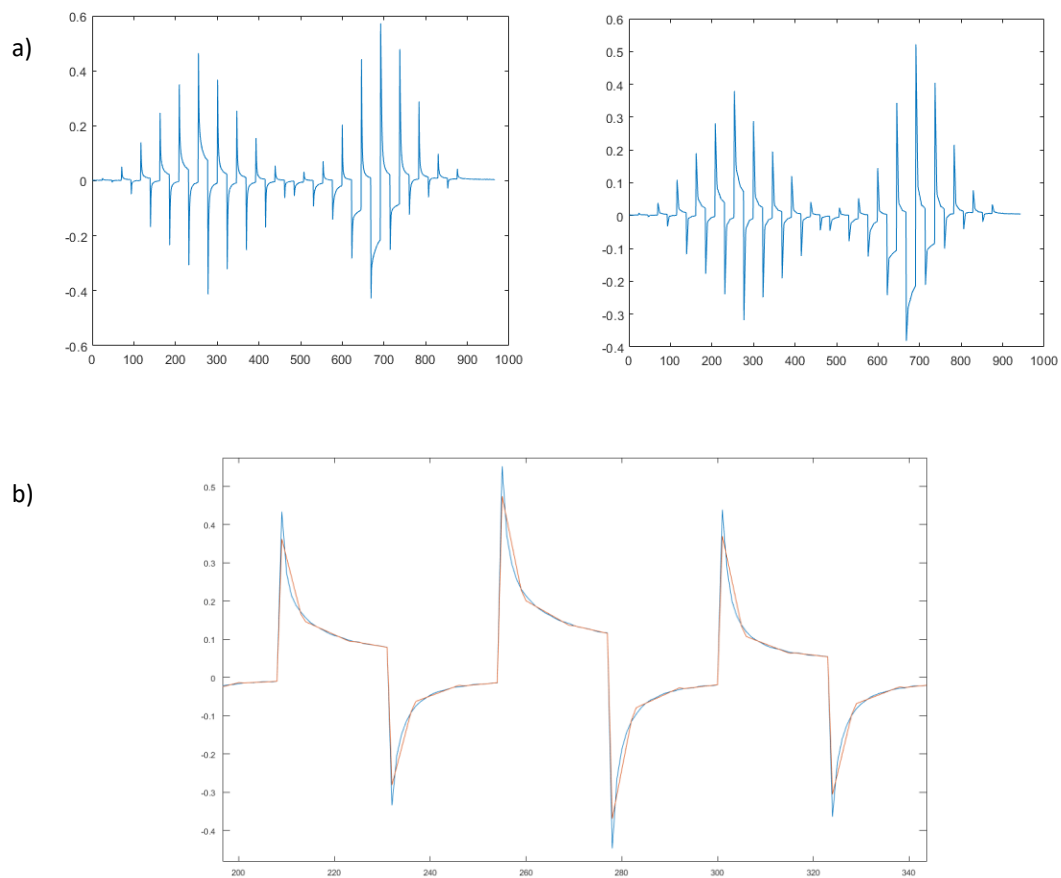


Figura 13. a) Forma de onda original (izquierda) y forma de onda reconstruida mediante pendientes locales (derecha). b) Forma de onda (azul) y reconstruida con tres pendientes locales (rojo).

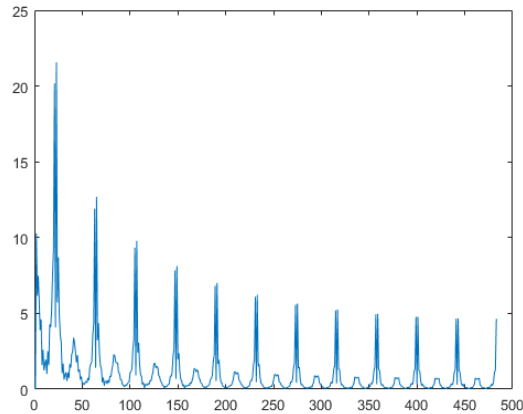


Figura 14. Componentes frecuenciales de la forma de onda anterior.

5.6. METODOLOGÍA

Cada uno de los modelos se construye con validación cruzada a nueve grupos balanceados. Es decir, por cada modelo se construyen realmente nueve, cada uno de ellos utilizando 8/9 de las 207 muestras (184) como entrenamiento, y 1/9 (23) como validación. Estos nueve grupos se construyen inicialmente y son los mismos durante todo el estudio.

Los modelos se comparan de acuerdo al Área bajo la Curva ROC (AUC) (explicada en el apartado 5.6.1.) sobre los conjuntos de validación, obteniendo por tanto nueve estadísticos por modelo. Se considera mejor modelo aquel que presenta un AUC medio (de los nueve) superior, siempre y cuando el inferior de ellos sea superior a 0.5. Esto es debido a que un AUC inferior a 0.5 indica una capacidad de clasificación inferior a un modelo aleatorio. Para su comparación gráfica se acude a gráficos *Box-Whiskers*.

Se tienen cinco bloques de variables:

- Puntos de corriente (7728 variables)
- Componentes frecuenciales (3864 variables)
- Corrientes en OCP (16 variables)
- Áreas bajo cada pulso (320 variables)
- Pendientes locales (960 variables)

Se escalan los bloques de variables mediante escalado Pareto (explicado en el apartado 5.6.3.).

5.6.1. ÁREA BAJO LA CURVA ROC

La curva ROC (acrónimo de *Receiver Operating Characteristic* o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. En la *Figura 15* se muestra un ejemplo.

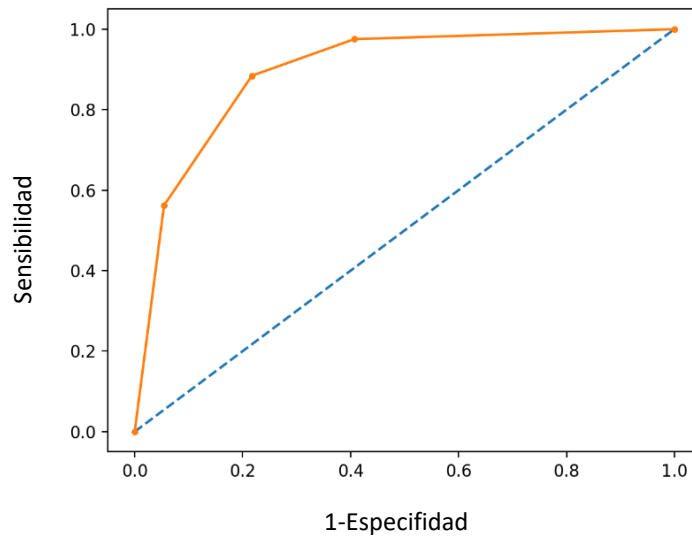


Figura 15. Ejemplo de curva ROC con AUC = 0.9 (naranja) y un clasificador aleatorio (azul). Fuente: (2019, August 19). How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python. Retrieved from <https://machinelearningmastery.com/roc-c>

Un clasificador aleatorio se aproxima a un segmento con pendiente unitaria y origen nulo, siendo su AUC (acrónimo de *Area Under (ROC) Curve* o Área Bajo la Curva (ROC)) de aproximadamente 0.5. Un clasificador perfecto, por su parte, tiene una sensibilidad del 100% para todo el eje de ordenadas, siendo por tanto su AUC unitario. Los AUC inferiores a 0.5 denotan clasificadores binarios peores que un clasificador aleatorio, por lo que un clasificador será mejor a mayor AUC presente, normalmente comprendido entre 0.5 y 1.

5.6.2. SENSIBILIDAD Y ESPECIFICIDAD

En la clasificación binaria se definen dos valores importantes para determinar la calidad del estimador: la sensibilidad y la especificidad. La sensibilidad indica la capacidad del modelo clasificador para acertar en diagnósticos positivos respecto al número total de positivos presentes en la muestra. En este caso, es la capacidad de identificar que una muestra de orina de una paciente con cáncer de vejiga efectivamente lo tiene. Por otro lado, la especificidad es la capacidad de este modelo de acierto en diagnósticos negativos respecto al número total de negativos presentes en la muestra, es decir, la capacidad de identificar las muestras de orina de pacientes sanos como tales [10]. Se una forma más clara se muestra el cálculo de estos dos valores, junto con el gráfico explicativo de la *Figura 16*, donde los elementos (muestras) circunscritos a la elipse son los clasificados como positivos, cuando estos son realmente positivos o negativos dependiendo de si se encuentran a la izquierda o la derecha del rectángulo, respectivamente.

$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}$$

$$\text{Especificidad} = \frac{\text{Verdaderos negativos}}{\text{Verdaderos negativos} + \text{Falsos positivos}}$$

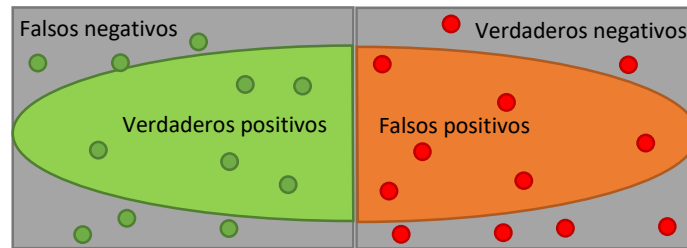


Figura 16. Representación de la clasificación de un modelo (elipse) señalando los falsos negativos (gris izquierda), verdaderos negativos (gris derecha), verdaderos positivos (verde claro) y falsos positivos (naranja) ante muestras negativas (rojas) y positivas (verdes).

5.6.3. ESCALADO PARETO POR BLOQUES

Cuando se cuenta con bloques de variables, es decir, valores de las muestras asociados entre sí porque responden a una misma naturaleza, los bloques con mayor número de variables tendrán, en conjunto, más varianza asociada.

El autoescalado de las variables únicamente hace que cada una de las variables tenga la misma varianza, sin atender a su agrupación por bloques. Esto puede invisibilizar la varianza que aporten las variables en bloques menos numerosos.

El extremo opuesto es el autoescalado por bloques: se iguala la varianza de cada uno de los bloques, igualando la aportación a la variabilidad por cada una de las fuentes o conjuntos de variables. Esto, sin embargo, va en detrimento de la información que puedan aportar bloques de variables más numerosos. En el caso del presente proyecto, por ejemplo, el bloque de variables de las corrientes en los OCP cuenta con 16 variables, mientras que el de los puntos de corriente cuenta con 7728 variables. La varianza unitaria de los bloques significa que, a priori, se le ofrece un peso equivalente a un valor de corriente en OCP que a 483 puntos de corriente.

Entre ambos extremos existe una opción que subsana las consecuencias negativas de ambos: el escalado tipo Pareto. Cada bloque de variables se escala para que la varianza del bloque sea la raíz cuadrada del número de variables. De esta forma, los bloques con mayor número de variables cuentan con mayor varianza, pero sin invisibilizar el resto de bloques con menor número de variables. La transformación a aplicar a cada una de las variables es la división entre su desviación típica (autoescalado de la variable) y entre la raíz cuadrada del número de variables del bloque. En la *Figura 17* se muestra una representación de cada una de las tres opciones ante cuatro bloques de variables.

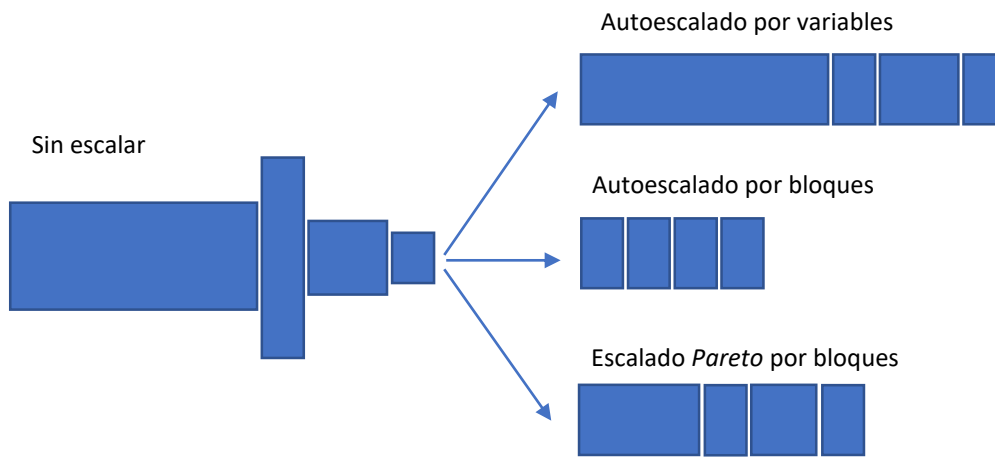


Figura 17. Representación de los distintos escalados.

5.6.4. CLASIFICADOR BINOMIAL

Un clasificador binomial es un algoritmo (también denominado modelo) que, dados unos datos de entrada (normalmente estructurados, como lo son los que se utilizan en el presente proyecto) que componen una muestra, arroja una predicción binaria con cierta tasa de acierto, sensibilidad y especificidad.

Para ser capaces de clasificar nuevas muestras, los modelos deben nutrirse de un conjunto limitado de muestras, de una muestra representativa de la población que se pretende poder clasificar con el modelo. A este proceso se le denomina *entrenamiento*, y al proceso de obtener los valores de AUC, tasa de acierto, sensibilidad y especificidad, se le denomina *validación*.

Los modelos clasificadores deben entrenarse con un conjunto de datos distinto al que luego se va a utilizar como validación, de forma que se evite el sobreajuste. Este fenómeno consiste en que el modelo se adapta no a los patrones que permiten discernir de forma general entre muestras positivas y negativas, sino a las muestras concretas del conjunto de entrenamiento, memorizando sus particularidades para poder identificarlas como positivas o negativas cuando se le vuelvan a presentar. Como es obvio, el sobreajuste es un fenómeno a evitar, porque limita o imposibilita el uso del modelo fuera del conjunto de entrenamiento.

Para poder proporcionar el resultado los modelos clasificadores utilizan distintos operadores sobre los datos de entrada, de forma que se atienden a distintas relaciones entre los mismos. Esto permite la existencia de modelos clasificadores distintos. En este proyecto se va a hacer uso de la regresión en componentes principales (PCR), máquina de soporte vectorial (SVM), redes neuronales artificiales (ANN) con y sin reducción de la dimensión (mediante PCA y PFA), y sNPLSDA, por lo que a continuación se explican cada una de las técnicas.

5.6.4.1. PRINCIPAL COMPONENTS REGRESSION (PCR)

Este modelo clasificador se basa en la obtención de las *componentes principales* de la matriz compuesta por las muestras y variables, y con estas realizar una regresión lineal, un *Principal Components Regression* (PCR).

El uso de una regresión lineal con los datos sin reducir la dimensión (sin obtener las componentes principales mediante un *Análisis de Componentes Principales* (PCA)) presenta el inconveniente de que la covarianza de las variables afecta a la estabilidad de los parámetros de la regresión, por lo que se hace obligatoria la selección de variables, con la consecuente pérdida de información.

El PCA permite la reducción de la dimensionalidad del conjunto de variables de acuerdo a direcciones de variabilidad decreciente, por lo que permite convertir el punto flaco de la regresión lineal, la covarianza de las variables, en un pilar para la construcción de variables independientes, evitando la selección de variables.

Esta técnica de reducción de la dimensión, así como el *Análisis de Factores Principales* (PFA), pueden utilizarse seguidamente con otras técnicas de clasificación, como SVM y ANN, tal y como se hará durante el desarrollo del trabajo, y no exclusivamente con regresión lineal.

¿Pero qué es un PCA? El análisis de componentes principales (PCA) es una técnica multivariante que analiza una matriz de datos en la que las observaciones se describen mediante varias variables dependientes cuantitativas interrelacionadas. Su objetivo es extraer la información importante de la matriz, representarla como un conjunto de nuevas variables ortogonales llamadas componentes principales [11], de forma que puede mostrarse el patrón de similitud de las observaciones y de las variables como puntos en mapas, denominados *score plots* y *loading plots*, respectivamente.

Las componentes principales son una combinación lineal de las variables de entrada, por lo que el PCA se encuentra limitado a describir relaciones lineales entre las variables.

Como ejemplo sencillo, a continuación, en la *Figura 18*, se muestra el ejemplo de una componente principal en un espacio tridimensional [12].

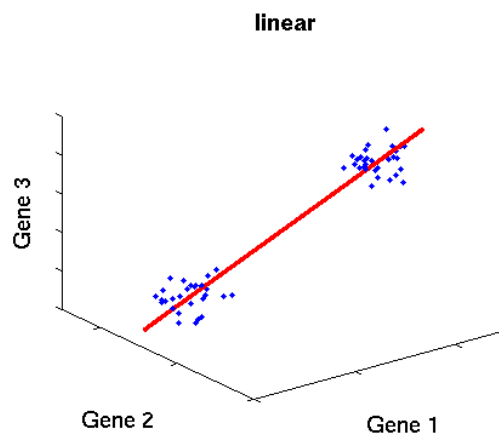


Figura 18. Representación de la primera componente principal en un espacio tridimensional. Fuente: Scholz, M. (2006).

5.6.4.2. PRINCIPAL FACTORS REGRESSION (PFR)

El PFA es un PCA que no construye sus “componentes principales” mediante relaciones lineales de las variables de entrada, sino mediante otro tipo de relaciones. Con el fin de evidenciar las diferencias y no entrar en contradicción con la definición de *componente principal*, las “componentes principales” de un PFA, como su nombre indica, se denominan *factores principales*. Estos factores principales pueden atender a distintas relaciones entre las variables dependiendo del *kernel* del PFA, es decir, de su algoritmo interno. Como ejemplo, en la *Figura 19* se muestra el PFA desarrollado por M. Scholz (2006), que es fácilmente comparable con la *Figura 18*.

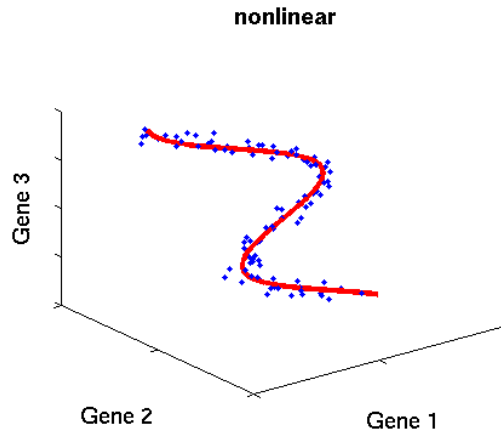


Figura 19. Representación del primer factor principal en un espacio tridimensional. Fuente: Scholz, M. (2006).

Como se puede intuir, existen infinidad de *kernels*, tantos como relaciones no-lineales se pretenda captar con el PFA. En el presente proyecto se han utilizado los *kernel* de la librería *kernelab* (Apartado 9, *Recursos Informáticos Utilizados*, Referencia 3) polinomiales de segundo, tercer y cuarto orden, así como el *kernel* ANOVA con base radial de tercer orden.

Una función *kernel* k calcula el producto escalar de dos matrices x y x' , en un mapa de características $\phi: X \rightarrow H$.

Para los *kernels* polinomiales el parámetro *scale* permanece unitario, y el *offset* nulo, siendo *degree* dos, tres o cuatro, de acuerdo al orden del polinomio:

$$k(x, x') = (\text{scale} \cdot \langle x, x' \rangle + \text{offset})^{\text{degree}}$$

Para el *kernel* con base radial ANOVA de tercer orden, n es igual a tres, y d es unitario.

$$k(x, x') = \left(\sum_{k=1}^n \exp(-\sigma(x^k - x'^k)^2) \right)^d$$

De acuerdo con Schölkopf, Smola, y Müller (1998) [13], el PFA, ejecutado con distintos *kernel*, realiza una transformación no-lineal del sistema de coordenadas encontrando los factores principales que están no-linealmente relacionados con las variables de entrada. Dado un conjunto de observaciones centradas x_k , $k = 1, \dots, M$, $x_k \in \mathbb{R}^N$, PCA diagonaliza la matriz de covarianza:

$$C = \frac{1}{M} \sum_{j=1}^M x_j x_j^T$$

Resolviendo el problema de los valores propios $\lambda v = Cv$.

La misma computación puede hacerse en el espacio del producto escalar F , que está posiblemente relacionado con el espacio de entrada por un mapa no-lineal $\phi: \mathbb{R}^N \rightarrow F$, $x \rightarrow X$. Asumiendo que se trata de datos centrados y se utiliza la matriz de covarianza F :

$$\hat{C} = \frac{1}{C} \sum_{j=1}^M \Phi(x_j) \Phi(x_j)^T$$

Por lo que los factores principales se obtienen con los vectores propios de la matriz *kernel* centrada:

$$K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$$

5.6.4.3. SUPPORT VECTOR MACHINE (SVM)

Las máquinas de soporte vectorial (SVM) son un conjunto de algoritmos de aprendizaje supervisado, cuyo desarrollador principal fue Vladimir Vapnik en AT&T. Las SVM pueden utilizarse tanto en predicción como en clasificación. En este apartado se explica su aplicación en clasificación, en particular en el caso de la clasificación binomial.

De forma intuitiva, en las SVM la clasificación se realiza mediante la separación de las muestras según su clase mediante una hipersuperficie, o, si es necesario, un conjunto de hipersuperficies, proyectando si hace falta las muestras a espacios de dimensionalidad superior (teóricamente tendiendo a infinito, pero en la práctica limitadas por la función de coste).

En el caso de que el conjunto de datos binomial sea lineal y separable, la hipersuperficie será un hiperplano, este viene (total o parcialmente) definido por los dos puntos de las dos clases más cercanas al mismo, configurando el denominado *vector de soporte*. Estos dos puntos más cercanos al vector de soporte deben situarse a la máxima distancia (o margen) posible, parámetro que ajusta el hiperplano.

Así, dado un conjunto separable $C = \{(x_1, y_1), \dots, (x_n, y_n)\}$, donde $x_i \in \mathbb{R}^d$ e $y \in \{1, -1\}$, el hiperplano que separa ambas clases se define como una función lineal: $D(x) = (w_1x_1 + \dots + w_dx_d) + b$, w_i son números reales para todo b real e $i = 1, \dots, d$, cumpliendo con $y_i D(x_i) \geq 0 \forall i \in [1, n] \cap \mathbb{N}$ [14].

Al tratarse de un algoritmo de aprendizaje supervisado (es decir, que realiza un entrenamiento iterativo ajustándose progresivamente al conjunto de datos) se corre el riesgo del sobreajuste de la SVM al conjunto de datos. Esto sucede particularmente en aquellos conjuntos de datos donde una separación perfecta no es posible, porque las pautas de variabilidad de los atributos tienen comportamientos probabilísticos que ocasionan falsos positivos y negativos atendiendo a las relaciones que es capaz de establecer el *kernel* en cuestión. El parámetro C (de coste) permite establecer un margen blando, de forma que los errores son posibles, aunque penalizados. El objetivo de la SVM con margen blando es minimizar la función de coste en vez de perfilar una clasificación perfecta del conjunto de entrenamiento, por lo que seleccionándolo adecuadamente permite que la SVM se ajuste a la estructura latente de discriminación y no a la estructura particular del conjunto de entrenamiento.

Para ello, se generan un conjunto de variables artificiales reales y positivas $\xi_i, \forall i \in [1, n] \cap \mathbb{N}$, de forma que si esta variable artificial es nula significa que se ha realizado correctamente la clasificación, y si esta es positiva, introduce una penalización en la función de coste [14]. Esta función queda descrita como:

$$\min \quad \frac{1}{2} |w|^2 + C \sum_{i=1}^n \xi_i$$

Las hipersuperficies, por defecto, son hiperplanos (o conjuntos de hiperplanos), pero dependiendo del *kernel* de la SVM pueden ser polinomiales o de base radial gaussiana, dotando a la SVM de mayor versatilidad, aunque requiriendo mayor capacidad de cómputo. Esto se consigue mediante la función $\Phi: X \rightarrow F$, que al igual que como se ha explicado en el apartado anterior sobre PFR, convierte cada punto de entrada x en un punto en el espacio F . La función de clasificación, por tanto, es: $D(x) = (w_1\Phi(x_1) + \dots + w_d\Phi(x_d)) + b = \langle w, \Phi(x) \rangle + b$.

A modo de ejemplo ilustrativo, puesto que es el *kernel* que se utiliza en el desarrollo del proyecto, se muestra el *kernel* de base radial gaussiana y un ejemplo de su aplicación en clasificación binomial dependiendo del parámetro *sigma*.

$$K(x_i, x_j) = e^{-\frac{(x_i - x_j)^2}{2 \cdot \sigma^2}}$$

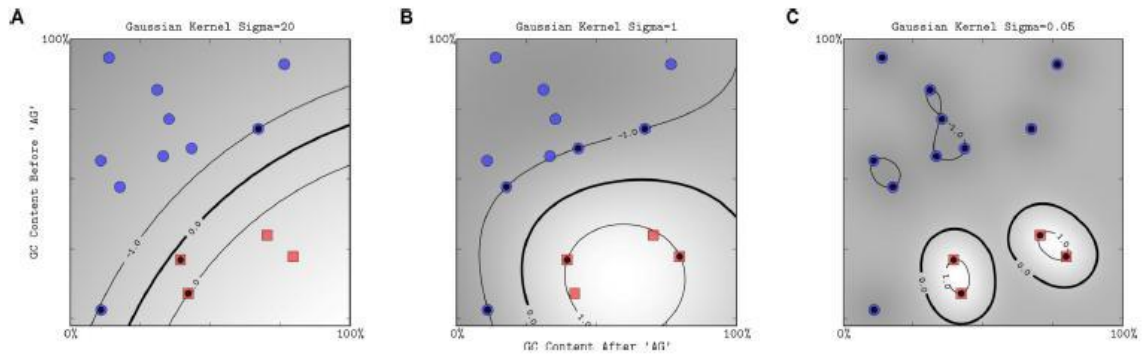


Figura 20. Ejemplo de clasificación binomial con SVM con *kernel* base radial gaussiana con distintos valores para el parámetro *sigma*.

El resultado de la nueva dimensión puede verse representado a continuación, y esta creación de dimensiones superiores puede repetirse las veces que sea necesario.

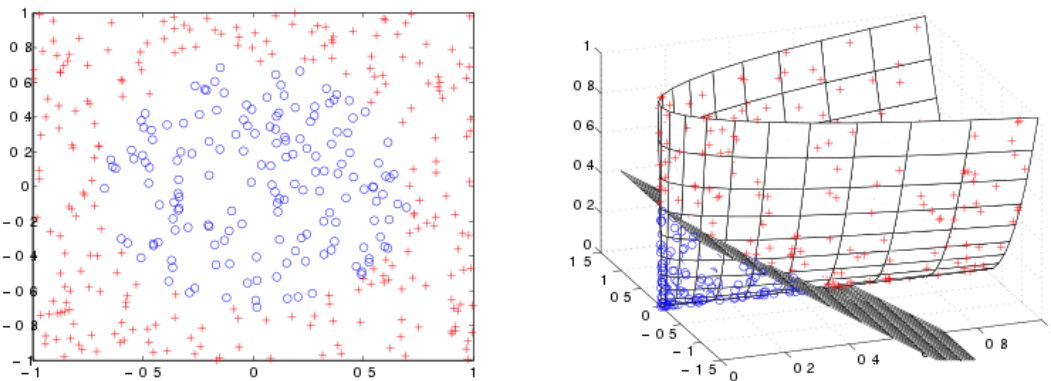


Figura 21. Ejemplo de creación de dimensión no-lineal con SVM. Fuente: Machine Learner (2014).

Las dimensiones originales del conjunto de datos se conocen como *atributos*, mientras las generadas a partir de ellos se conocen como *características*. La creación de un número elevado de características lleva al paso necesario de seleccionar aquellas que cumplen mejor la función de clasificación.

5.6.4.4. SPARSE N-WAY PARTIAL LEAST SQUARES – DISCRIMINANT ANALYSIS (sNPLSDA)

Los mínimos cuadrados parciales (PLS) tiene relación con el PCA, pero en vez de encontrar las direcciones de máxima variabilidad del conjunto de datos, encontrando relaciones fundamentales entre dos matrices: la de variables regresoras (X) y la de variables respuesta (Y). En el caso de que la matriz Y sea binaria, se conoce como PLSDA, (siendo DA el acrónimo de *Análisis Discriminante*), obteniéndose un algoritmo clasificador.

Mientras que el PCA únicamente obtiene como componentes principales las direcciones de máxima varianza de su espacio X, PLS obtiene las direcciones que maximizan la covarianza entre las variables

latentes de X y de Y. Estas direcciones de máxima covarianza se denominan *variables latentes*, por lo que, tras entrenar el modelo con un subconjunto de muestras, es posible proyectar nuevas muestras en el espacio de variables latentes, obteniendo su clasificación y qué variaciones en los valores independientes se encuentran asociados a cada clasificación. El algoritmo más empleado para la obtención de las variables latentes es NIPALS (*Nonlinear estimation by Iterative Partial Least Square*) [15].

El término *sparse* del PLSDA hace referencia a una selección de variables. Como se ha comentado, cada variable latente es una composición lineal de variables X. Debido al ruido inherente en las muestras, aquellas variables que no tengan contribución en una variable latente es probable que tengan una pequeña contribución en la variable latente distinta de cero. Con el fin de reducir la influencia del ruido, y obtener clasificadores más robustos, pueden establecerse con contribución cero para todas las variables latentes aquellas variables que tengan una influencia total en la contribución en todas las variables latentes inferior a cierto umbral. Si el umbral se compara con el sumatorio de las contribuciones en valor absoluto por variable, se denomina una penalización-L1, que es la que se aplica en este proyecto. Por eficiencia experimental, en el presente proyecto el número de variables a permanecer se establece de antemano, de forma que el umbral varía hasta eliminar el número de variables necesario [16].

Atendiendo a la naturaleza tridimensional del conjunto de datos, el método *sparse* puede aplicarse no al conjunto de variables sin estructura, sino a cada uno de los modos del conjunto de datos. El algoritmo NPLS aplica PLS a estructuras N-dimensionales, ejecutando NIPALS en cada uno de los modos de forma sucesiva. De esta forma se obtienen idénticos resultados que con PLS para un mismo conjunto de datos, pero se facilita su comprensión (puesto que las contribuciones se pueden observar por cada modo) [16]. Además, presenta otra ventaja: que para los modelos *sparse* es posible elegir el número de variables por modo a permanecer: en vez de elegir entre 1 a 7728 variables, que es lo que serían necesario hacer en sPLSDA, pueden elegirse el número de electrodos y de puntos de corriente, pudiéndose identificar si hay tramos de las formas de onda que especialmente aportan información, o electrodos que son los que presentan mayor covarianza con la matriz Y; arrojando información valiosa sobre el proceso de discriminación. Cómo no, esto sólo es posible si se obtiene un modelo sNPLSDA con buena capacidad de clasificación.

5.6.4.5. ARTIFICIAL NEURAL NETS (ANN)

Las redes neuronales artificiales (ANN por sus siglas en inglés) son modelos computacionales que han sido concebidos inspirados en el funcionamiento de las redes neuronales biológicas.

Consiste en un conjunto de unidades, denominadas *perceptrones* o *neuronas*, conectadas entre sí en distintas arquitecturas. En el presente proyecto únicamente se utilizan *Multilayer Perceptron* (MLP, redes multicapa), que consisten en una capa de neuronas de entrada (una por cada variable de entrada), un conjunto de capas ocultas, y una capa de salida (una por cada variable respuesta de la que emitir una predicción o clasificación).

Cada capa emite señales a la siguiente, y recibe señales de la anterior, de forma que todas las neuronas de una capa emiten a todas las neuronas de la siguiente, y todas las neuronas de una capa reciben información de la anterior (a excepción de las capas de entrada y de salida, por razones obvias). Esto puede verse más claramente con la *Figura 22*.

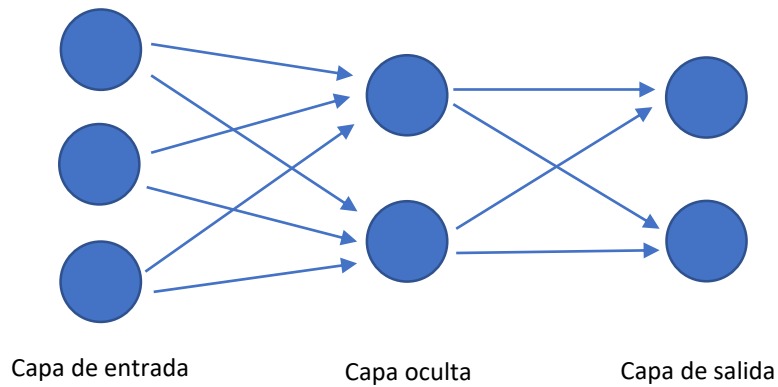


Figura 22. Ejemplo de MLP con una capa oculta.

De esta forma, la información va de las capas de entrada a las de salida a través de las capas ocultas, realizándose un procesamiento en cada una de las neuronas de los datos entrantes, que se emiten a las neuronas de la siguiente capa procesados. Este procesamiento se encuentra caracterizado por un *peso* (que debe ajustarse mediante entrenamiento), que multiplica el valor de entrada y una función de activación (que modifica el efecto del peso).

La función de activación utilizada en este proyecto es la función *sigmoid*, puesto que la función de la red neuronal puede describirse como una suma ponderada no lineal $f(x) = k(\sum_i w_i g_i(x))$, donde k es la función de activación. Esta función de activación (aunque también otras) permite que la red neuronal en su conjunto cumpla con el teorema de aproximación universal, de forma que con una o más capas ocultas con un número finito de neuronas puede aproximar funciones continuas en subconjuntos limitados en \mathbb{R}^n . Esto es gracias a que la función de activación cumple ciertas características [17].

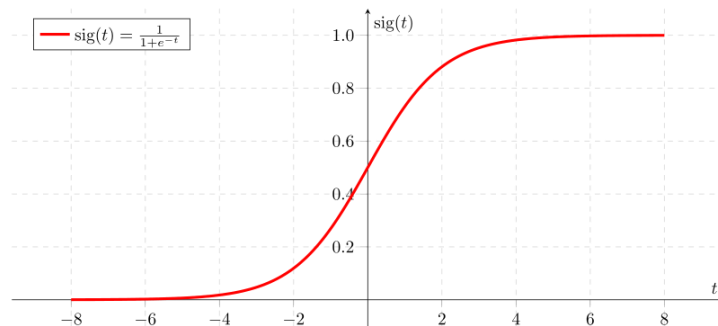


Figura 23. Función de activación *sigmoid*.

Los pesos de cada uno de los perceptrones son los que dotan a la ANN de la capacidad de clasificación, pero estos deben ajustarse. La forma de hacerlo es mediante el *entrenamiento*. Este entrenamiento se realiza mediante *propagación hacia atrás de errores*, en el cual se aplica un patrón a la entrada de la ANN como estímulo, propagándose desde la entrada hasta la salida. La salida ofrecida por la ANN se compara con la que realmente pertenece a los valores de entrada (en el caso de clasificación, la etiqueta de clasificación, en el caso de binomial, cero o uno), de forma que se obtiene el error en la salida. Este error se propaga hacia las neuronas que han contribuido al resultado de forma proporcional, modificando proporcionalmente sus pesos. Este proceso se repite, de forma que conforme más entrenada está la red, más se adecúa la estructura de la misma para reconocer patrones en el espacio de entrada, y puedan clasificar correctamente conjuntos de datos que no pertenecen al conjunto de entrenamiento [17].

El valor que intenta minimizarse mediante el entrenamiento de la ANN es la *función de pérdida*, en el caso particular de la clasificación binaria, es la *entropía binaria cruzada*. De una forma intuitiva, la entropía mide la incertidumbre de una fuente de información, o, dicho de otra forma, la cantidad de información promedio que contienen los símbolos usados. Si todos los símbolos contienen la misma información, se dice que la entropía es máxima. En aras de reducir la cantidad de símbolos necesarios para realizar una clasificación, la reducción (o minimización, de hacerse una reducción al máximo) de la entropía consiste en configurar el sistema simbólico de forma que un número reducido de símbolos aporte la mayor parte de la información, simplificando por tanto el modelo de clasificación. Relacionado con la incertidumbre, la entropía se puede asociar a la cantidad de ruido presente en el modelo: a menor entropía, mayor ratio señal-ruido, más fácilmente pueden identificarse las variables (tangibles o latentes) que permiten la clasificación. La entropía cruzada cuantifica la diferencia entre dos distribuciones de probabilidad: la distribución verdadera de, en este caso, el vector respuesta binario (presencia o ausencia de cáncer, 1 o 0), p , respecto a la obtenida con el modelo clasificador, q . Por tanto, permite la cuantificación del error entre la predicción y los valores reales mediante la comparación entre la distribución obtenida con el modelo clasificador q y la verdadera distribución del conjunto de variables a predecir, p , de acuerdo a la siguiente expresión:

$$H(p, q) = E_p[-\log(q)] = H(p) + D_{KL}(p||q)$$

Donde $H(p)$ es la entropía de p y $D_{KL}(p||q)$ es la divergencia de Kullback-Leibler o entropía relativa. Al ser p y q variables discretas:

$$H(p, q) = - \sum_x p(x) \log(q(x))$$

Cada una de las tandas de entrenamiento se conocen como *epochs*. El número de *epochs* determina hasta qué punto van a ajustarse los parámetros, un número demasiado limitado evitará que la ANN pueda entrenarse completamente, mientras que uno demasiado grande hará que la ANN se sobreentrene. Como ya se ha comentado, el problema de los algoritmos supervisados iterativos es que pueden *sobreentrenarse*, al igual que las SVM, sobreajustarse al conjunto de datos de entrenamiento.

La forma en la que se va a establecer el número de *epochs*, de forma que se evite el sobreentrenamiento, es implementando una validación tras cada *epoch*, es decir, durante el desarrollo, con un subconjunto de datos distinto al de validación, denominándose este subconjunto *de desarrollo*.

El proceso de entrenamiento tiene, por tanto, tandas de entrenamiento de las capas – validación, de forma que es posible identificar cuando se empieza a sobreentrenar, puesto que el conjunto de *desarrollo* empieza a presentar una función de pérdida que empieza a aumentar mientras se minimiza la arrojada por el conjunto de entrenamiento. Con el fin de evitar aumentos puntuales de la función de pérdida del subconjunto de desarrollo o identificar mínimos locales sin dar oportunidad a encontrar mínimos globales o locales mejores, se implementa el parámetro *paciencia*, que especifica el número de *epochs* a esperar antes de considerar que se está sobreentrenando (es decir, número de *epochs* en el que se está empeorando la función de pérdida para el subconjunto de desarrollo), restableciéndose los pesos que arrojaron mejores resultados una vez se para el proceso de entrenamiento [17].

La facilidad para que el entrenamiento sea eficiente vendrá condicionada por la estructura de la ANN: el número de capas y neuronas por capa. Para esto no hay reglas aplicables, por lo que deben explorarse distintas arquitecturas. Además, hay otras técnicas que pueden mejorar el proceso de entrenamiento. Por un lado, el uso de *dropouts* entre capas: que aleatoriamente, un porcentaje de las señales de la capa anterior en la que se ubica el *dropout* se anulan, normalmente entre un 10% y 50%. Esto es especialmente

útil ante variables de entrada correlacionadas. Otra técnica que permite mejorar el proceso de entrenamiento es la introducción de regularizadores en las capas: estos introducen una penalización a más neuronas tengan pesos distintos a cero en la estructura, de forma que se dificulta el sobreentrenamiento, manteniendo el número de neuronas activas en un mínimo y disminuyendo la influencia del ruido en los datos durante el entrenamiento [17].

5.6.5. ANOVA E INTERVALOS DE TUKEY

El análisis de la varianza o ANOVA permite determinar si existen diferencias estadísticamente significativas entre poblaciones, basándose en muestras de las mismas, siempre y cuando se cumplan las siguientes hipótesis:

- La variable dependiente debe estar comprendida en un intervalo.
- Las observaciones son independientes.
- Los residuos se distribuyen normalmente.
- Existe homogeneidad entre las varianzas.

Realizándose mediante la tabla que se muestra a continuación donde la F obtenida se compara con el valor de la F de *Snedecor* para los grados de libertad en los factores y entre ellos, con un p-Valor [18].

Tabla 2. Ejemplo de tabla ANOVA.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Factores	SC_F	$n_F - 1$	$CM_F = SC_F / gdl_F$	CM_F / CM_R
Residuos	SC_R	$n_T - n_F$	$CM_T = SC_R / gdl_R$	
Total	SC_T	$n_T - 1$		

Si el valor de la F calculada es mayor que la F de *Snedecor* de la tabla significa que existen diferencias significativas de acuerdo al p-Valor elegido (normalmente 5%).

¿Y si la comparación es entre más de una pareja de medias en conjuntos equilibrados (con mismo número de muestras)? Para, tras el análisis ANOVA que señale que hay diferencias significativas entre las medias, pueden utilizarse los intervalos de confianza de *Tukey*, también conocidos como HSD (*Honestly-significant-difference*), un test conservador mediante el que puede identificarse qué medias son significativamente distintas de cada una de las otras, puesto que se comparan de forma múltiple [18].

6. COMPARACIÓN DE MODELOS

El objetivo es el desarrollo de un clasificador binomial, de forma que sea capaz de distinguir entre pacientes sanos y con cáncer, independientemente del estadio de desarrollo tumoral. Esto presenta la ventaja de ser capaz de detectar los estadios incipientes, pero a su vez, debido a la necesidad de diferenciar entre estos y los sanos, que presentarán pautas multivariantes más similares entre sí, puede disminuir en sensibilidad hacia los casos de mayor estadio, así como de especificidad.

Para el clasificador binomial se exploran los modelos tanto sin variables adicionales, como con ellas.

Se valúan modelos clasificadores basados en *Principal Component Regression* (PCR) y *Principal Factor Regression* (PFR, PCR con *kernel* no-lineal), *Support Vector Machine* (SVM) sin y con reducción de la

dimensión (utilizando PCA y PFA, como en las regresiones anteriores), *sparse N-way Partial Least Squares Discriminant Analysis* (sNPLSDA) y *Artificial Neural Nets* (ANN) sin y con reducción de la dimensión, cada uno con los cinco pretratamientos comentados en 5.3.

6.1. CLASIFICACIÓN BINOMIAL SIN VARIABLES ADICIONALES

6.1.0. MODELO BASE: PLSDA

El modelo que se escoge para establecer la marca inicial es un PLSDA.

Se obtiene, con validación cruzada a nueve grupos balanceados, un AUC medio máximo con 10 variables latentes de 0.597, con una tasa de aciertos asociada de 0.584.

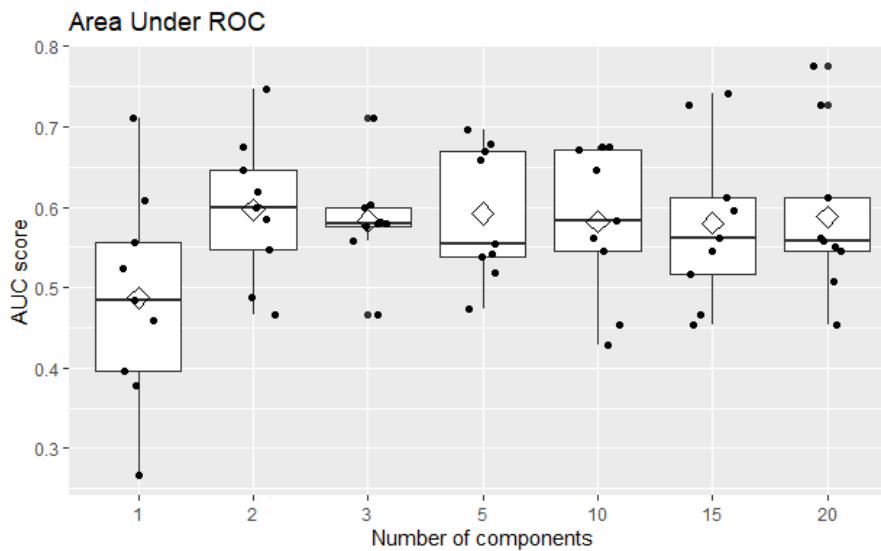


Figura 24. AUC ("AUC score" valor en el eje Y) para PLSDA con distinto número de variables latentes ("Number of components" en el eje X).

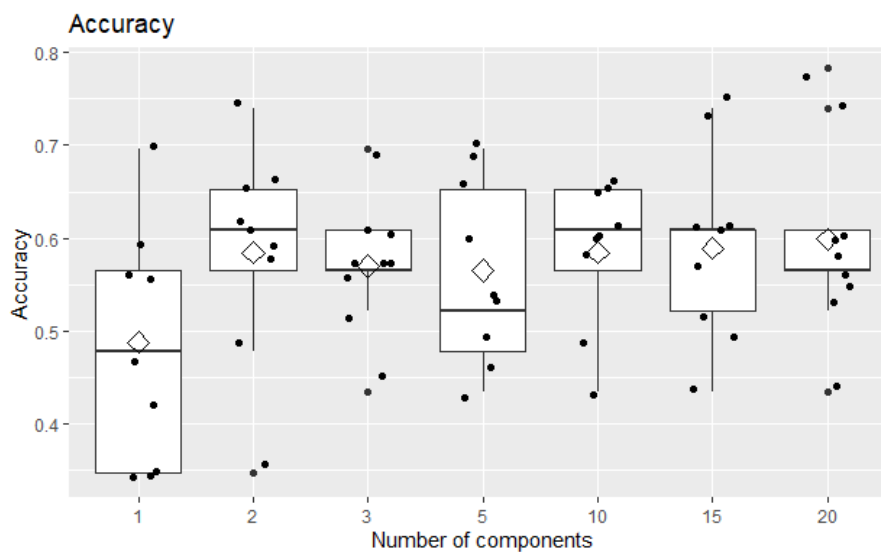


Figura 25. Tasa de acierto ("Accuracy" en el eje Y) para PLSDA con distinto número de variables latentes ("Number of components" en el eje X).

6.1.1. MODELOS BASADOS EN PCR Y PFR

Se muestran a continuación 975 modelos clasificadores, correspondientes al uso de 2 a 40 variables latentes para los cinco pretratamientos y cinco *kernels* de reducción de la dimensión (PCA, polinomial de segundo orden, polinomial de tercer orden, polinomial de cuarto orden y ANOVA).

Se seleccionan 25 modelos, correspondientes al número de variables latentes que maximiza la AUC media de la validación cruzada a nueve grupos, para cada uno de los cinco pretratamientos y cinco técnicas de reducción de la dimensión.

Tabla 3. AUC medio máximo para número de variables latentes.

	PCR	PFR pol. 2ºG	PFR pol. 3ºG	PFR pol. 4ºG	PFR ANOVA
Autoescalado	0.606	0.617	0.648	0.589	0.659
Centrado	0.626	0.562	0.580	0.543	0.659
Modo 2	0.631	0.561	0.549	0.542	0.649
Modo 3	0.622	0.559	0.579	0.534	0.657
Modos 2 y 3	0.629	0.559	0.550	0.537	0.651

Se muestran en detalle en los anexos 1 a 25.

6.1.2. MODELOS BASADOS EN SVM

Se distinguen los modelos sin reducción de la dimensión, y los modelos con reducción de la dimensión. Para los segundos, se sigue la misma filosofía que los modelos PCR y PFR.

Se muestran a continuación 980 modelos clasificadores, correspondientes al uso de 2 a 40 variables latentes para los cinco pretratamientos y cinco *kernels* de reducción de la dimensión (PCA, polinomial de segundo orden, polinomial de tercer orden, polinomial de cuarto orden y ANOVA), y a los cinco sin reducción de la dimensión.

Se seleccionan los cinco modelos sin reducción de la dimensión y, del resto, como se ha comentado, se sigue el mismo criterio de selección que para los modelos PFR y PCR.

Tabla 4. AUC medio máximo para número de variables latentes.

	Sin reducción	PCA	PFA pol. 2ºG	PFA pol. 3ºG	PFA pol. 4ºG	PFA ANOVA
Autoescalado	0.629	0.631	0.605	0.576	0.581	0.586

Centrado	0.613	0.671	0.580	0.580	0.578	0.630
Modo 2	0.629	0.662	0.573	0.591	0.567	0.651
Modo 3	0.629	0.667	0.587	0.578	0.589	0.658
Modos 2 y 3	0.629	0.659	0.563	0.592	0.595	0.646

Se muestran en detalle en los anexos 26 a 51.

6.1.3. MODELOS BASADOS EN sNPLSDA

Los resultados para los cinco pretratamientos con sNPLSDA se muestran a continuación:

Tabla 5. AUC medio máximo para cada pretratamiento.

Autoescalado	Centrado	Modo 2	Modo 3	Modos 2&3
0.654	0.611	0.633	0.677	0.645

Se muestran en detalle en los anexos 52 a 56.

6.1.4. MODELOS BASADOS EN ANN

Para los modelos basados en redes neuronales artificiales (ANN) se distinguen los modelos sin reducción de la dimensión, y los modelos con reducción de la dimensión. Para los segundos, se reduce la dimensión con las técnicas de PCA y PFA de los anteriores, pero introduciendo los cuarenta primeros factores de cada una de las técnicas como variables de entrada en las ANN, haciendo un total de 200 variables de entrada.

Para todos los modelos clasificadores basados en ANN se utiliza un *dropout* entre las capas internas de 0.25, excepto de los nodos de entrada a la primera capa oculta en los modelos con reducción de la dimensión. La función de activación es *sigmoid* para todas las capas, puesto que es una función de activación que cumple el teorema de la aproximación universal, según el cual para redes neuronales establece que cada función continua que mapea intervalos de números reales a algún intervalo de salida de números reales puede ser aproximado arbitrariamente por un perceptrón (neurona) multicapa con una sola capa escondida.

Además de la reducción de la dimensión se distingue el número de capas (una, dos o tres), el regularizador L^2 empleado y el número de neuronas empleadas por capa. Se escogen aquellos modelos que no presentan grupos con AUC por debajo de 0.5, siempre que los haya.

Las redes neuronales artificiales, además, tienen la particularidad respecto al resto de modelos que deben ser entrenadas. Es por ello que, en vez de utilizar las 184 muestras para entrenar, utilizando como conjunto de desarrollo el mismo conjunto de datos de validación, se opta por entrenar con 150 muestras del conjunto de datos de entrenamiento original, utilizando como conjunto de datos de validación durante

el entrenamiento (de desarrollo) los 24 restantes. Esta selección se realiza aleatoriamente por cada red neuronal artificial entrenada, velando porque tanto el subconjunto de datos de entrenamiento como los de desarrollo se encuentren balanceados.

Esto último es particularmente importante puesto que permite, como se ha comentado., realizar un *ensemble* de modelos sin que las redes neuronales hayan visto, previamente, el conjunto de datos de validación, lo que llevaría a sobreentrenar el modelo clasificador y que no se encontrase ajustado de acuerdo a la estructura latente inferida de los datos, sino directamente sobreentrenada a todo el conjunto disponible.

Tabla 6. AUC medio máximo cada conjunto de ANN.

	Sin reducción una capa	Sin reducción dos capas	Sin reducción tres capas	Con reducción una capa	Con reducción dos capas	Con reducción tres capas
Autoescalado	0.597	0.607	0.632	0.587	0.640	0.611
Centrado	0.622	0.593	0.678	0.650	0.665	0.693
Modo 2	0.626	0.591	0.654	0.642	0.591	0.614
Modo 3	0.631	0.587	0.632	0.628	0.585	0.647
Modos 2 y 3	0.632	0.615	0.651	0.583	0.587	0.630

Se muestran en detalle en los anexos 57 a 86.

6.1.5. MODELOS ENSEMBLE

El teorema del *Juicio de Condorcet* afirma que la probabilidad de tasa de aciertos tiende a 100% ante un sistema de votación simple conformado por N modelos clasificadores independientes, si N es impar y tiende a infinito. Los clasificadores mostrados en los apartados anteriores, como es de esperar por basarse en el mismo conjunto de datos, no son independientes (y no se cuenta con suficientes muestras como para poder implementar tal aproximación). Sin embargo, el uso de distintas técnicas de clasificación arroja distintos resultados: hay variabilidad porque cada modelo clasificador se construye en base a la observación de distintas relaciones en el conjunto de datos, bien sea por el pretratamiento aplicado (de un total de cinco), la técnica de reducción de la dimensión (ninguna, PCA o cuatro PFA distintos), o el modelo clasificador (regresión lineal, SVM, sNPLSDA, ANN con distintas arquitecturas). Además, el método para dar una respuesta final no tiene por qué estar limitado a un sistema de votación, por lo que en la jerarquía superior del modelo clasificador pueden situarse otra clase de modelos (promedio, *K vecinos más cercanos* (KNN), SVM, *random forest*, ANN). De esta forma pueden explorarse modelos que traten de seleccionar las clasificaciones que acierten más para cierta clase de muestras, fortaleciéndose los modelos de base unos a los otros.

Para el *ensemble* se utilizan las predicciones de los 90 modelos seleccionados en los anteriores apartados como variables de entrada, manteniendo la validación cruzada en nueve grupos. Es decir, se mantiene

durante todo el proceso de construcción un conjunto de datos de validación oculto para el entrenamiento de los modelos, a excepción de para la ANN en la etapa del *ensemble*, puesto que en su proceso de entrenamiento debe ver periódicamente el conjunto de validación para comprobar si realmente se está ajustando a la estructura latente de los datos o se está sobreajustando a las muestras de entrenamiento.

Los modelos clasificadores *ensemble* que no son ANN son susceptibles de empeorar ante la incorporación de modelos que no aportan información adicional a la clasificación. Es por ello que se realiza una regresión *stepwise*, en la que se empieza con el modelo clasificador que arroja mayor AUC medio (de los nueve grupos de validación cruzada), y se van añadiendo los modelos clasificadores simples que mejoran más el AUC del modelo *ensemble* final. Esto se muestra en la *Figura 26*.

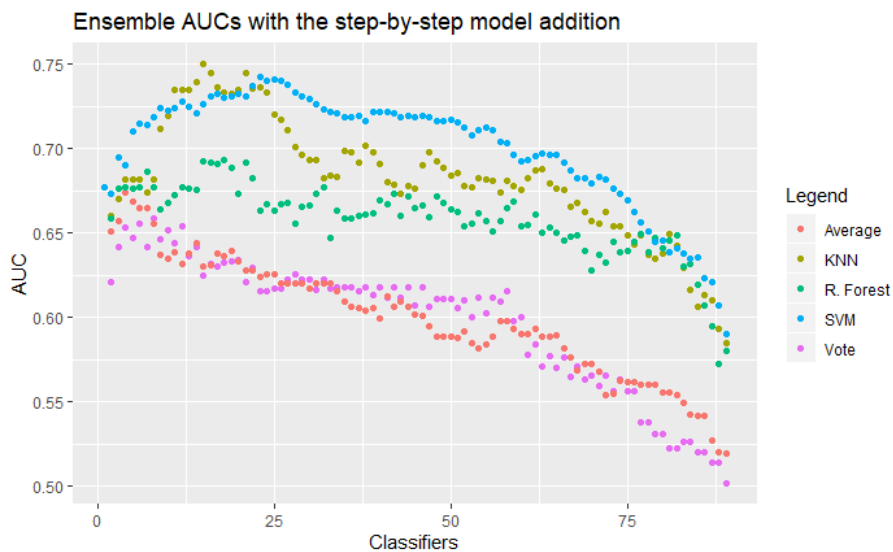


Figura 26. AUC del *ensemble* con la inclusión de clasificadores, dependiendo de la técnica de clasificación superior.

Los modelos *ensemble* de votación y de promedio no ofrecen mejores resultados en ningún caso que el mejor de los modelos clasificadores de la base, el sNPLSDA escalado en el modo 3 con 50 puntos de corriente y tres electrodos. Los modelos *ensemble* han arrojado un AUC medio máximo de 0.658 y 0.673 para votación y promedio, respectivamente, contra 0.677 del sNPLSDA.

Paralelamente, se entrena una ANN con todos los modelos clasificadores a su entrada. Se evalúan 27 arquitecturas de dos capas distintas, con 8, 16 y 32 neuronas por capa, con regularizadores L^2 de 0.1, 0.001 y 10^{-6} .

El que mejores resultados arroja es con un L^2 de 0.1, 32 neuronas en la primera capa y 8 neuronas en la segunda, con un AUC medio de 0.554. Un valor realmente bajo, sustancialmente por debajo del 0.677 del mejor de los modelos clasificadores de la base.

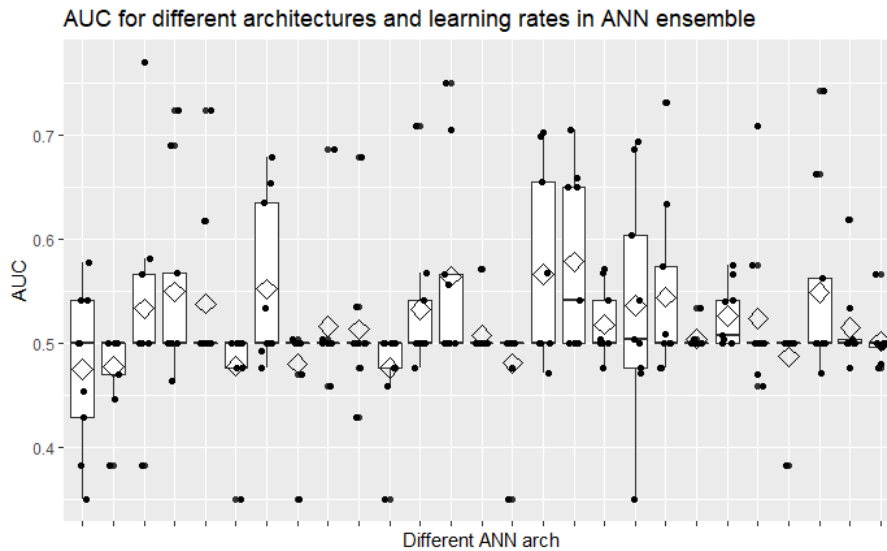


Figura 27. AUC para cada una de las arquitecturas y learning rates de ANN en ensemble.

Se escoge el mejor de los modelos de acuerdo al número de clasificadores básicos incluidos, para las técnicas de clasificación KNN, RF y SVM, y se comparan de acuerdo a sus AUC y tasas de aciertos para cada uno de los grupos de la validación cruzada, primero de acuerdo a los clasificadores de base que maximizan el AUC promedio, y luego para los que maximizan la tasa de aciertos.

Los resultados que se muestran son la ampliación de los valores calculados en la regresión *stepwise* anterior para cada uno de los modelos, pero ampliando en los puntos especificados para cada uno de los modelos *ensemble*.

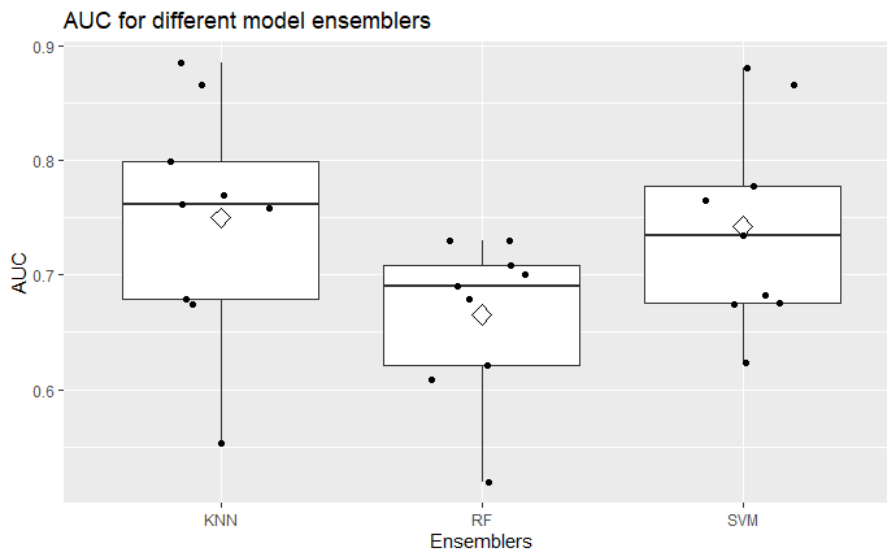


Figura 28. Modelos *ensemble* de acuerdo a su AUC.

Como puede observarse, el *random forest* presenta un valor de AUC promedio inferior al resto de modelos, presentando una media y mediana menor que el resto, por lo que de momento es el primero que se descarta. KNN y SVM presentan resultados similares, teniendo mayor media de AUC entre grupos SVM, pero KNN mayor mediana. Para salir de dudas, se muestran las tasas de aciertos de los modelos

para el conjunto de clasificadores de base que maximizan el AUC, así como para el conjunto que maximizan la tasa de aciertos de cada uno de los modelos.

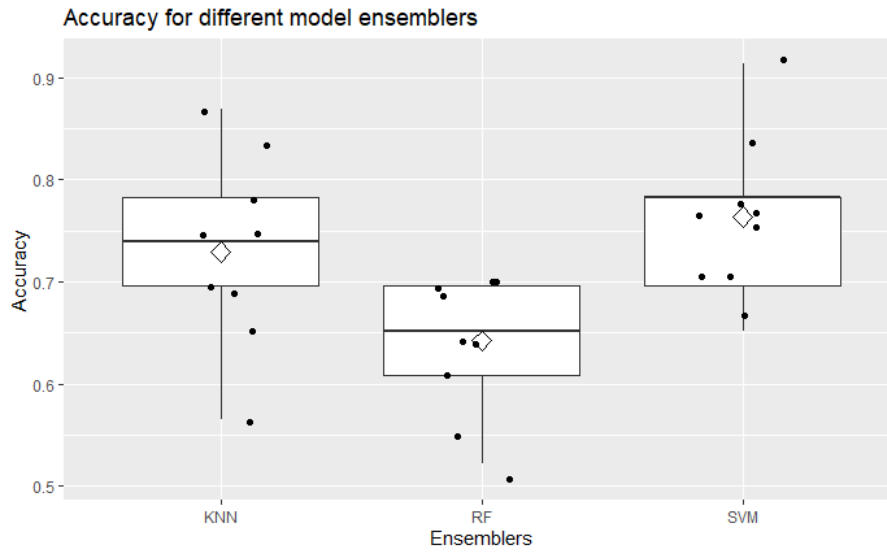


Figura 29. Tasas de aciertos por modelo *ensemble* para los modelos clasificadores de base que maximizan AUC promedio.

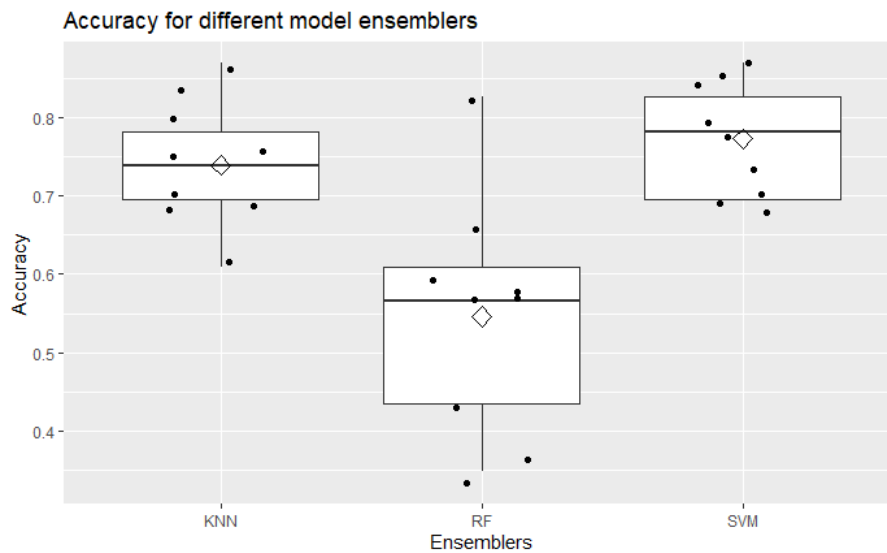


Figura 30. Tasas de aciertos por modelo *ensemble* para los modelos clasificadores de base que maximizan tasa de aciertos promedio.

Para comparar los resultados y observar la magnitud y significación estadística de las diferencias de los métodos en las tasas de aciertos maximizadas, se realiza un ANOVA.

Tabla 7. ANOVA de los modelos *ensemble*.

	GdL	SC	SCM	Valor-F	Valor-P
Modelos	2	0.2701	0.13506	11.76	0.000275
Residuos	24	0.2756	0.01148		

Existen diferencias estadísticamente significativas entre la media de tasa de aciertos de los tres modelos *ensemble*. Como puede observarse a continuación con los intervalos de *Tukey*, esta diferencia sólo es significativa del KNN y SVM respecto a RF.

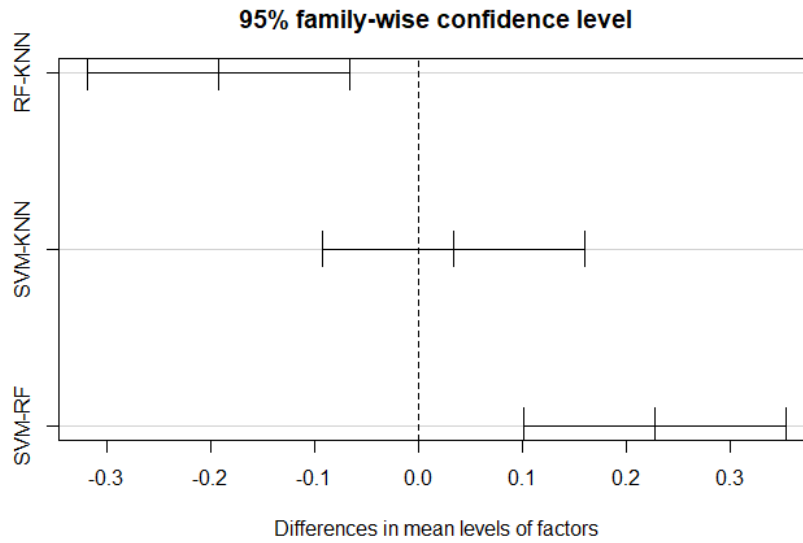


Figura 31. Tukey HSD para los modelos *ensemble*.

Como modelo final se escoge el SVM, aunque su tasa de aciertos no es estadísticamente superior a la obtenida con KNN.

6.1.6. MODELO FINAL

A continuación, se muestran el AUC, tasa de aciertos, sensibilidad y especificidad del modelo final.

El cálculo de los intervalos de confianza, al 95%, de cada uno de los estadísticos, se realiza en base al valor crítico de la distribución t de Student para 9 muestras.

$$\bar{x} - t_{n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1} \cdot \frac{s}{\sqrt{n}}$$

Tabla 8. Modelo final.

	Estimación	Amplitud del intervalo de confianza (95%)
AUC	0.726	±0.062
Tasas de aciertos	0.773	±0.058
Sensibilidad	0.764	±0.115
Especificidad	0.775	±0.126

6.2. CLASIFICACIÓN BINOMIAL CON VARIABLES ADICIONALES

6.2.0. MODELO BASE: PLSDA

El modelo que se escoge para establecer la marca inicial es un PLSDA.

Se obtiene, con validación cruzada a nueve grupos balanceados, un AUC medio máximo con 10 variables latentes de 0.595, con una tasa de aciertos asociada de 0.589.

Los resultados son muy similares al obtenido sin variables adicionales.

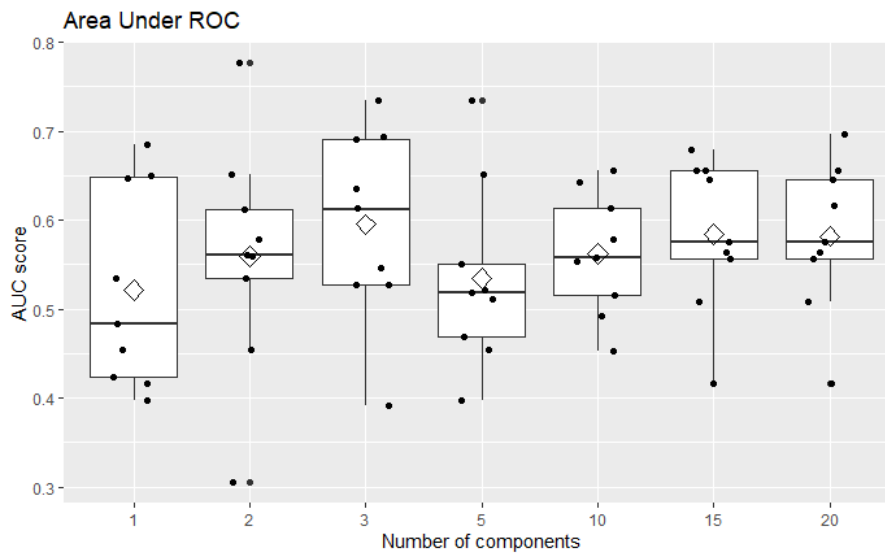


Figura 32. AUC para PLSDA con distinto número de variables latentes.

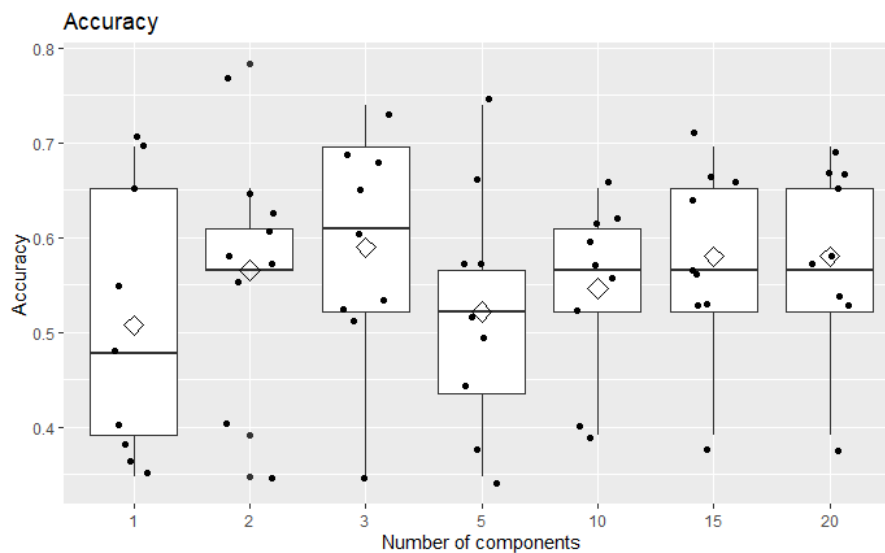


Figura 33. Tasa de acierto para PLSDA con distinto número de variables latentes.

6.2.1. MODELOS BASADOS EN PCR Y PFR

Tal y como se mostró con los modelos sin variables adicionales, se muestran a continuación 975 modelos clasificadores, correspondientes al uso de 2 a 40 variables latentes para los cinco pretratamientos y cinco *kernels* de reducción de la dimensión (PCA, polinomial de segundo orden, polinomial de tercer orden, polinomial de cuarto orden y ANOVA).

Se seleccionan 25 modelos, correspondientes al número de variables latentes que maximiza la AUC media de la validación cruzada a nueve grupos, para cada uno de los cinco pretratamientos y cinco técnicas de reducción de la dimensión.

Tabla 9. AUC medio máximo para número de variables latentes.

	PCR	PFR pol. 2ºG	PFR pol. 3ºG	PFR pol. 4ºG	PFR ANOVA
Autoescalado	0.638	0.598	0.598	0.581	0.632
Centrado	0.655	0.570	0.604	0.595	0.651
Modo 2	0.655	0.571	0.612	0.563	0.644
Modo 3	0.665	0.556	0.607	0.584	0.645
Modos 2 y 3	0.661	0.567	0.609	0.554	0.643

En los anexos 87 a 111 se muestran en detalle.

6.2.2. MODELOS BASADOS EN SVM

Se distinguen los modelos sin reducción de la dimensión, y los modelos con reducción de la dimensión. Para los segundos, se sigue la misma filosofía que los modelos PCR y PFR.

Se muestran a continuación 980 modelos clasificadores, correspondientes al uso de 2 a 40 variables latentes para los cinco pretratamientos y cinco *kernels* de reducción de la dimensión (PCA, polinomial de segundo orden, polinomial de tercer orden, polinomial de cuarto orden y ANOVA), y a los cinco sin reducción de la dimensión.

Se seleccionan los cinco modelos sin reducción de la dimensión y, del resto, se sigue el mismo criterio de selección que para los modelos PFR y PCR.

Tabla 10. AUC medio máximo para número de variables latentes.

Sin reducción	PCA	PFA pol. 2ºG	PFA pol. 3ºG	PFA pol. 4ºG	PFA ANOVA
------------------	-----	-----------------	-----------------	-----------------	--------------

Autoescalado	0.613	0.596	0.595	0.608	0.606	0.614
Centrado	0.629	0.638	0.605	0.596	0.591	0.624
Modo 2	0.629	0.638	0.605	0.596	0.591	0.623
Modo 3	0.629	0.623	0.626	0.589	0.591	0.628
Modos 2 y 3	0.629	0.630	0.601	0.600	0.596	0.624

En los anexos 112 a 137 se muestran en detalle.

6.2.3. MODELOS BASADOS EN sNPLSDA

Los resultados para los cinco pretratamientos con sNPLSDA se muestran a continuación

Tabla 11. AUC medio máximo para cada pretratamiento.

Autoescalado	Centrado	Modo 2	Modo 3	Modos 2 y 3
0.626	0.652	0.624	0.674	0.649

Se muestran en detalle en los anexos 138 a 142.

6.2.4. MODELOS BASADOS EN ANN

De la misma forma que para los modelos basados en ANN sin variables adicionales, con estas se distinguen los modelos sin reducción de la dimensión, y los modelos con reducción de la dimensión. Para todos los modelos clasificadores basados en ANN se utiliza un *dropout* entre las capas internas de 0.25, excepto de los nodos de entrada a la primera capa oculta en los modelos con reducción de la dimensión. La función de activación es *sigmoid* para todas las capas.

También de la misma forma que para sin variables adicionales se distingue el número de capas (una, dos o tres), el regularizador L^2 empleado y el número de neuronas empleadas por capa. No se utilizan las 184 muestras para entrenar, sino 150 muestras del conjunto de datos de entrenamiento original, utilizando como conjunto de datos de validación durante el entrenamiento (de desarrollo) los 24 restantes. Esta selección se realiza aleatoriamente por cada red neuronal artificial entrenada, velando porque tanto el subconjunto de datos de entrenamiento como los de desarrollo se encuentren equilibrados.

Tabla 12. AUC medio máximo cada conjunto de ANN.

Sin reducción una capa	Sin reducción dos capas	Sin reducción tres capas	Con reducción una capa	Con reducción dos capas	Con reducción tres capas
---------------------------------------	--	---	---------------------------------------	--	---

Autoescalado	0.629	0.639	0.641	0.603	0.594	0.601
Centrado	0.644	0.641	0.644	0.658	0.650	0.618
Modo 2	0.648	0.651	0.633	0.620	0.576	0.610
Modo 3	0.639	0.660	0.658	0.593	0.605	0.584
Modos 2 y 3	0.618	0.650	0.635	0.581	0.654	0.628

En los anexos 143 a 171 se muestran en detalle.

6.2.5. MODELOS ENSEMBLE

Tal y como se ha explicado en el apartado análogo para modelos sin variables adicionales, se utiliza el teorema del *Juicio de Condercet* como principio para el desarrollo de modelos *ensemble*. Para la jerarquía superior del modelo clasificador se prueban los clasificadores de votación, promedio, *K vecinos más cercanos* (KNN), SVM, *random forest* y ANN.

Para el *ensemble* se utilizan las predicciones de los 90 modelos seleccionados en los anteriores apartados como variables de entrada, manteniendo la validación cruzada en nueve grupos. Es decir, se mantiene durante todo el proceso de construcción un conjunto de datos de validación oculto para el entrenamiento de los modelos, a excepción de para la ANN en la etapa del *ensemble*, puesto que en su proceso de entrenamiento debe ver periódicamente el conjunto de validación para comprobar si realmente se está ajustando a la estructura latente de los datos o se está sobreajustando a las muestras de entrenamiento.

Los modelos clasificadores *ensemble* que no son ANN son susceptibles de empeorar ante la incorporación de modelos que no aportan información adicional a la clasificación. Es por ello que se realiza una regresión *stepwise*, en la que se empieza con el modelo clasificador que arroja mayor AUC medio (de los nueve grupos de validación cruzada), y se van añadiendo los modelos clasificadores simples que mejoran más el AUC del modelo *ensemble* final. Esto se muestra en la *Figura 34*.

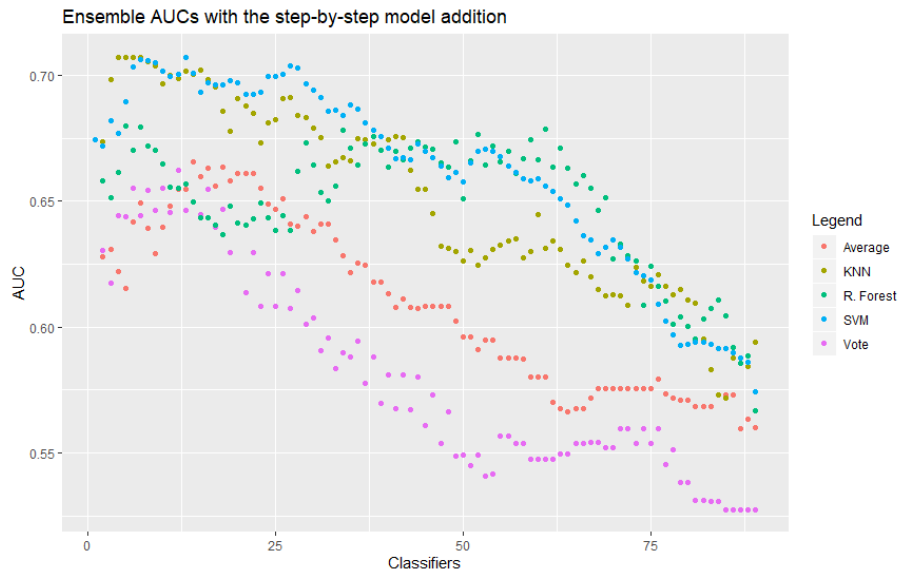


Figura 34. AUC del ensemble con la inclusión de clasificadores, dependiendo de la técnica de clasificación superior.

Los modelos *ensemble* de votación y de promedio no ofrecen mejores resultados en ningún caso que el mejor de los modelos clasificadores de la base, el sNPLSDA escalado en el modo 3 con 50 puntos de corriente y tres electrodos, el mismo que sin variables adicionales. Los modelos *ensemble* han arrojado un AUC medio máximo de 0.666 en ambos, contra 0.675 del sNPLSDA.

Paralelamente, de la misma forma que para sin variables adicionales, se entrena una ANN con todos los modelos clasificadores a su entrada. Se evalúan 27 arquitecturas de dos capas distintas, con 8, 16 y 32 neuronas por capa, con regularizadores L^2 de 0.1, 0.001 y 10^{-6} .

El que mejores resultados arroja es con un L^2 de 10^{-6} , 16 neuronas a una capa, con un AUC medio de 0.559. Un valor realmente bajo, sustancialmente por debajo del 0.675 del mejor de los modelos clasificadores de la base.

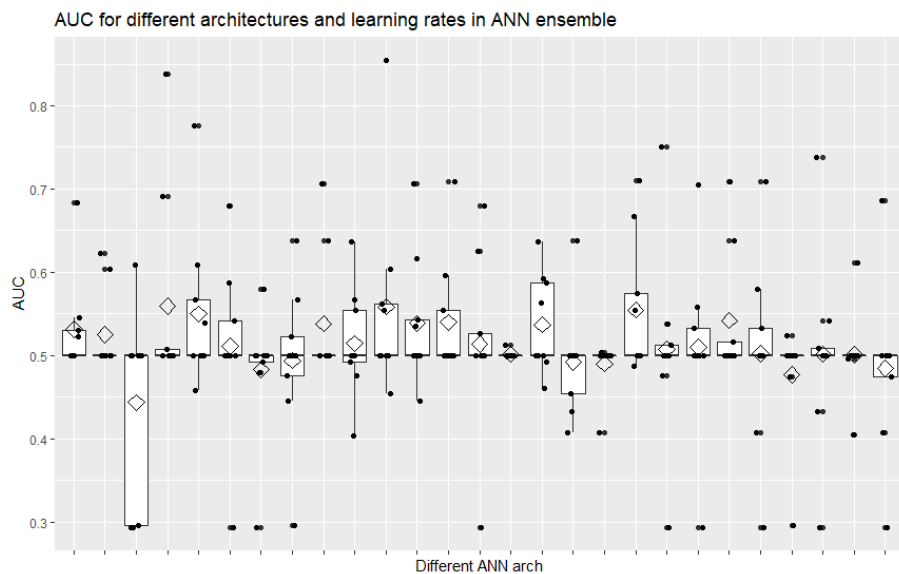


Figura 35. AUC para cada una de las arquitecturas y learning rates de ANN en ensemble.

Se escoge el mejor de los modelos de acuerdo al número de clasificadores básicos incluidos, para las técnicas de clasificación KNN, RF y SVM, y se comparan de acuerdo a sus AUC y tasas de aciertos para cada uno de los grupos de la validación cruzada, primero de acuerdo a los clasificadores de base que maximizan el AUC promedio, y luego para los que maximizan la tasa de aciertos.

Los resultados que se muestran son la ampliación de los valores calculados en la regresión *stepwise* anterior para cada uno de los modelos, pero ampliando en los puntos especificados para cada uno de los modelos *ensemble*.

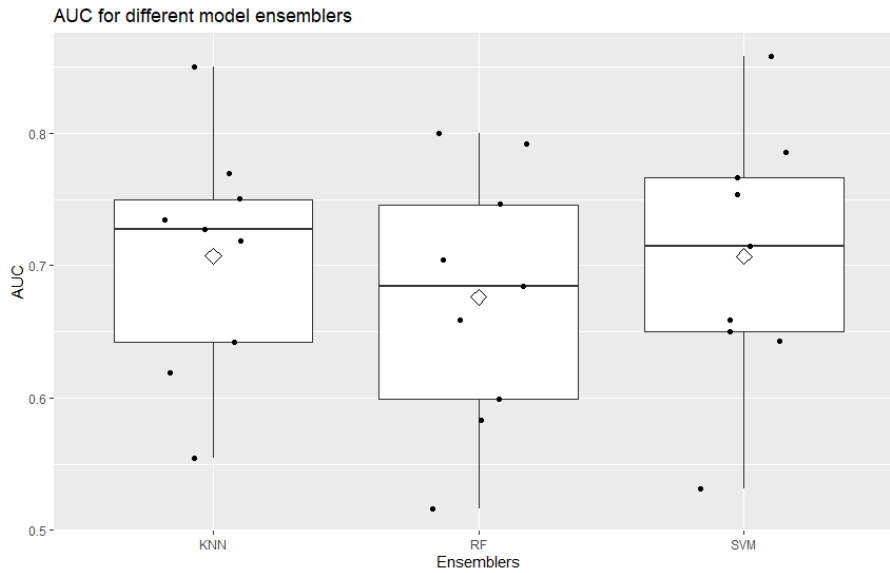


Figura 36. Modelos ensemble de acuerdo a su AUC.

Como puede observarse, los tres modelos presentan un valor de AUC promedio similar, presentando KNN una media y mediana ligeramente superior que el resto. Para salir de dudas, se muestran las tasas de aciertos de los modelos para el conjunto de clasificadores de base que maximizan el AUC, así como para el conjunto que maximizan la tasa de aciertos de cada uno de los modelos.

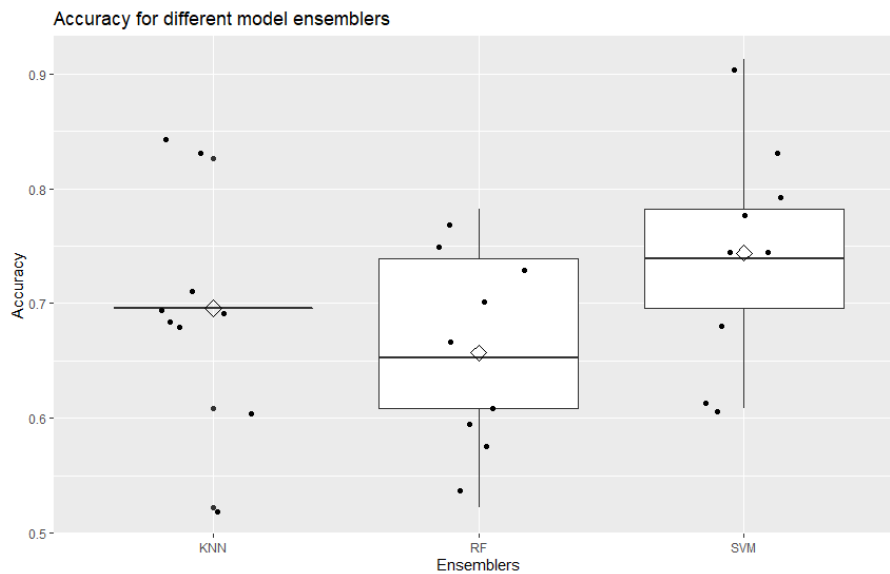


Figura 37. Tasas de aciertos por modelo ensemble para los modelos clasificadores de base que maximizan AUC promedio.

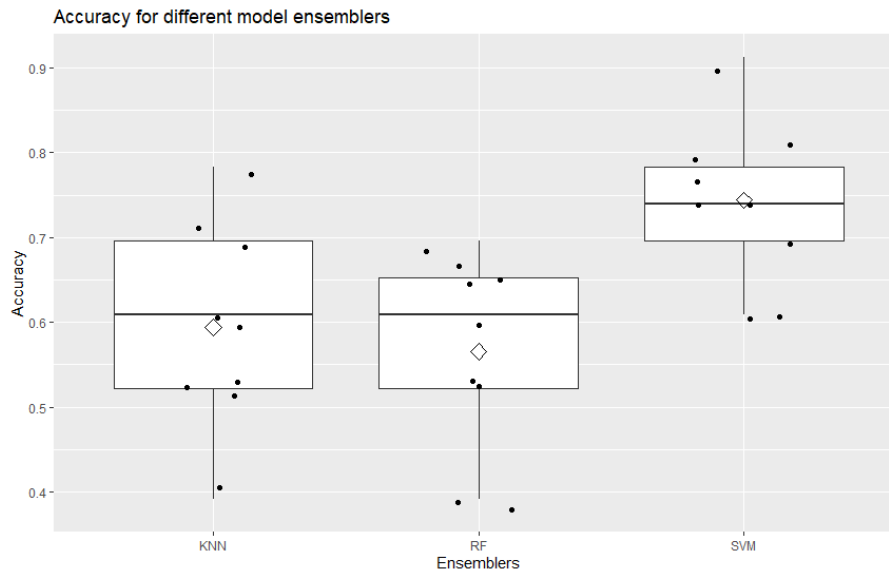


Figura 38. Tasas de aciertos por modelo ensemble para los modelos clasificadores de base que maximizan tasa de aciertos promedio.

Para comparar los resultados y observar la magnitud y significación estadística de las diferencias de los métodos en las tasas de aciertos maximizadas, se realiza un ANOVA.

Tabla 13. ANOVA de los modelos *ensemble*.

	GdL	SC	SCM	Valor-F	Valor-P
Modelos	2	0.1656	0.08283	6.703	0.00487
Residuos	24	0.2966	0.01236		

Existen diferencias estadísticamente significativas entre la media de tasa de aciertos de los tres modelos *ensemble*. Como puede observarse a continuación con los intervalos de *Tukey*, esta diferencia sólo es significativa del SVM a los otros dos modelos.

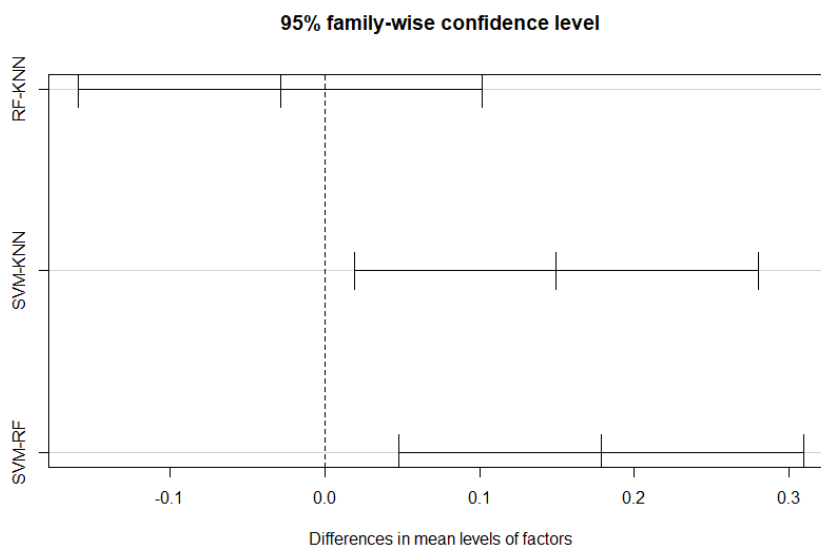


Figura 39. Tukey HSD para los modelos *ensemble*.

Por tanto, como modelo final se escoge el SVM.

6.2.6. MODELO FINAL

A continuación, se muestran el AUC, tasa de aciertos, sensibilidad y especificidad del modelo final.

El cálculo de los intervalos de confianza, al 95%, de cada uno de los estadísticos, se realiza en base al valor crítico de la distribución t de Student para 9 muestras.

$$\bar{x} - t_{n-1} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1} \cdot \frac{s}{\sqrt{n}}$$

Tabla 14. Modelo final.

	Estimación	Amplitud del intervalo de confianza (95%)
AUC	0.707	±0.079
Tasas de aciertos	0.744	±0.080
Sensibilidad	0.739	±0.150
Especificidad	0.740	±0.119

7. CONCLUSIONES

En el presente proyecto se ha desarrollado un algoritmo clasificador binomial para la detección precoz y seguimiento, con bajo coste económico y temporal y de forma no invasiva, del cáncer de vejiga.

Para ello, en primer lugar, se ha explorado el uso de distintas técnicas de clasificación, con distintos pretratamientos, de acuerdo a la estructura tridimensional de los datos (muestras, electrodos y puntos de corriente) y distintas técnicas de reducción de la dimensión (PCA, PFA polinomial de segundo, tercer y cuarto orden y ANOVA).

Estas técnicas han consistido en regresión de componentes principales (PCR), regresión de factores principales (PFR), máquinas de soporte vectorial (SVM) sin y con reducción de la dimensión (PCA, PFA, dando lugar a PCSVM, PFSVM), redes neuronales artificiales (ANN) sin y con reducción de la dimensión (similar a con SVM), y análisis discriminante N-dimensional mediante regresión en mínimos cuadrados parciales dispersa (sNPLSDA). Muchas de estas técnicas mejoraban el análisis discriminante mediante regresión de mínimos cuadrados parciales (PLSDA) sobre el set de datos sin selección de variables ni eliminación de *outliers*, siendo el que presentaba mayor diferencia en el AUC medio en una validación cruzada a nueve grupos el sNPLSDA con escalado en el tercer modo, 50 puntos de corriente y tres electrodos, con un resultado de 0.677 frente a los 0.597 del modelo de línea de base.

Mediante el *ensemble* de los mejores modelos clasificadores mediante SVM ha sido posible mejorar el AUC promedio hasta el 0.726, obteniendo una tasa de aciertos del 77.3%, una sensibilidad del 76.4% y

una especificidad del 77.5%, obteniéndose con KNN resultados sin diferencias estadísticamente significativas.

Sin embargo, existía la posibilidad que la información discriminante se encontrara en valores asociados a los puntos de corriente de una forma no-lineal, es por eso que se obtuvieron variables adicionales de frecuencia, pendientes locales, áreas bajo pulsos y valores de corriente iniciales. Estos bloques de variables, junto a los puntos, se sometieron a un escalado Pareto para reducir la influencia de los bloques con más variables respecto a los de menos, y se repitieron los métodos clasificadores anteriores.

Con variables adicionales, el PLSDA utilizado como modelo de base arroja resultados similares, con un AUC medio de 0.595 con diez variables latentes. El mejor de los modelos clasificadores sin *ensemble* vuelve a ser el mismo, el sNPLSDA escalado en el modo 3, con tres electrodos y 50 puntos de corriente, con un AUC medio de 0.675. Hasta este punto los resultados son equivalentes a los modelos sin variables adicionales. En la combinación de modelos, el modelo *ensemble* basado en SVM es significativamente superior a KNN y RF (que no presentan diferencias significativas entre sí). Sin embargo, los resultados obtenidos son ligeramente inferiores a los obtenidos por el *ensemble* SVM sin variables adicionales: un AUC medio de 0.707, una tasa de aciertos de 74.4%, sensibilidad de 73.9% y especificidad de 74.0%. Las variables adicionales sólo introducen ruido en el modelo, que hace que empeore ligeramente su poder clasificador.

Comparando los resultados obtenidos en este proyecto (los del modelo clasificador *ensemble* basado en SVM sin variables adicionales) con los resultados obtenidos por *Loras et al. (2016)* [7], se ha obtenido una menor sensibilidad (76.4% contra 86.4%) y especificidad (77.5% contra 80.9%), no batiendo los resultados obtenidos mediante espectroscopia de RMN. La ventaja de la lengua electrónica voltamétrica, pese a su menor poder de clasificación, es su capacidad de medición *in situ* e instantánea, de forma económica y sencilla (sin necesidad de personal altamente cualificado), características que no comparte con la espectroscopia de RMN, por lo que puede ser utilizada como método de cribado o *screening*, para posteriormente utilizar otras técnicas más sofisticadas, más invasivas o con mayor sensibilidad y especificidad.

8. TRABAJOS FUTUROS

La aproximación de la lengua electrónica voltamétrica como instrumento universal, que permite la detección de una amplísima gama de compuestos y multitud de aplicaciones tiene serias limitaciones si el algoritmo clasificador o predictor de la misma se basa en trenes de pulsos arbitrarios o estándar.

Para mejorar la capacidad de detección, y en consecuencia de clasificación, del instrumento de medida, se hace imperioso no sólo mejorar los algoritmos clasificadores, sino realizar un estudio sobre qué electrodos presentan reactividad a los metabolitos relacionados con la presencia de cáncer, y en base a qué potenciales se presentan con mayor claridad las variaciones de corriente asociadas, en ausencia o con la menor contribución posible de otros metabolitos también presentes.

La potencialidad de la lengua electrónica para esta aplicación queda más que demostrada con el presente trabajo, pero para su aplicación clínica debe atenderse a los metabolitos identificados en los últimos estudios del grupo [5-7], rediseñando el tren de pulsos a aplicar y, si es necesario, los electrodos que la componen. Para una traslación de esta estrategia como apoyo al diagnóstico en el entorno clínico resultará necesario también la utilización de electrodos desechables con alta reproducibilidad para evitar los procesos de acondicionamiento necesarios en la lengua electrónica utilizada en este estudio.

9. RECURSOS INFORMÁTICOS UTILIZADOS

1. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
2. David Hervas (2018). sNPLS: NPLS Regression with L1 Penalization. R package version 0.3.31. <https://CRAN.R-project.org/package=sNPLS>
3. Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9), 1-20. URL <http://www.jstatsoft.org/v11/i09/>
4. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.1. <https://CRAN.R-project.org/package=dplyr>
5. Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5.
6. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
7. Andrie de Vries and Brian D. Ripley (2016). gg dendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'. R package version 0.1-20. <https://CRAN.R-project.org/package=ggdendro>
8. Jeremy VanDerWal, Lorena Falconi, Stephanie Januchowski, Luke Shoo and Collin Storlie (2019). SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1-221.1. <https://CRAN.R-project.org/package=SDMTools>
9. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-2. <https://CRAN.R-project.org/package=e1071>
10. Alexandros Karatzoglou, Alex Smola, Kurt Hornik, Achim Zeileis (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9), 1-20. URL <http://www.jstatsoft.org/v11/i09/>
11. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
12. Torgo, L. (2016). Data Mining with R, learning with case studies, 2nd edition Chapman and Hall/CRC. URL: <http://ltorgo.github.io/DMwR2>
13. Stephen Milborrow (2019). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.7. <https://CRAN.R-project.org/package=rpart.plot>
14. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. *R News* 2(3), 18--22.
15. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.
16. Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount and Shengqiao Li (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.3. <https://CRAN.R-project.org/package=FNN>
17. Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. <https://CRAN.R-project.org/package=caret>
18. Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. <https://CRAN.R-project.org/package=tidyverse>

19. Bjørn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2019). pls: Partial Least Squares and Principal Component Regression. R package version 2.7-1. <https://CRAN.R-project.org/package=pls>
20. Sebastien Le, Julie Josse, Francois Husson (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18. 10.18637/jss.v025.i01
21. Alboukadel Kassambara and Fabian Mundt (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. <https://CRAN.R-project.org/package=factoextra>
22. Daniel Adler, Duncan Murdoch and others (2019). rgl: 3D Visualization Using OpenGL. R package version .100.19. <https://CRAN.R-project.org/package=rgl>
23. Tal Galili (2015). dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics*. DOI: 10.1093/bioinformatics/btv428
24. Dongjun Chung, Hyonho Chun and Sunduz Keles (2019). spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R package version 2.2-3. <https://CRAN.R-project.org/package=spls>
25. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20), 7881. URL: <http://rocr.bioinf.mpi-sb.mpg.de>.
26. JJ Allaire and François Chollet (2019). keras: R Interface to 'Keras'. R package version 2.2.4.1. <https://CRAN.R-project.org/package=keras>
27. JJ Allaire and Yuan Tang (2019). tensorflow: R Interface to 'TensorFlow'. R package version 1.13.1. <https://CRAN.R-project.org/package=tensorflow>

10. BIBLIOGRAFÍA

1. Su, H., Jiang, H., Tao, T., Kang, X., Zhang, X., Kang, D., ... & Zhang, J. (2019). Hope and challenge: Precision medicine in bladder cancer. *Cancer medicine*, 8(4), 1806-1816.
2. Tang, D. H., & Chang, S. S. (2015). Management of carcinoma in situ of the bladder: best practice and recent developments. *Therapeutic advances in urology*, 7(6), 351-364.
3. Yang, B., Liao, G. Q., Wen, X. F., Chen, W. H., Cheng, S., Stolzenburg, J. U., ... & Neuhaus, J. (2017). Nuclear magnetic resonance spectroscopy as a new approach for improvement of early diagnosis and risk stratification of prostate cancer. *Journal of Zhejiang University-SCIENCE B*, 18(11), 921-933.
4. Tilki, D., Burger, M., Dalbagni, G., Grossman, H. B., Hakenberg, O. W., Palou, J., ... & Zlotta, A. R. (2011). Urine markers for detection and surveillance of non-muscle-invasive bladder cancer. *European urology*, 60(3), 484-492.
5. Loras, A., Trassiera, M., Sanjuan-Herráez, D., Martínez-Bisbal, M. C., Castell, J. V., Quintás, G., & Ruiz-Cerdá, J. L. (2018). Bladder cancer recurrence surveillance by urine metabolomics analysis. *Scientific reports*, 8(1), 9172.
6. Loras, A., Suárez-Cabrera, C., Martínez-Bisbal, M. C., Quintás, G., Paramio, J. M., Martínez-Máñez, R., ... & Ruiz-Cerdá, J. L. (2019). Integrative Metabolomic and Transcriptomic Analysis for the Study of Bladder Cancer. *Cancers*, 11(5), 686.
7. Loras, A., Martínez-Bisbal, M. C., Quintás, G., Gil, S., Martínez-Máñez, R., & Ruiz-Cerdá, J. L. (2019). Urinary Metabolic Signatures Detect Recurrences in Non-Muscle Invasive Bladder Cancer. *Cancers*, 11(7), 914.
8. Pascual, L., Campos, I., Vivancos, J. L., Quintás, G., Loras, A., Martínez-Bisbal, M. C., ... & Ruiz-Cerdà, J. L. (2016). Detection of prostate cancer using a voltammetric electronic tongue. *Analyst*, 141(15), 4562-4567.
9. Adnan, A., Hörsten, D. V., & Pawelzik, E. (2017). Rapid prediction of moisture content in intact green coffee beans using near infrared spectroscopy. *Foods*, 6(5), 38.

10. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Epidemiología clínica. Ciencia básica para la medicina clínica. 2ª ed. Madrid: *Editorial médica panamericana*; 1994.
11. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
12. Scholz, M. (2006). Approaches to analyse and interpret biological profile data.
13. Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5), 1299-1319.
14. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3), 1995.
15. Wold, S; Sjöström, M.; Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58(2): 109-130. doi:10.1016/S0169-7439(01)00155-1
16. Hervas, D., Prats-Montalbán, J. M., Lahoz, A., & Ferrer, A. (2018). Sparse N-way partial least squares with R package sNPLS. *Chemometrics and Intelligent Laboratory Systems*, 179, 54-63.
17. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
18. García García, J. A., López Alvarenga, J. C., Jiménez Ponce, F., Ramírez Tapia, Y., Lino Pérez, L., & Reding Bernal, A. (2014). Metodología de la investigación bioestadística y bioinformática en ciencias médicas y de la salud.

ANEXO: SIN VARIABLES ADICIONALES

Los anexos hacen referencia a la selección de número de componentes principales, factores, u otros parámetros de los modelos. Para ello, se atiende al AUC medio mediante validación cruzada (CV) a nueve grupos balanceados, comentándose adicionalmente si alguno de ellos, puntualmente, se encuentra por debajo de 0.5 (es decir, arroja peores resultados que un clasificador aleatorio).

A1. PCR AUTOESCALADO

Con 9 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.606, uno de ellos con un AUC cercano a 0.45.

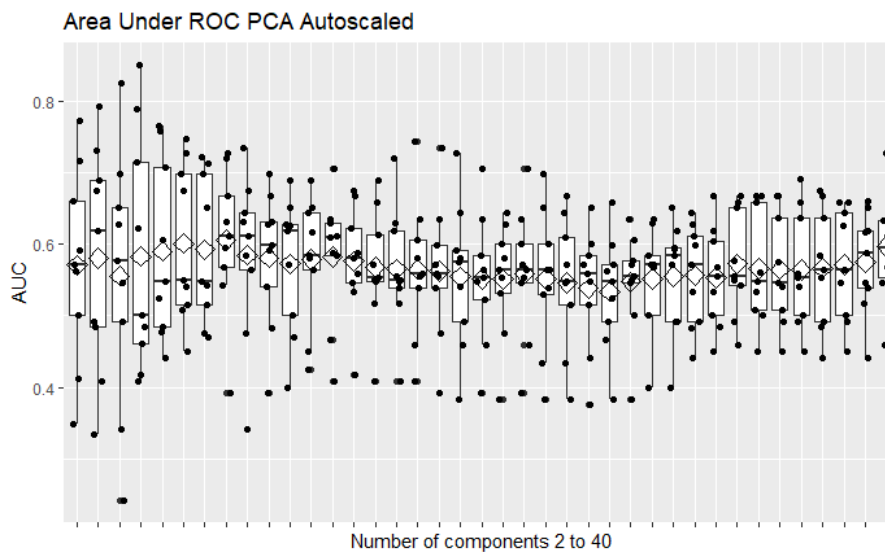


Figura 40. PCR Autoescalado.

A2. PFR POLINOMIAL DE ORDEN 2 AUTOESCALADO

Con 7 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.617, uno de ellos con un AUC cercano a 0.45.

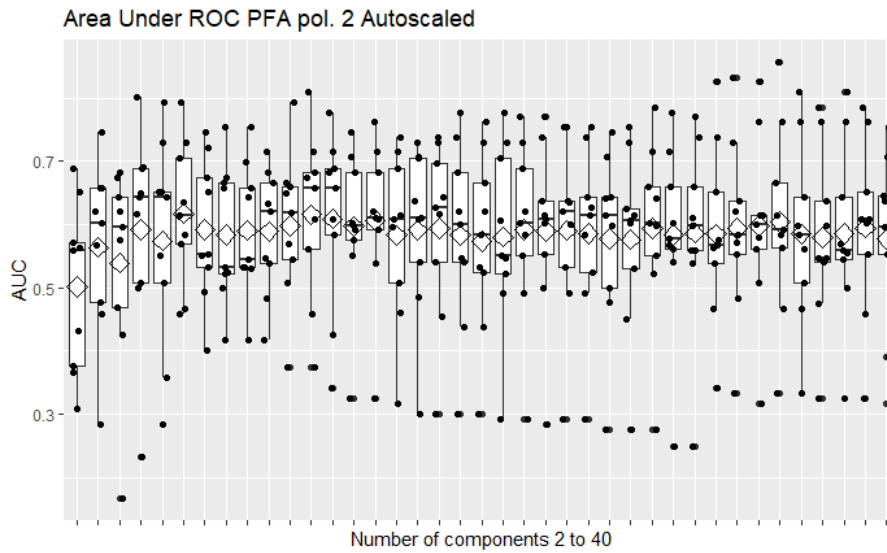


Figura 41. PFR pol. 2 autoescalado.

A3. PFR POLINOMIAL DE ORDEN 3 AUTOESCALADO

Con 17 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.648, dos de ellos con un AUC en torno a 0.3.

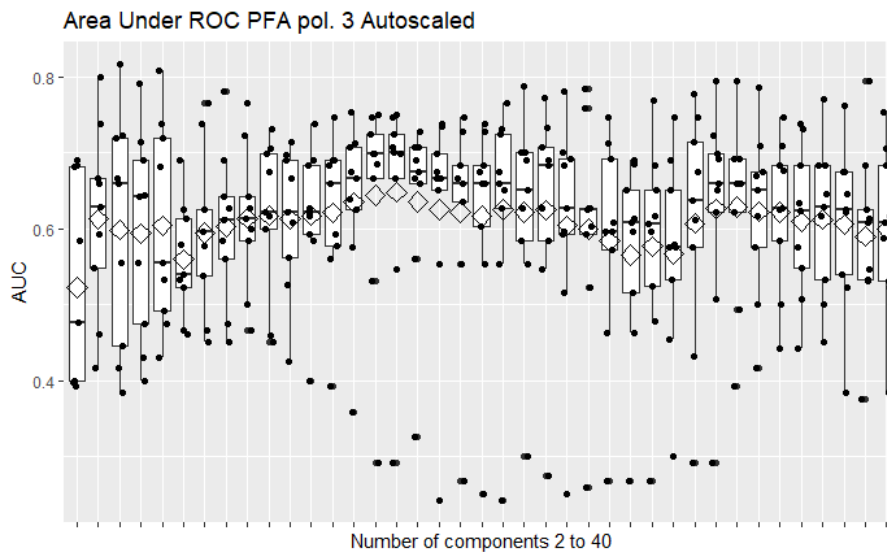


Figura 42. PFR pol. 3 autoescalado.

A4. PFR POLINOMIAL DE ORDEN 4 AUTOESCALADO

Con 13 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.589, tres de ellos con un AUC en torno a 0.4.

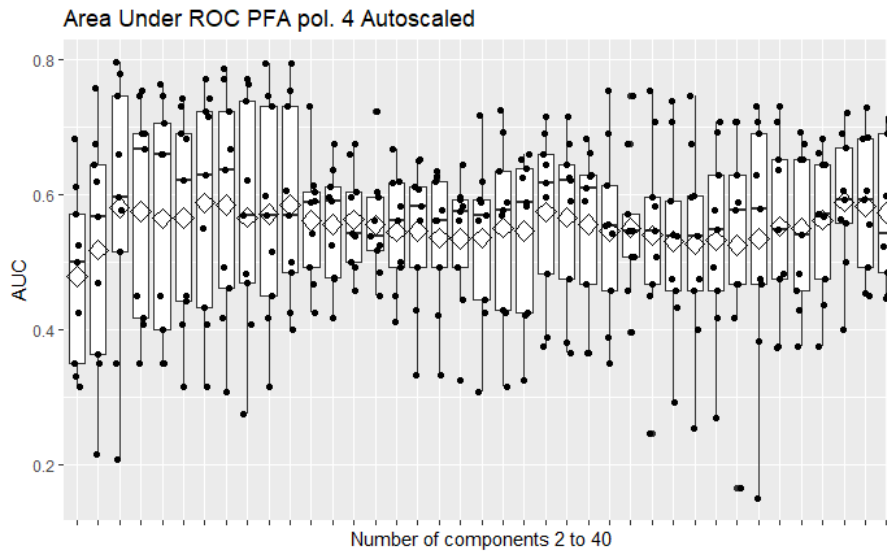


Figura 43. PFR pol. 4 autoescalado.

A5. PFR ANOVA AUTOESCALADO

Con 7 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.610, ninguno de ellos por debajo de 0.5.

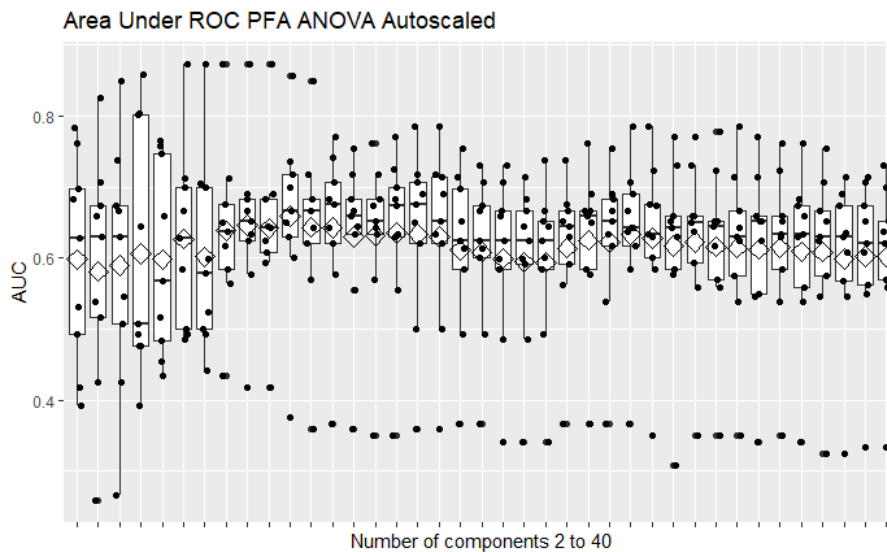


Figura 44. PFR ANOVA autoescalado.

A6. PCR SÓLO CENTRADO

Con 8 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.626, ninguno de ellos por debajo de 0.5.

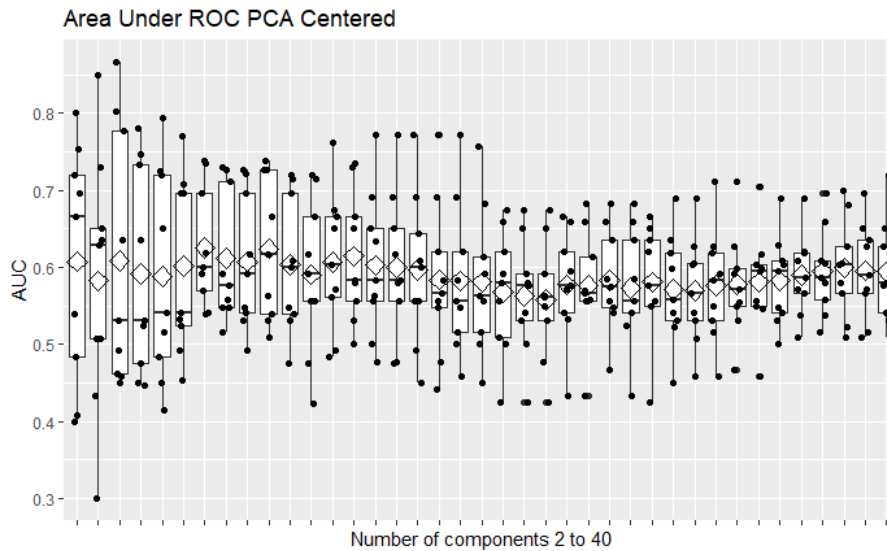


Figura 45. PCR centrado.

A7. PFR POLINOMIAL DE ORDEN 2 SÓLO CENTRADO

Con 2 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.562, dos de ellos en torno a 0.4 y otros dos en torno a 0.25.

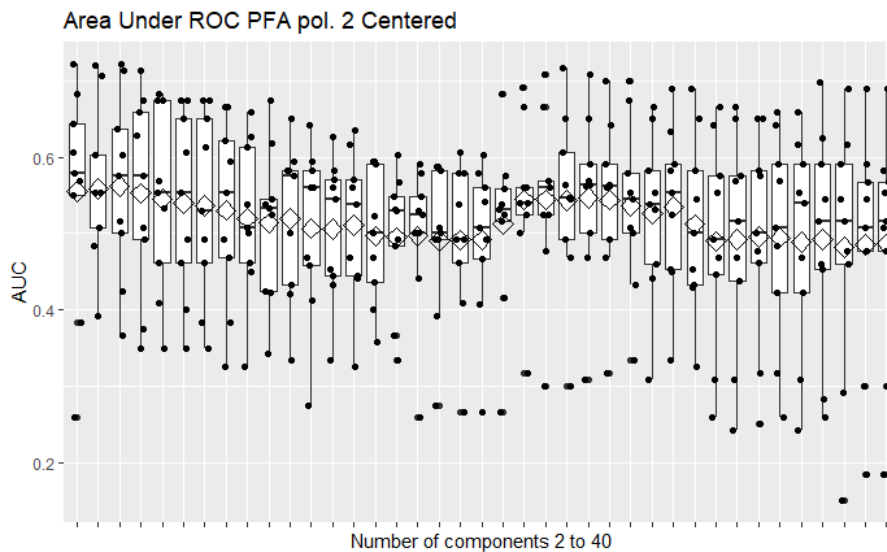


Figura 46. PFR pol. 2 centrado.

A8. PFR POLINOMIAL DE ORDEN 3 SÓLO CENTRADO

Con 8 variables latentes se obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.580, tres de ellos en torno a 0.45.

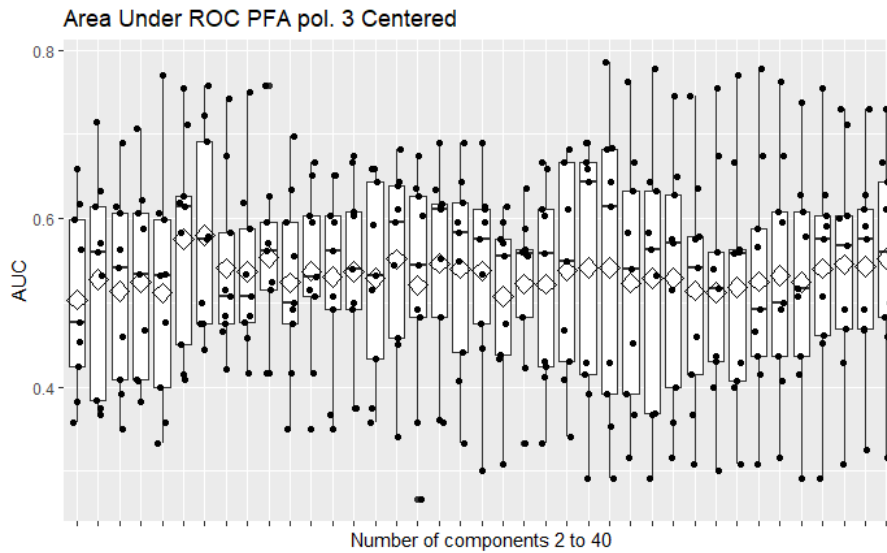


Figura 47. PFR pol. 3 centrado.

A9. PFR POLINOMIAL DE ORDEN 4 SÓLO CENTRADO

Con 5 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.543, uno de ellos en torno a 0.35 y otro en torno a 0.25.

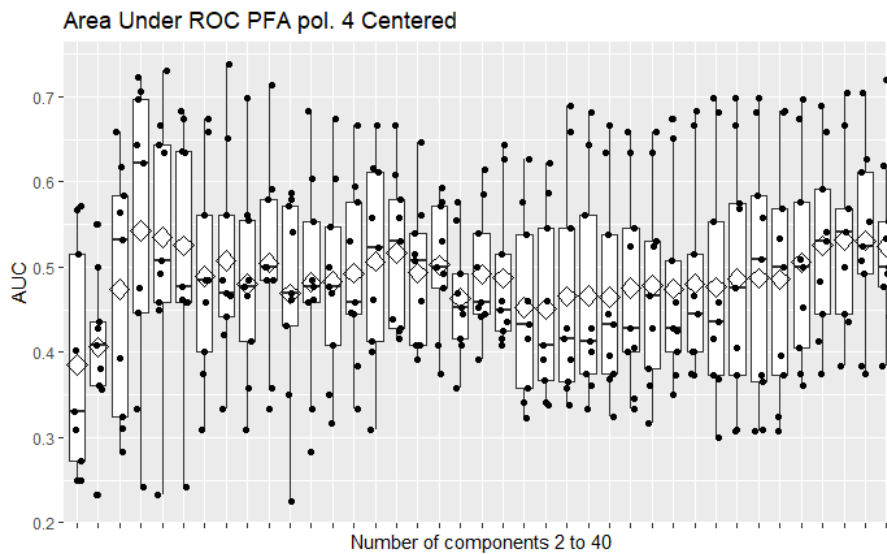


Figura 48. PFR pol. 4 centrado.

A10. PFR ANOVA SÓLO CENTRADO

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.62, de forma que sus AUC mínimos están en torno a 0.5.

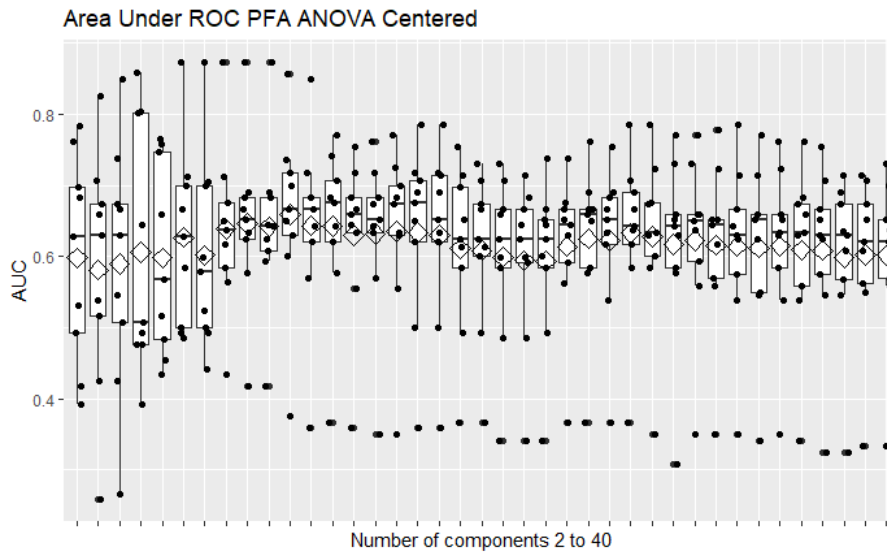


Figura 49. PFR ANOVA centrado.

A11. PCR ESCALADO EN EL MODO 2

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.631, ninguno por debajo de 0.5.

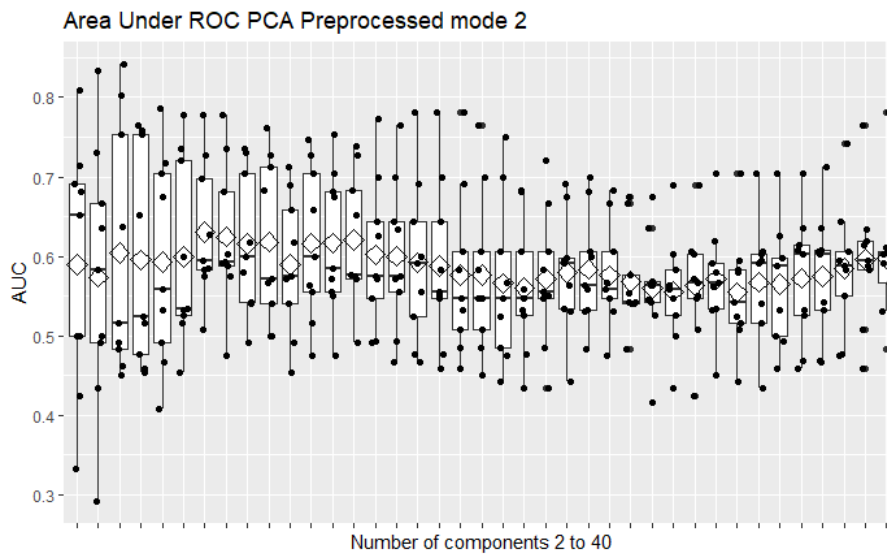


Figura 50. PCR escalado modo 2.

A12. PFR POLINOMIAL DE ORDEN 2 ESCALADO EN EL MODO 2

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.561, con uno de ellos en torno a 0.3 y otro en torno a 0.4.

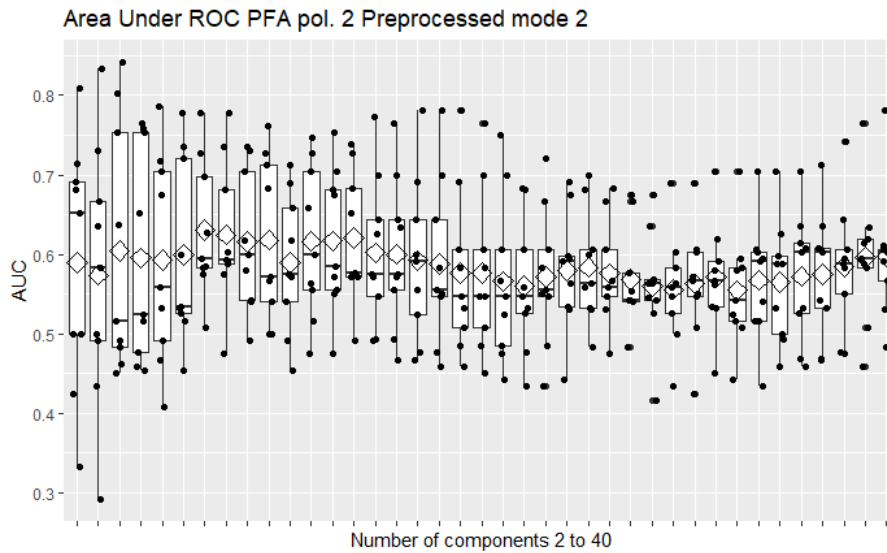


Figura 51. PFR pol. 2 escalado modo 2.

A13. PFR POLINOMIAL DE ORDEN 3 ESCALADO EN EL MODO 2

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.530, con algunos ligeramente por debajo de 0.5.

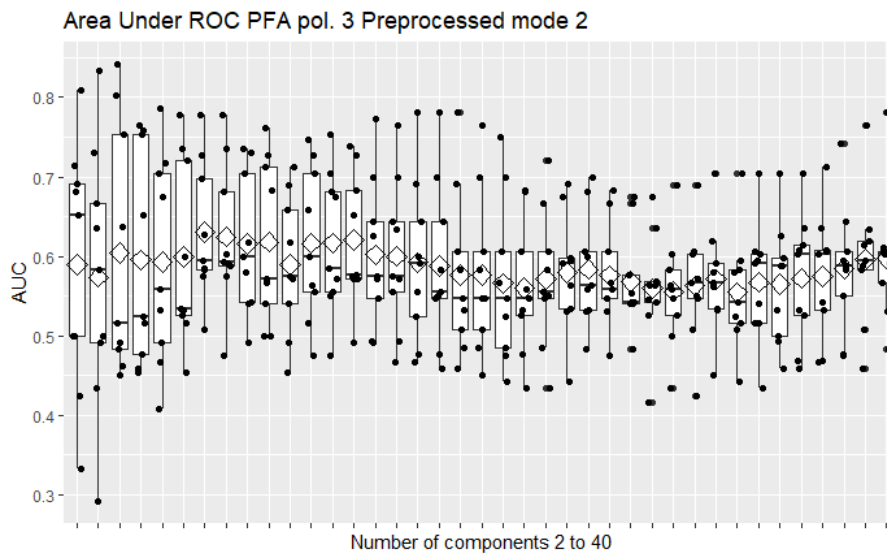


Figura 52. PFR pol. 3 escalado modo 2.

A14. PFR POLINOMIAL DE ORDEN 4 ESCALADO EN EL MODO 2

Con 17 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.549, con algunos ligeramente por debajo de 0.5.

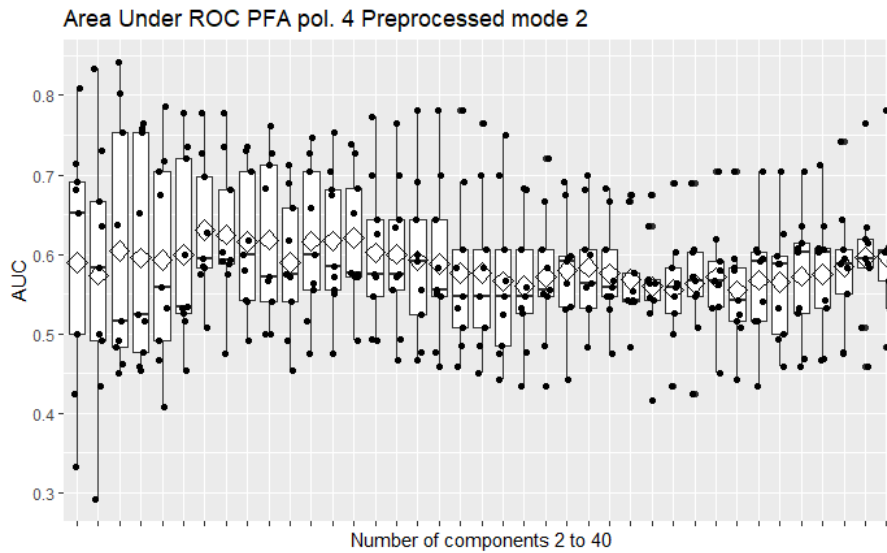


Figura 53. PFR pol. 4 escalado modo 2

A15. PFR ANOVA ESCALADO EN EL MODO 2

Con 9 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.649, todas por encima de 0.5.

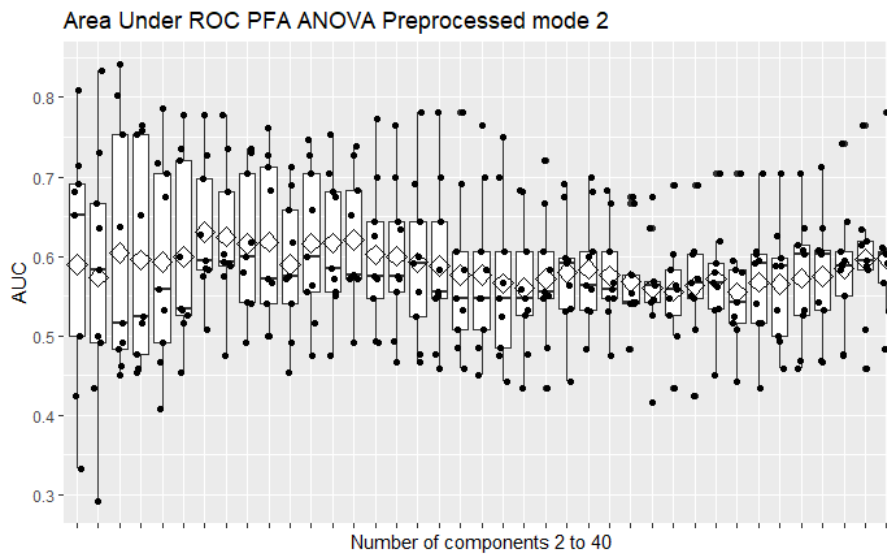


Figura 54. PFR ANOVA escalado modo 2.

A16. PCR ESCALADO EN EL MODO 3

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.622, todos por encima de 0.5.

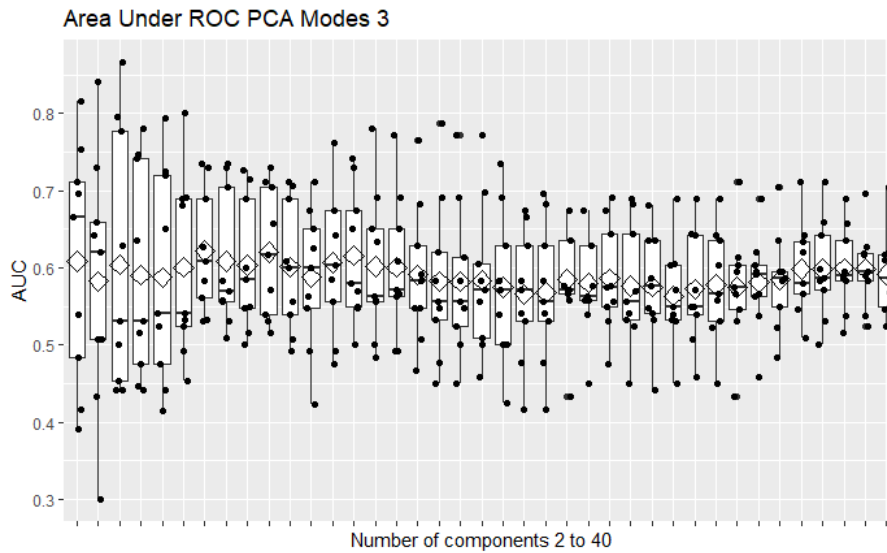


Figura 55. PCR escalado modo 2.

A17. PFR POLINOMIAL DE ORDEN 2 ESCALADO EN EL MODO 3

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.559, dos de ellos en torno a 0.4 y otros dos en torno a 0.25.

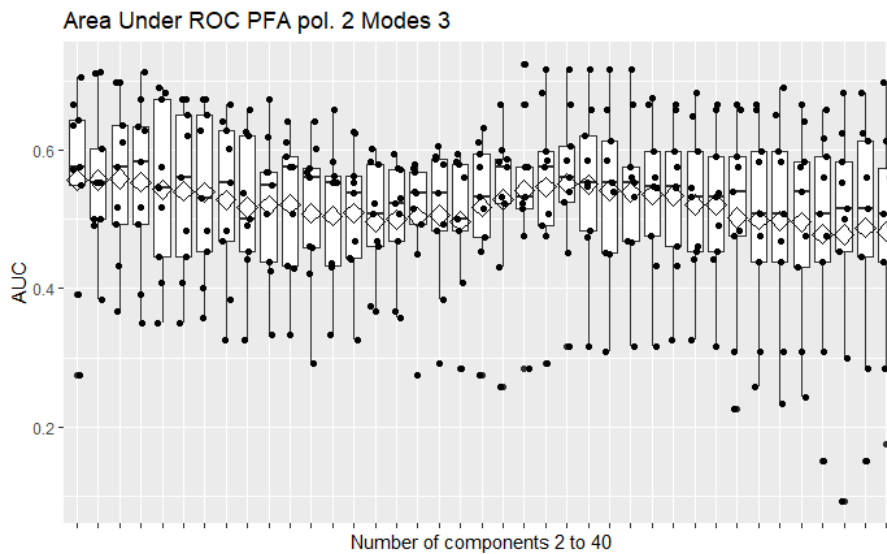


Figura 56. PFA pol. 2 escalado modo 3.

A18. PFR POLINOMIAL DE ORDEN 3 ESCALADO EN EL MODO 3

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.579, tres de ellos en torno a 0.45.

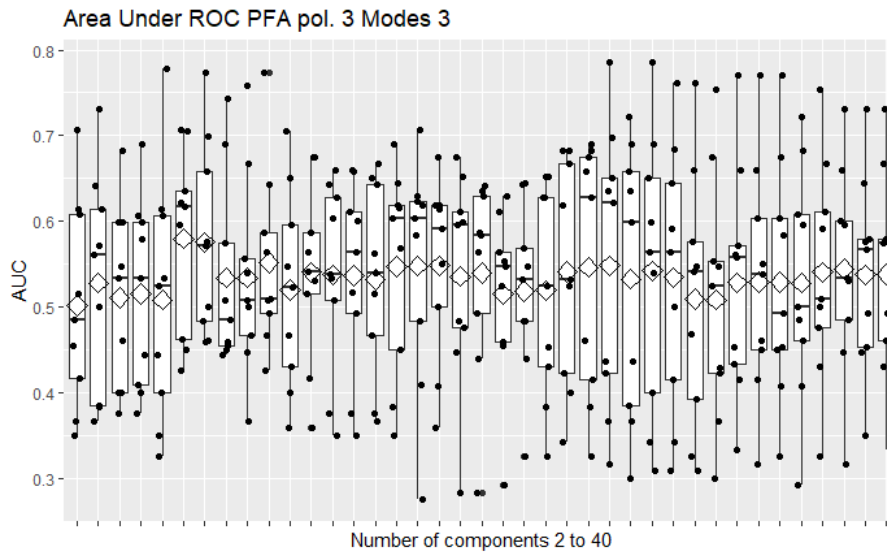


Figura 57. PFA pol. 3 escalado modo 3.

A19. PFR POLINOMIAL DE ORDEN 4 ESCALADO EN EL MODO 3

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.534, dos de ellos en torno a 0.2.

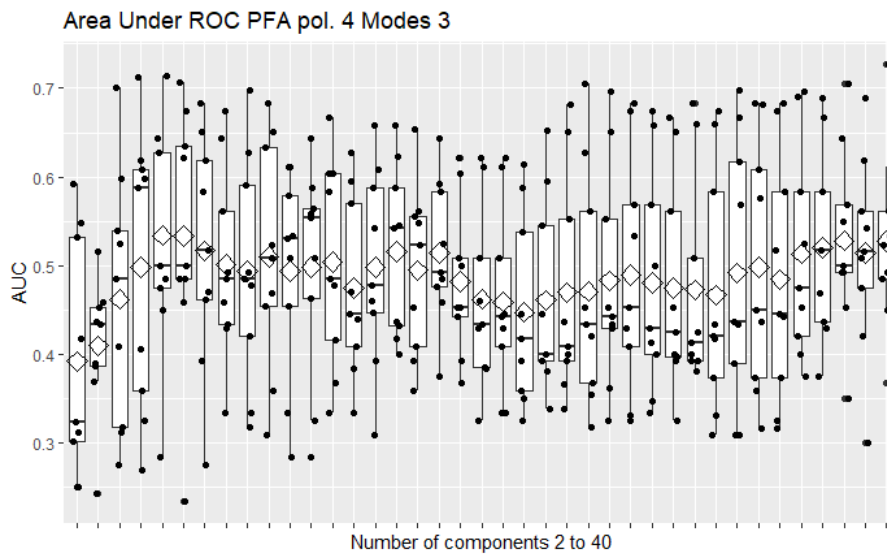


Figura 58. PFA pol. 4 escalado modo 3.

A20. PFR ANOVA ESCALADO EN EL MODO 3

Con 10 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.657, todas por encima de 0.5.

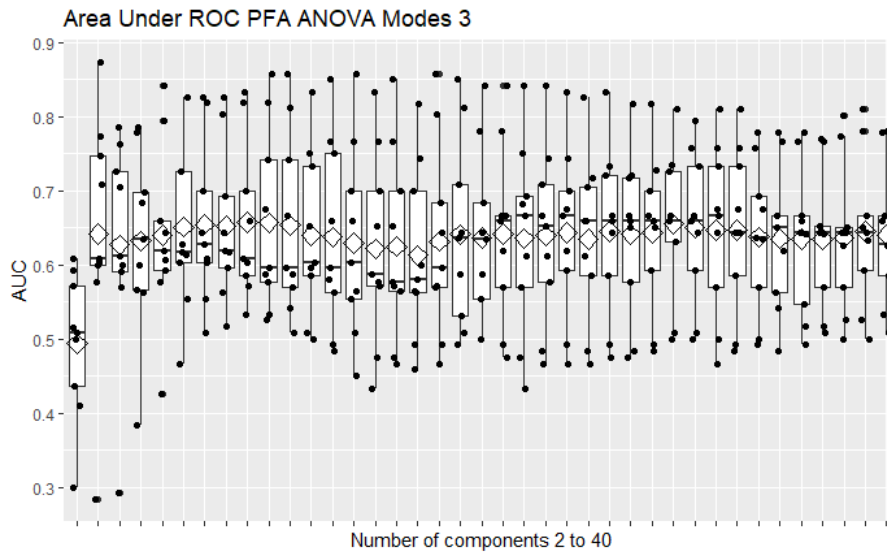


Figura 59. PFA ANOVA escalado modo 3.

A21. PCR ESCALADO EN LOS MODOS 2 Y 3

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.629, todas por encima de 0.5.

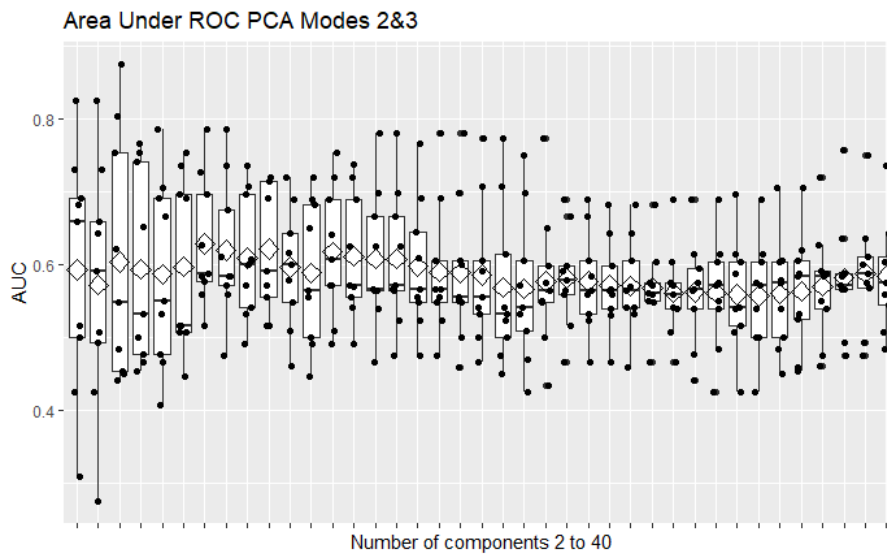


Figura 60. PCR escalado modos 2 y 3.

A22. PFR POLINOMIAL DE ORDEN 2 EN LOS MODOS 2 Y 3

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.559, tres de ellos en torno a 0.4.

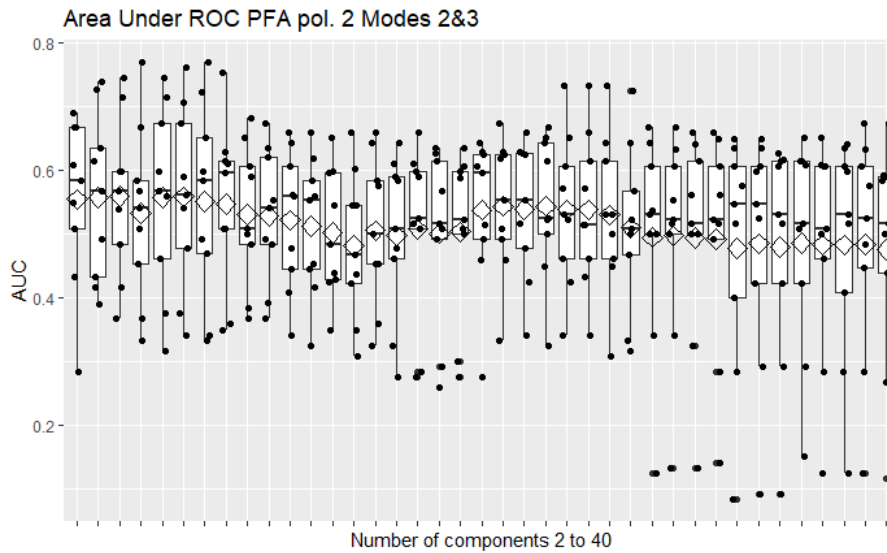


Figura 61. PFR pol. 2 escalado modos 2 y 3.

A23. PFR POLINOMIAL DE ORDEN 3 EN LOS MODOS 2 Y 3

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.52, dos de ellos en torno a 0.4 y otro en torno a 0.45.

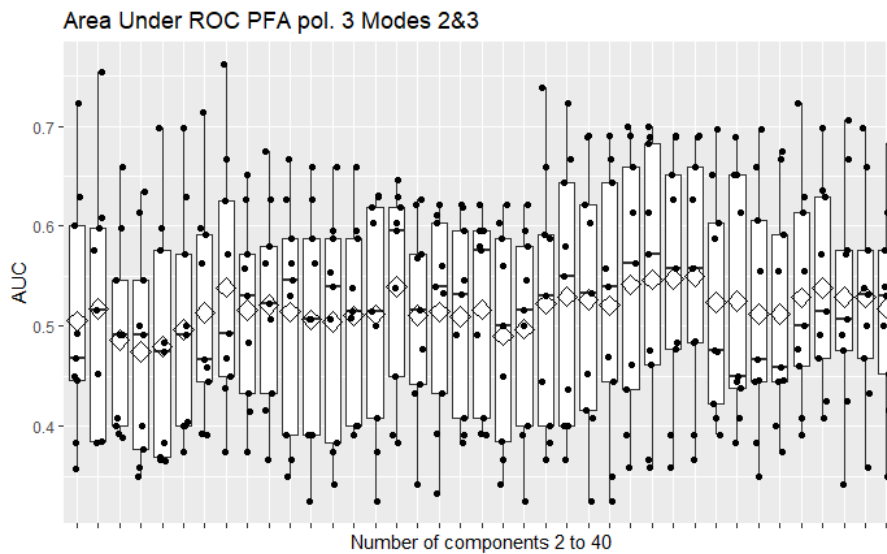


Figura 62. PFR pol. 3 escalado modos 2 y 3.

A24. PFR POLINOMIAL DE ORDEN 4 EN LOS MODOS 2 Y 3

Con 6 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.567, dos de ellos en torno a 0.45 y otro en torno a 0.3.

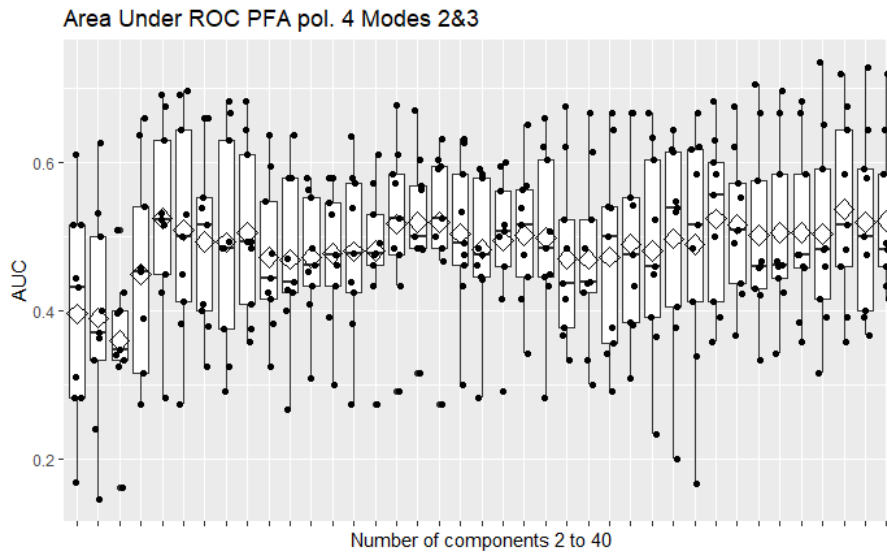


Figura 63. PFR pol. 4 escalado modos 2 y 3.

A25. PFR ANOVA EN LOS MODOS 2 Y 3

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, todos por encima de 0.5.

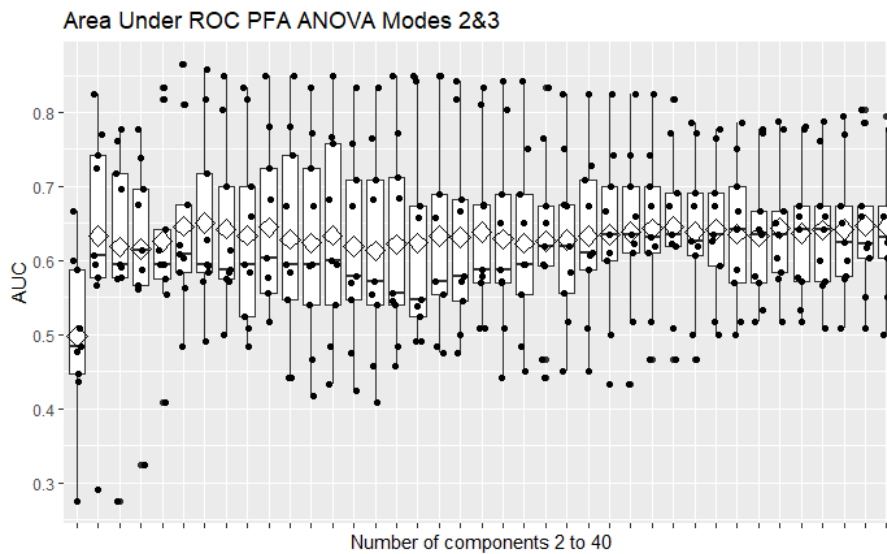


Figura 64. PFR ANOVA escalado modos 2 y 3.

A26. SVM SIN REDUCCIÓN DE LA DIMENSIÓN

Se puede observar que todos los modelos tienen AUC puntuales cercanos o por debajo a 0.4.

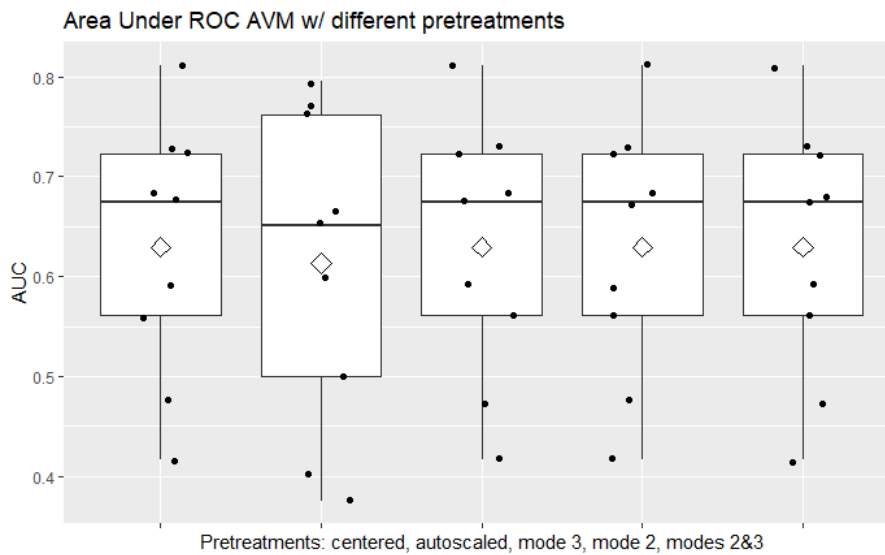


Figura 65. SVM con distintos pretratamientos, sin reducción de la dimensión.

A27. SVM PCA AUTOESCALADO

Con 18 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.631, con dos grupos en torno a 0.4.

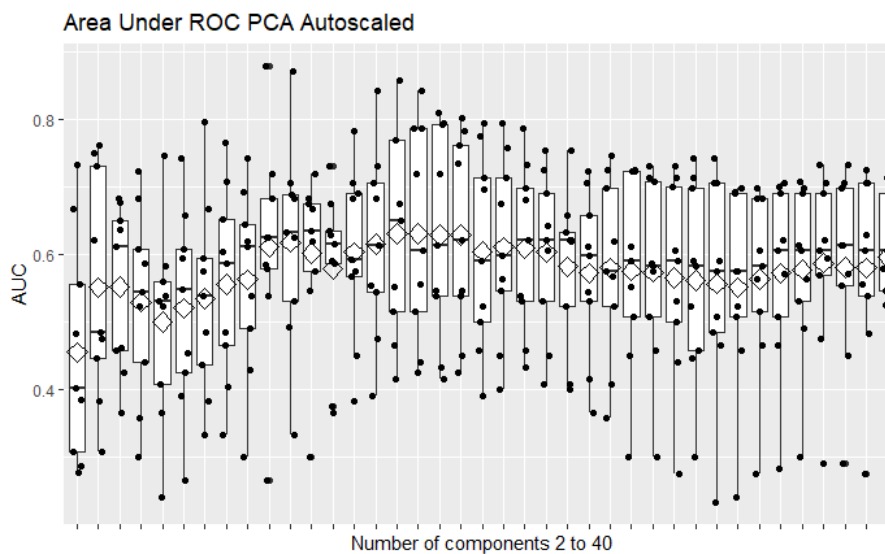


Figura 66. SVM PCA autoescalado.

A28. SVM PFA POL. 2ºG AUTOESCALADO

Con 10 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.605, con un grupo en torno a 0.4 y otro en torno a 0.3.

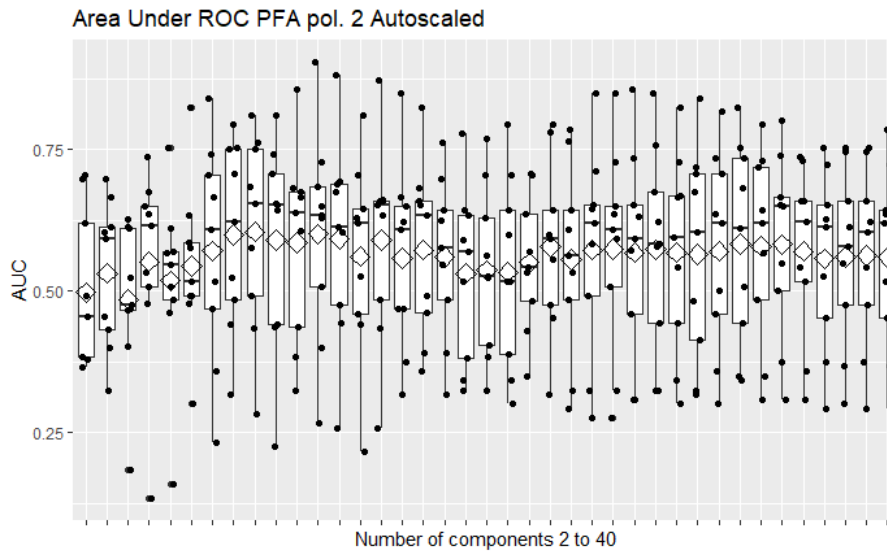


Figura 67. SVM PFA pol. 2 autoescalado.

A29. SVM PFA POL. 3ºG AUTOESCALADO

Con 9 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.576, con dos grupos en torno a 0.45.

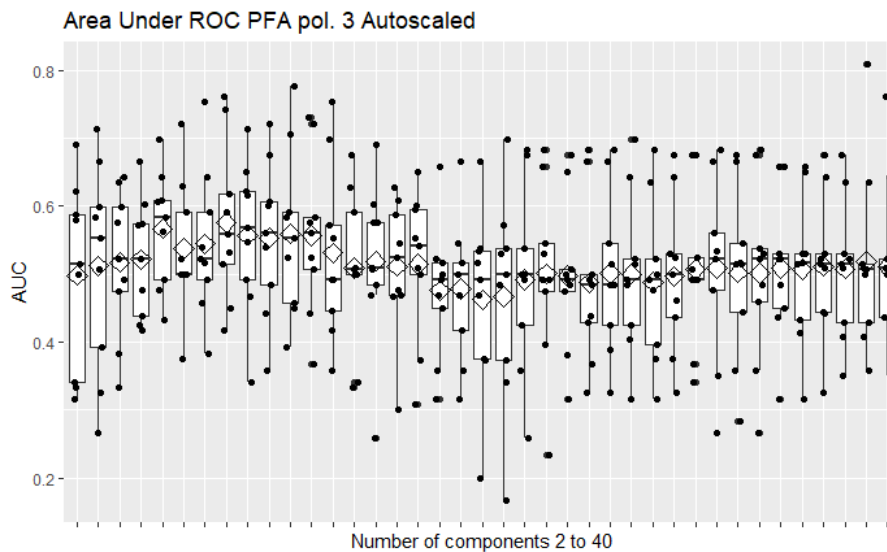


Figura 68. SVM PFA pol. 3 autoescalado.

A30. SVM PFA POL. 4ºG AUTOESCALADO

Con 16 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.581, con un grupo en torno a 0.4.

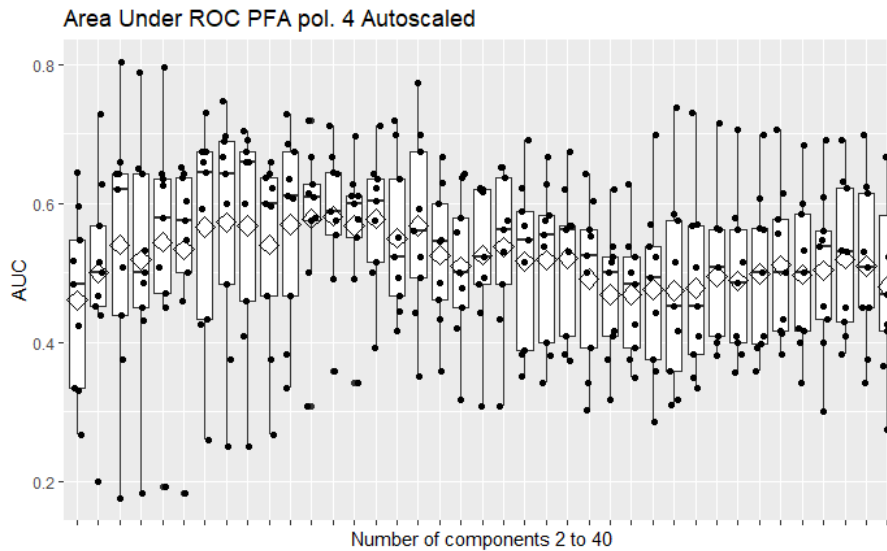


Figura 69. SVM PFA pol. 4 autoescalado.

A31. SVM PFA ANOVA AUTOESCALADO

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.586, con un grupo en torno a 0.4.

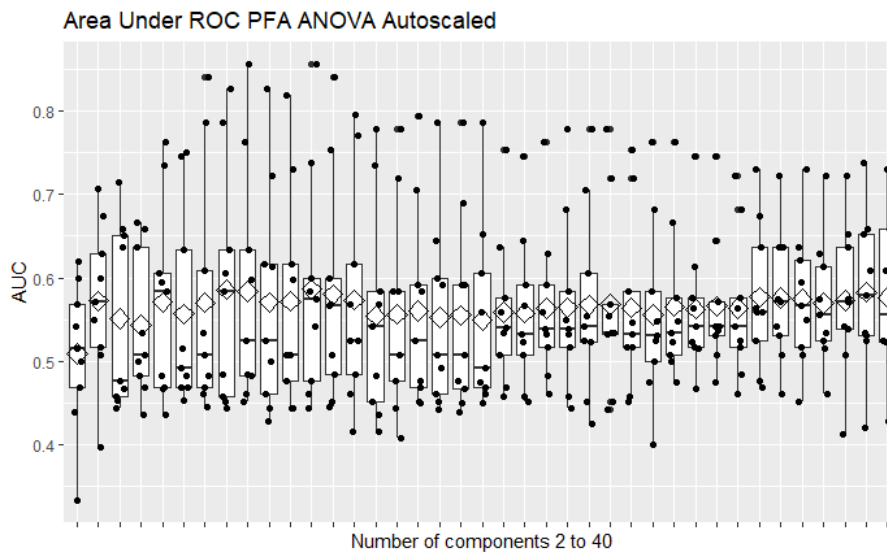


Figura 70. SVM PFA ANOVA autoescalado.

A32. SVM PCA CENTRADO

Con 20 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.671, con todos los grupos por encima de 0.5.

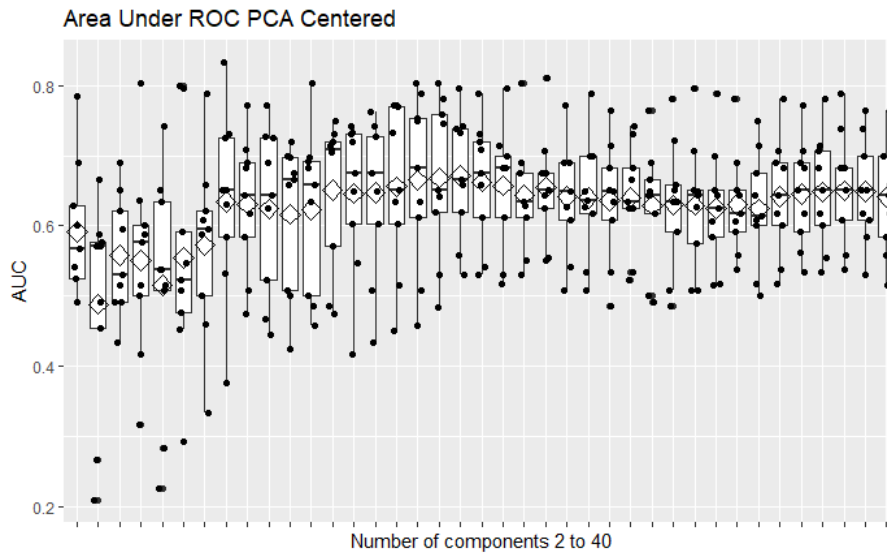


Figura 71. SVM PCA centrado.

A33. SVM PFA POL. 2ºG CENTRADO

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.580, con un grupo en torno a 0.4.

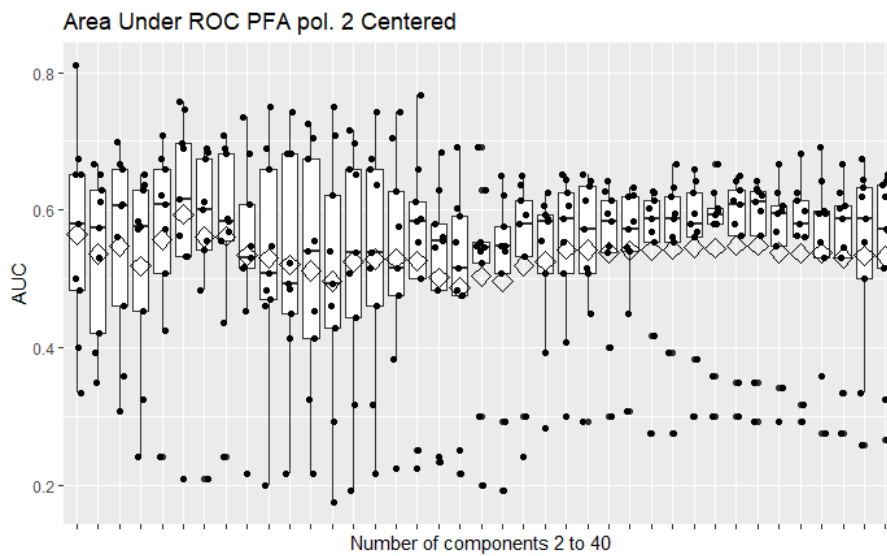


Figura 72. SVM PFA pol. 2 centrado.

A34. SVM PFA POL. 3ºG CENTRADO

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.580, con un grupo en torno a 0.4.

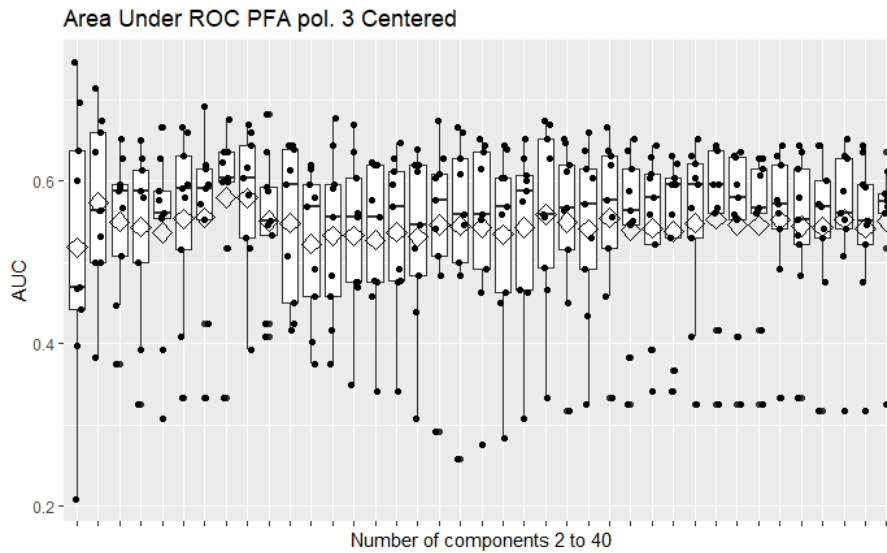


Figura 73. SVM PFA pol. 3 centrado.

A35. SVM PFA POL. 4ºG CENTRADO

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.578, con varios grupos por debajo de 0.5.

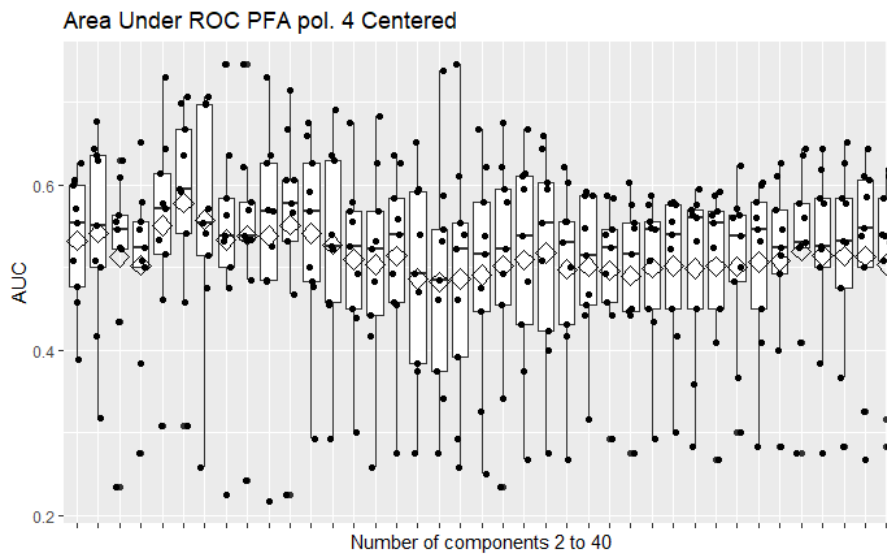


Figura 74. SVM PFA pol. 4 centrado.

A36. SVM PFA ANOVA CENTRADO

Con 12 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.630, con un grupo en torno a 0.35.

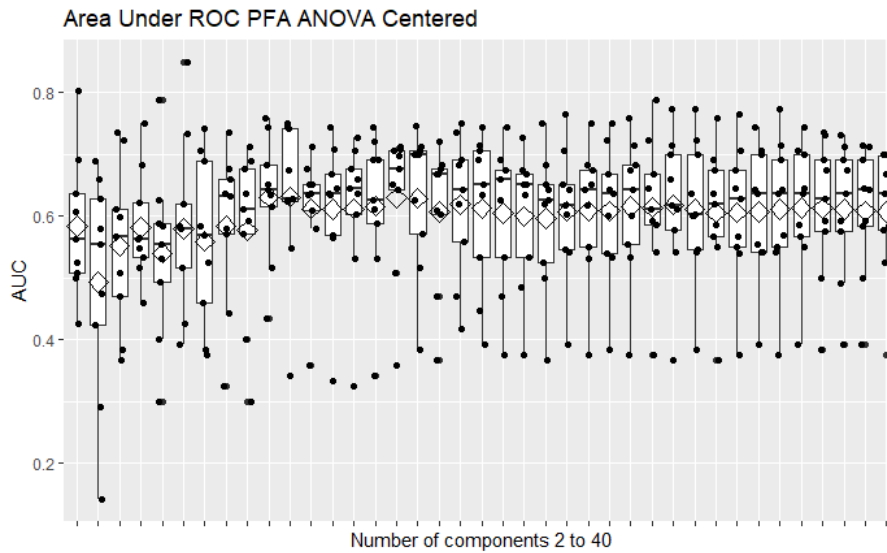


Figura 75. SVM PFA ANOVA centrado.

A37. SVM PCA ESCALADO MODO 2

Con 22 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.662, con todos los grupos por encima de 0.5.

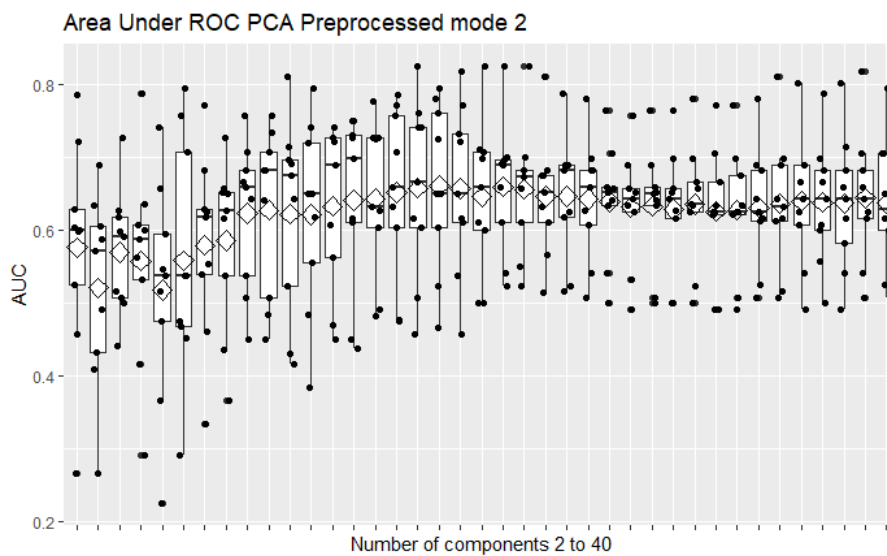


Figura 76. SVM PCA escalado modo 2.

A38. SVM PFA POL. 2ºG ESCALADO MODO 2

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.573, con dos grupos en torno a 0.2.

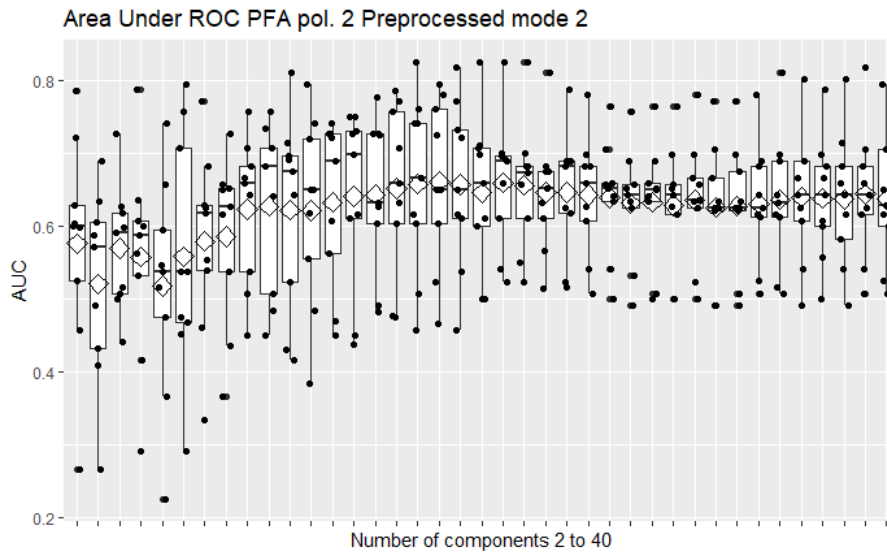


Figura 77. SVM PFA pol. 2 escalado modo 2.

A39. SVM PFA POL. 3ºG ESCALADO MODO 2

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.591, con dos grupos en torno a 0.35.

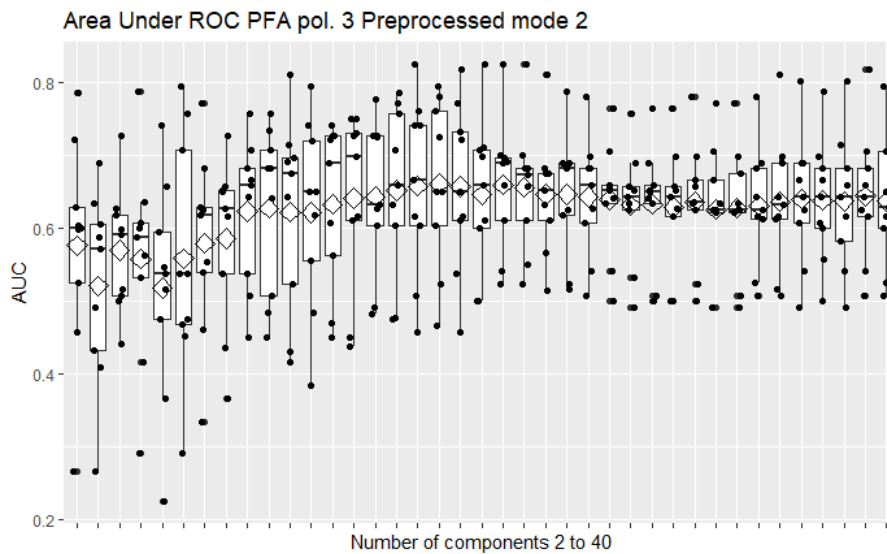


Figura 78. SVM PFA pol. 3 escalado modo 2.

A40. SVM PFA POL. 4ºG ESCALADO MODO 2

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.567, con dos grupos en torno a 0.2.

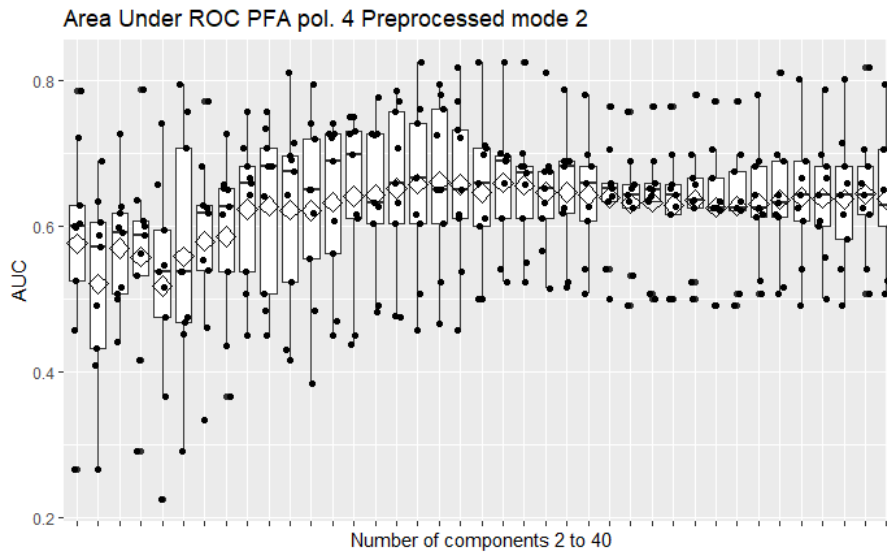


Figura 79. SVM PFA pol. 4 escalado modo 2.

A41. SVM PFA ANOVA ESCALADO MODO 2

Con 32 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, con todos los grupos por encima de 0.5.

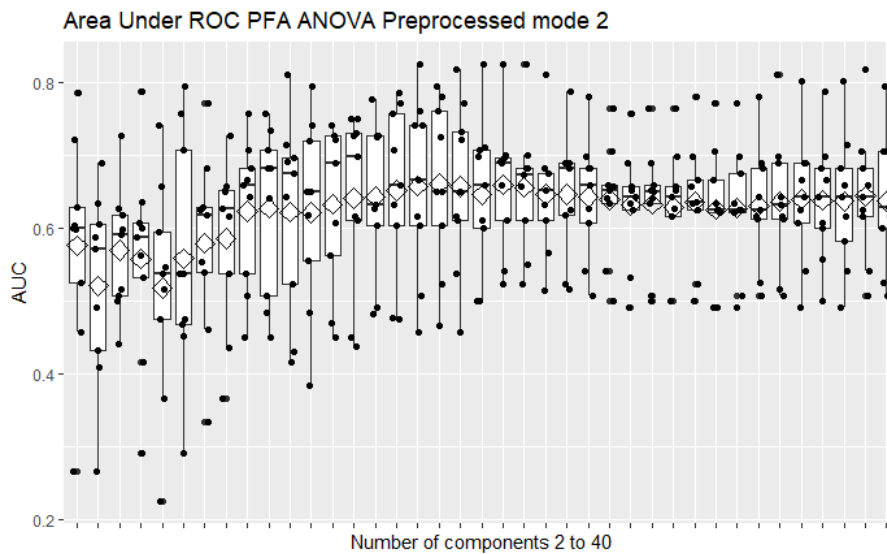


Figura 80. SVM PFA ANOVA escalado modo 2.

A42. SVM PCA ESCALADO MODO 3

Con 21 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.667, con todos los grupos superan 0.5.

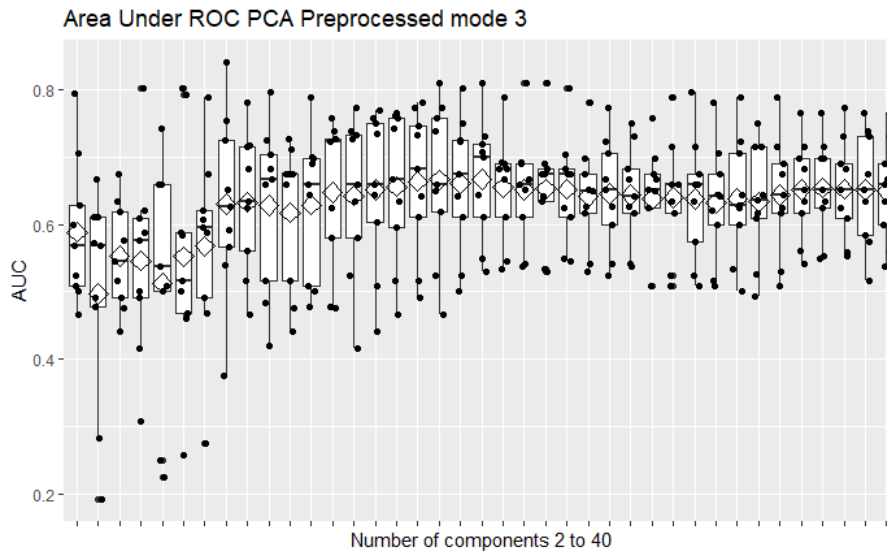


Figura 81. SVM PCA escalado modo 3.

A43. SVM PFA POL. 2ºG ESCALADO MODO 3

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.587, con dos grupos en torno a 0.2.

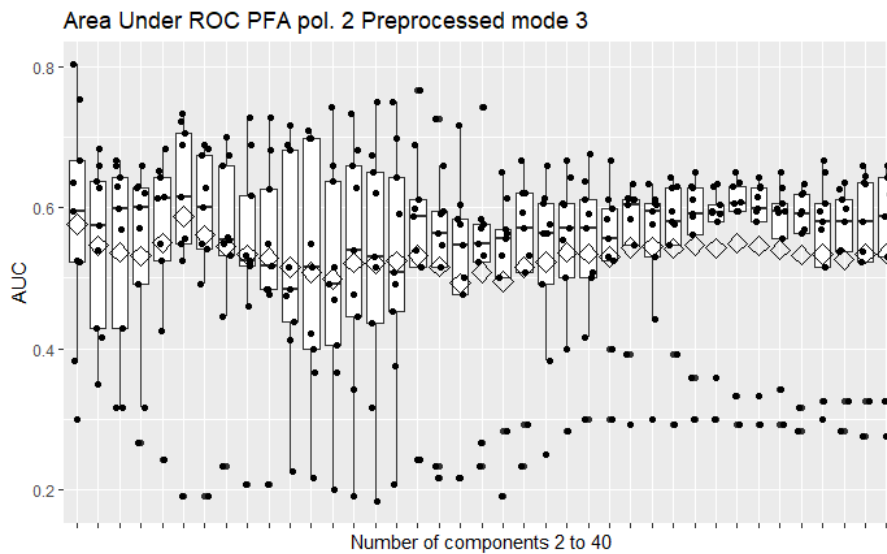


Figura 82. SVM PFA pol. 2 escalado modo 3.

A44. SVM PFA POL. 3ºG ESCALADO MODO 3

Con 10 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.578, con un grupo en torno a 0.4.

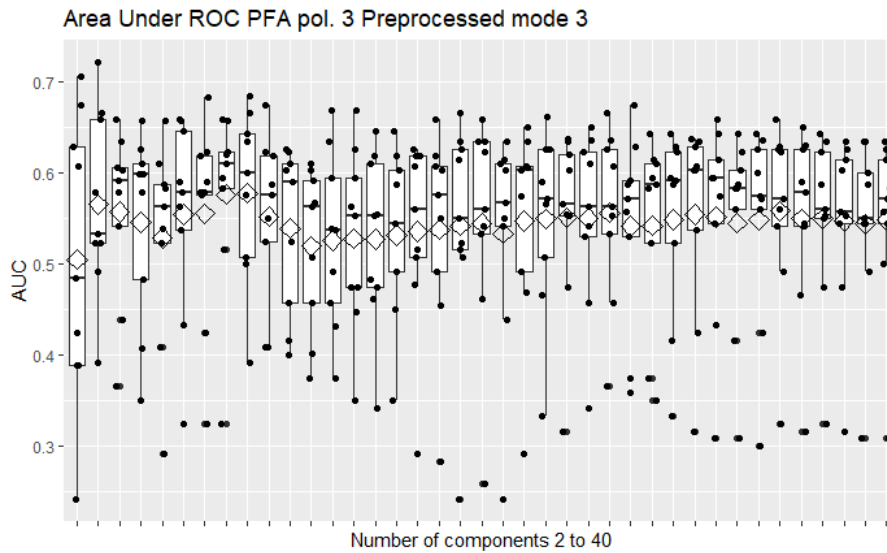


Figura 83. SVM PFA pol. 3 escalado modo 3.

A45. SVM PFA POL. 4ºG ESCALADO MODO 3

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.589, con dos grupos en torno a 0.3.

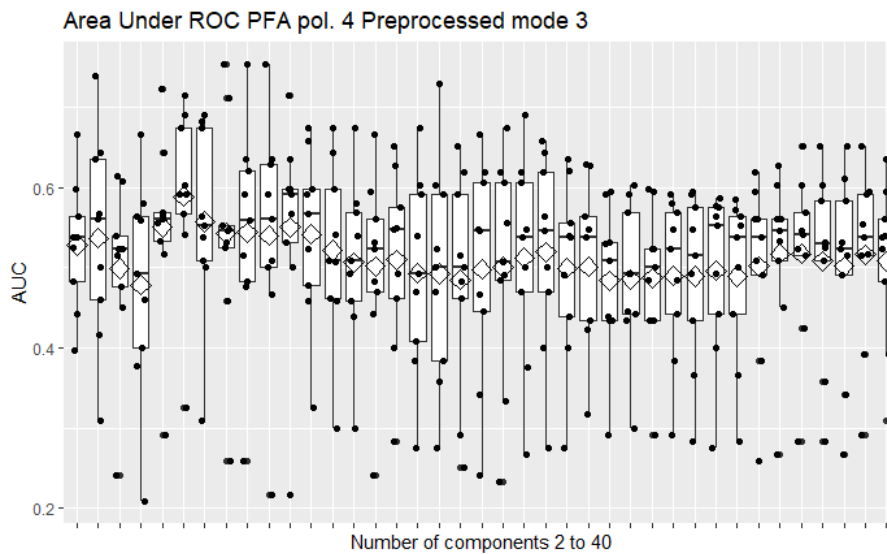


Figura 84. SVM PFA pol. 4 escalado modo 3.

A46. SVM PFA ANOVA ESCALADO MODO 3

Con 32 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.658, con todos los grupos por encima de 0.5.

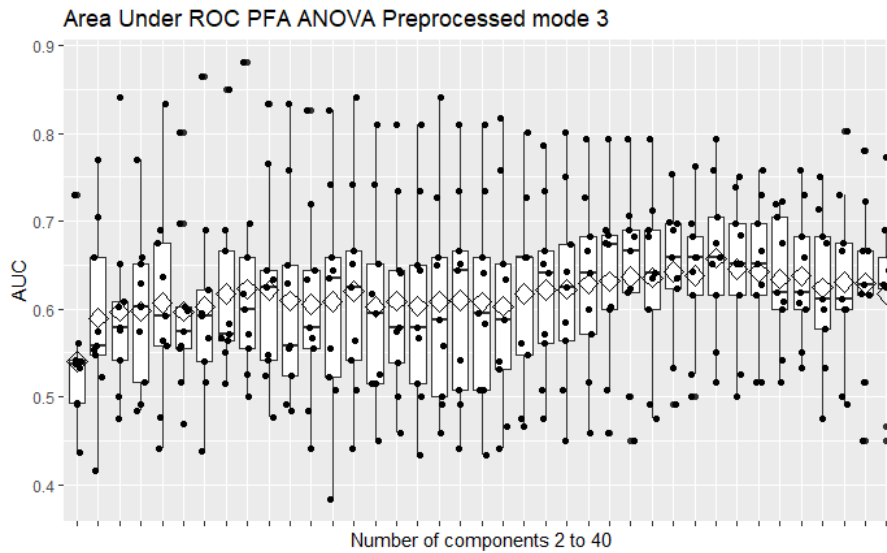


Figura 85. SVM PFA ANOVA escalado modo 3.

A47. SVM PCA ESCALADO MODOS 2 Y 3

Con 25 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.659, con todos los grupos por encima de 0.5.

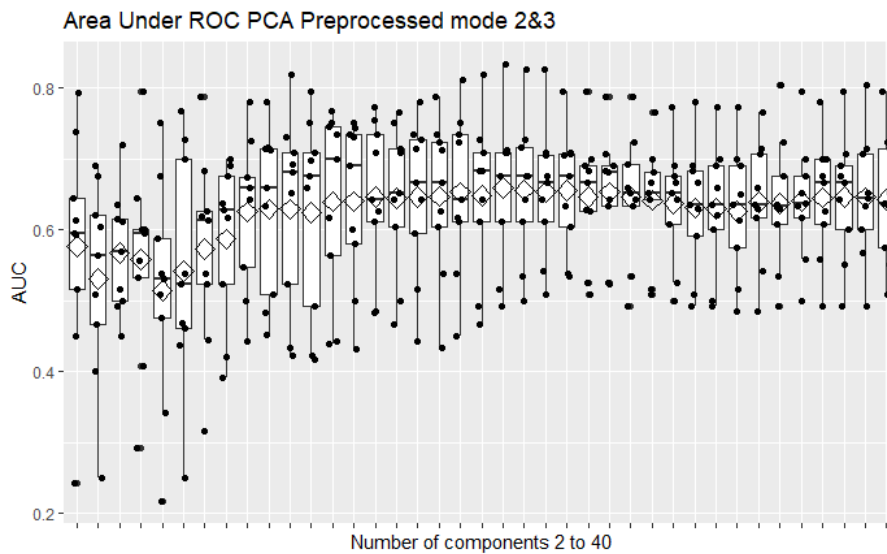


Figura 86. SVM PCA escalado modos 2 y 3.

A48. SVM PFA POL. 2ºG ESCALADO MODOS 2 Y 3

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.563, con un grupo en torno a 0.3 y otro en torno a 0.4.

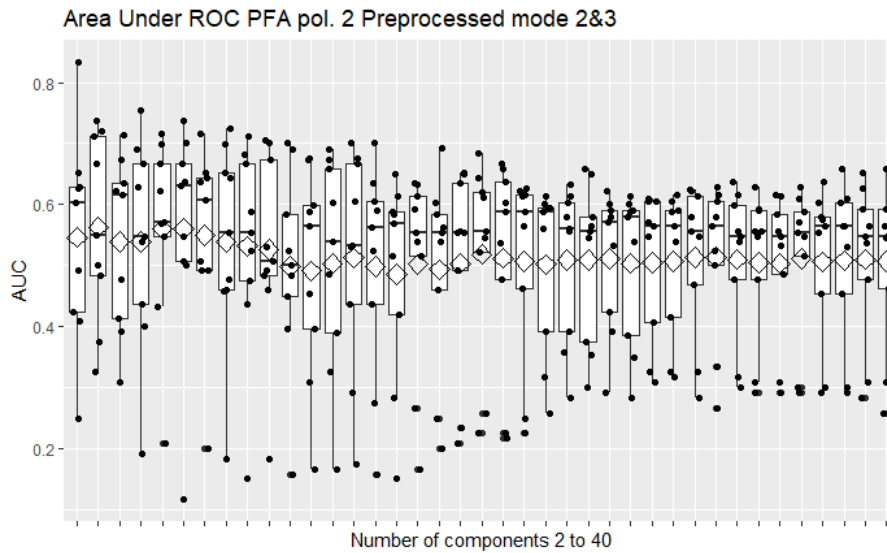


Figura 87. SVM PFA pol. 2 escalado modos 2 y 3.

A49. SVM PFA POL. 3ºG ESCALADO MODOS 2 Y 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.592, con dos grupos en torno a 0.3.

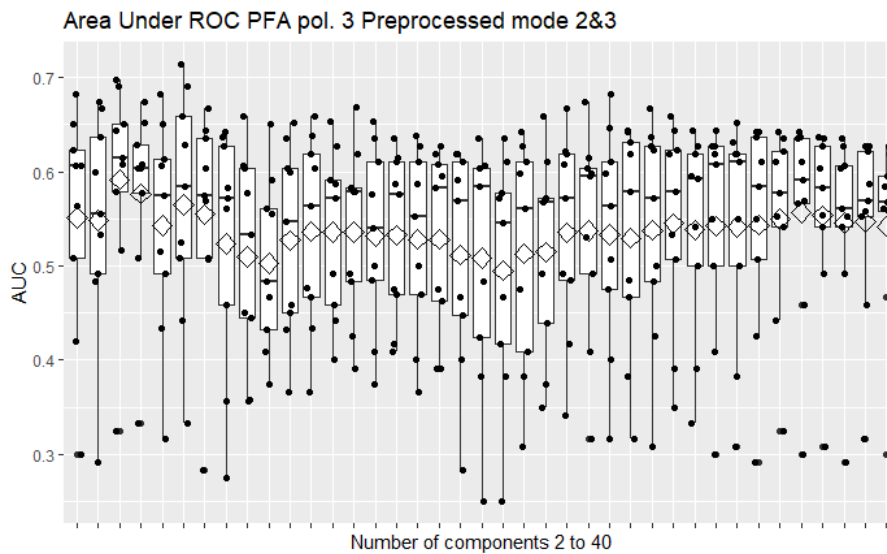


Figura 88. SVM PFA pol. 3 escalado modos 2 y 3.

A50. SVM PFA POL. 4ºG ESCALADO MODOS 2 Y 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.595, con dos grupos en torno a 0.45.

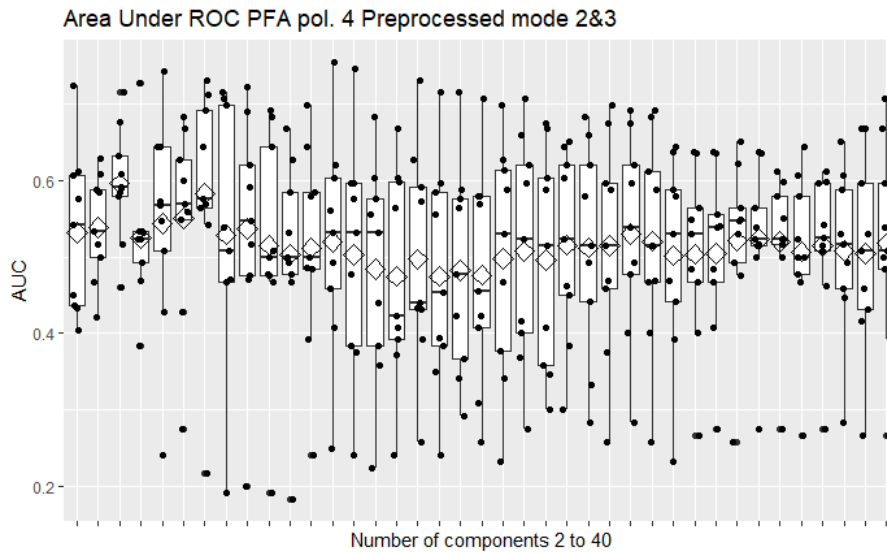


Figura 89. SVM PFA pol. 4 escalado modos 2 y 3.

A51. SVM PFA ANOVA ESCALADO MODOS 2 Y 3

Con 33 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.646, con todos los grupos por encima de 0.5.

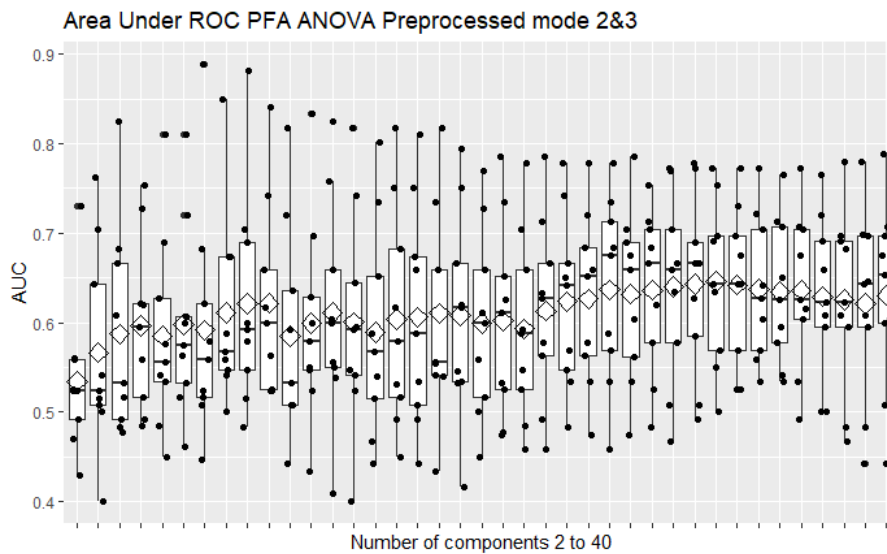


Figura 90. SVM PFA ANOVA escalado modos 2 y 3.

A52. SNPLSDA AUTOESCALADO

Con 50 puntos de corriente y 2 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.654, con algunos grupos por debajo de 0.5.

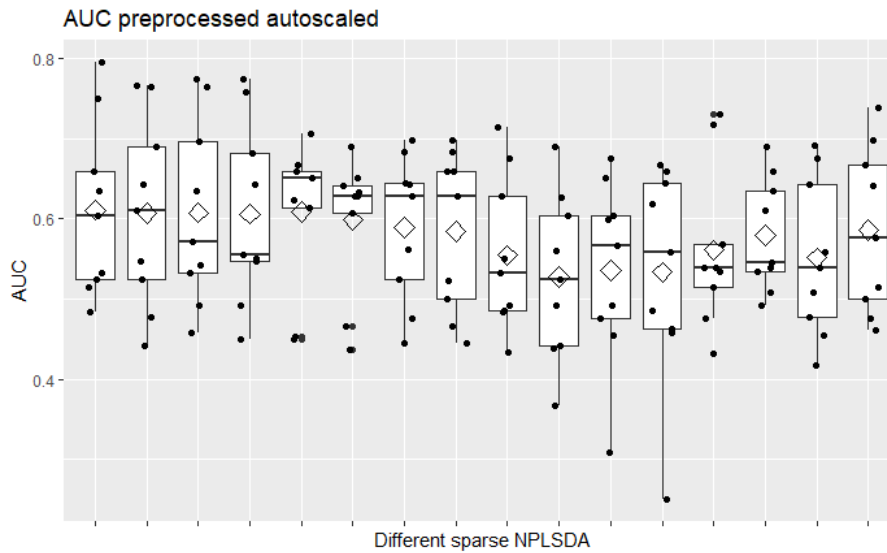


Figura 91. sNPLSDA 50 puntos de corriente y dos sensores, autoescalado.

A53. SNPLSDA CENTRADO

Con 50 puntos de corriente y 5 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.611, con todos los grupos por encima de 0.5.

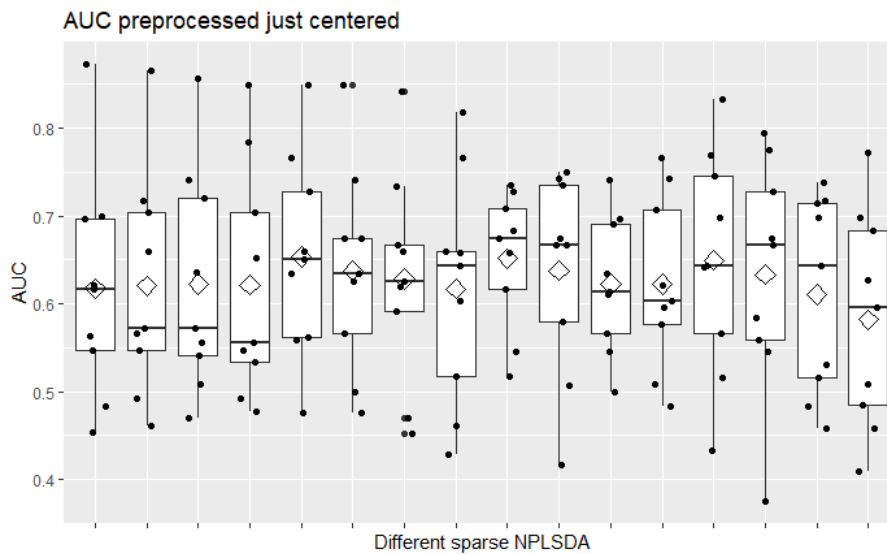


Figura 92. sNPLSDA 50 puntos de corriente y cinco sensores, centrado.

A54. SNPLSDA ESCALADO EN EL MODO 2

Con 50 puntos de corriente y 3 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.633, con todos los grupos por encima de 0.5.

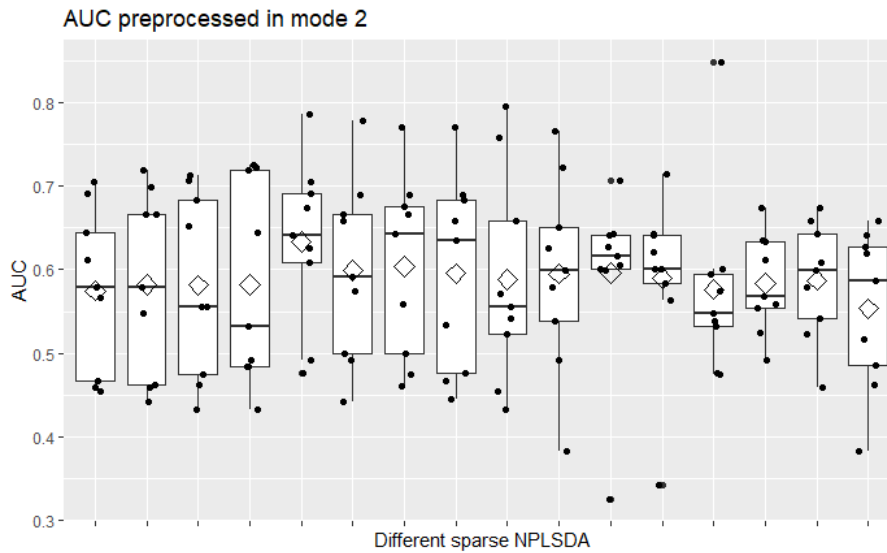


Figura 93. sNPLSDA 50 puntos de corriente y tres sensores, escalado modo 2.

A55. SNPLSDA ESCALADO EN EL MODO 3

Con 50 puntos de corriente y 3 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.677, con todos los grupos por encima de 0.5.



Figura 94. sNPLSDA 50 puntos de corriente y tres sensores, escalado modo 3.

A56. SNPLSDA ESCALADO EN LOS MODOS 2 Y 3

Con 50 puntos de corriente y 3 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.645, con todos los grupos por encima de 0.5.

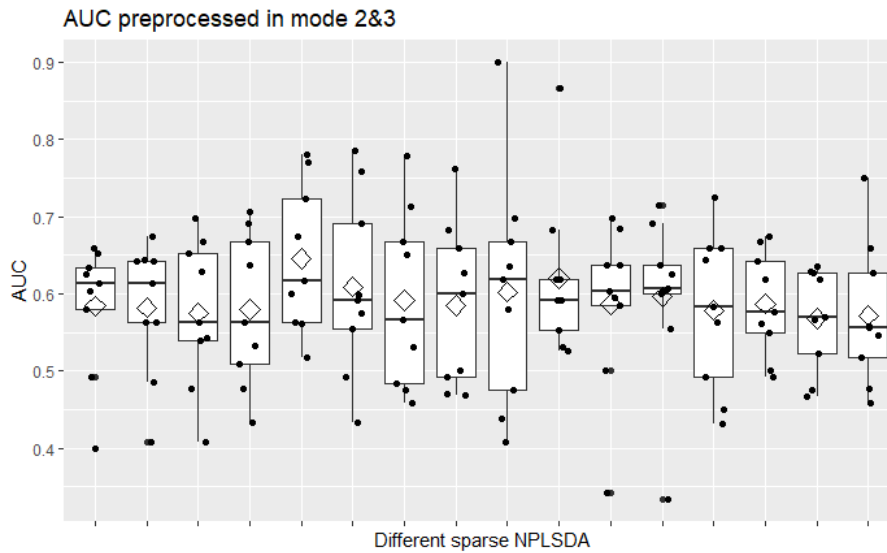


Figura 95. sNPLSDA 50 puntos de corriente y tres sensores, escalado modos 2 y 3.

A57. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de: 0.597, con todos los grupos por encima de 0.5.

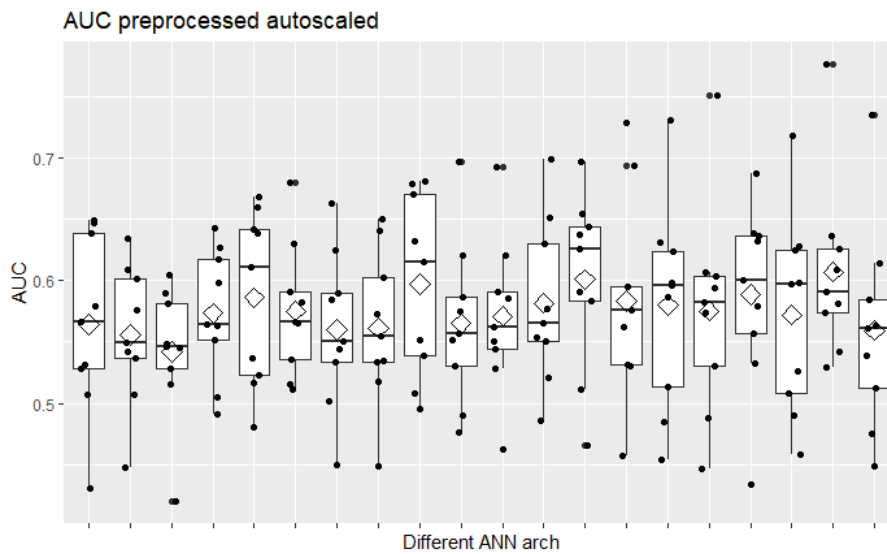


Figura 96. AUC de ANN de una capa, sin reducción de la dimensión, autoescalado.

A58. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.622, con un grupo por debajo de 0.5.

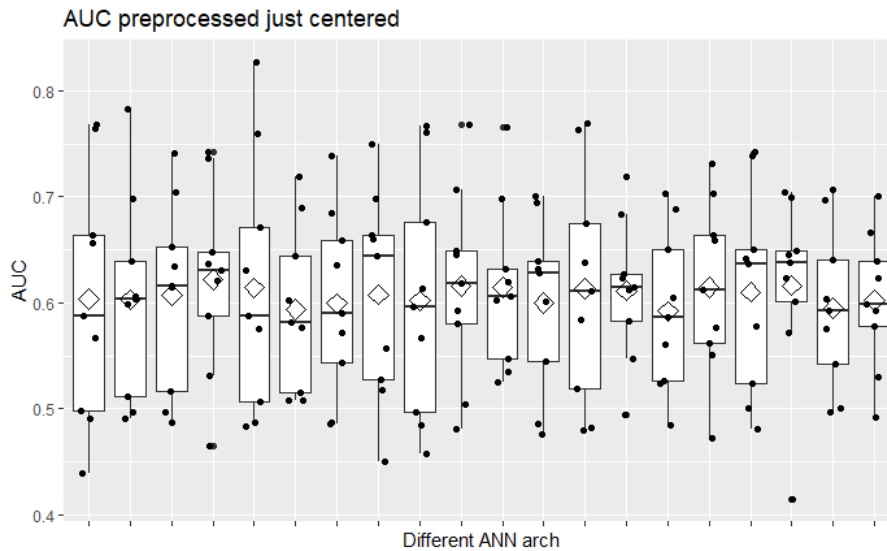


Figura 97. AUC de ANN de una capa, sin reducción de la dimensión, centrado.

A59. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO EN EL MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.626, con todos los grupos por encima de 0.5.

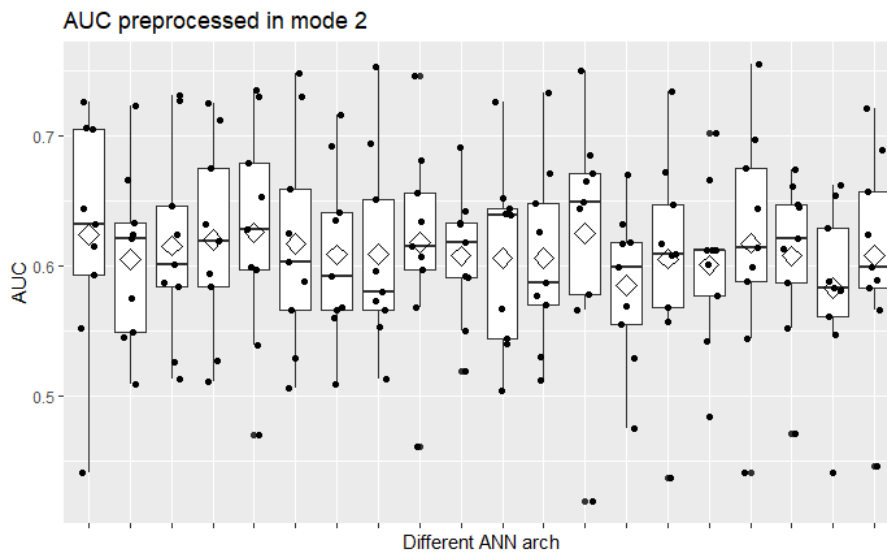


Figura 98. AUC de ANN de una capa, sin reducción de la dimensión, escalado modo 2.

A60. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO EN EL MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.631, con todos los grupos por encima de 0.5.

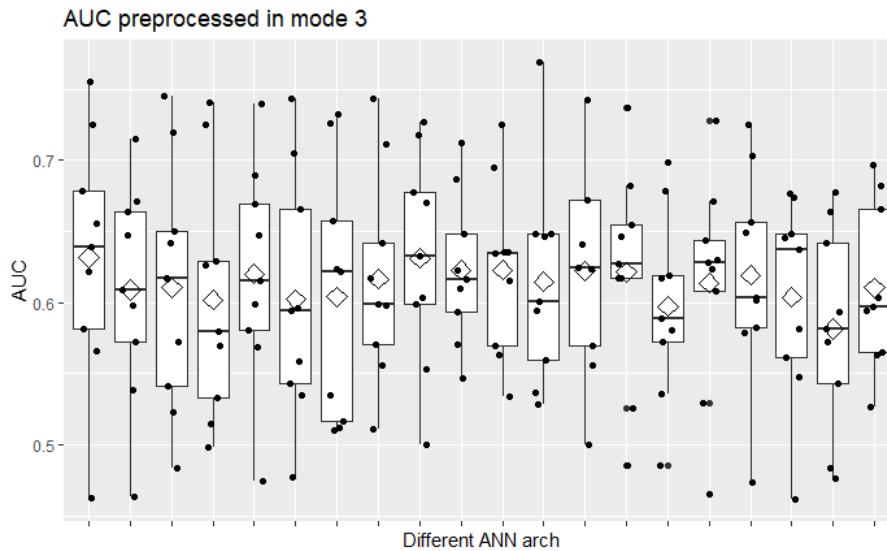


Figura 99. AUC de ANN de una capa, sin reducción de la dimensión, escalado modo 3.

A61. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO EN LOS MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.632, con un grupo por debajo de 0.5.

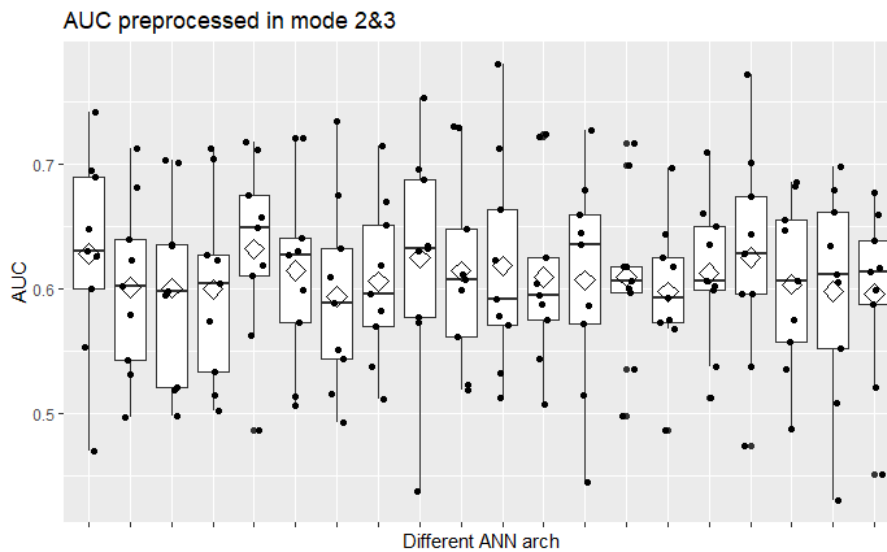


Figura 100. AUC de ANN de una capa, sin reducción de la dimensión, escalado modos 2&3.

A62. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.607, con un grupo por debajo de 0.5.

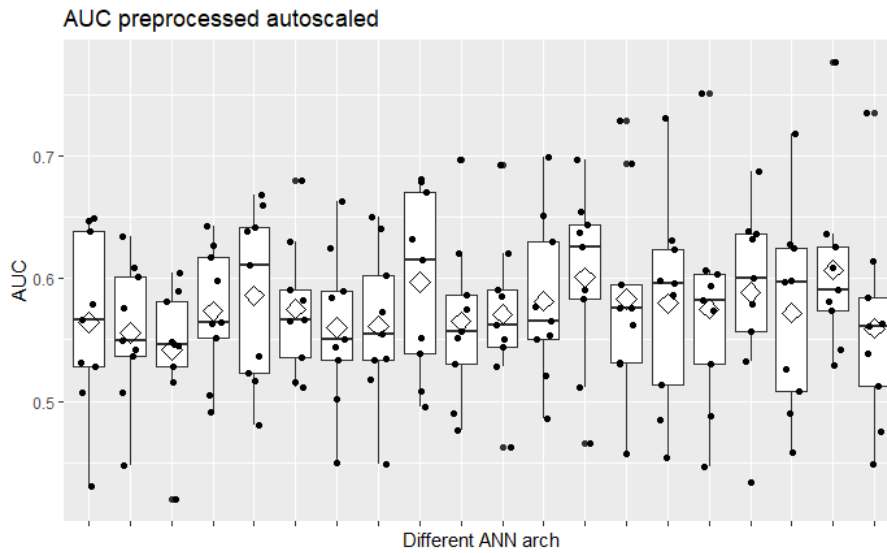


Figura 101. AUC de ANN de dos capas, sin reducción de la dimensión, autoescalado.

A63. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.633, con un grupo por debajo de 0.5.

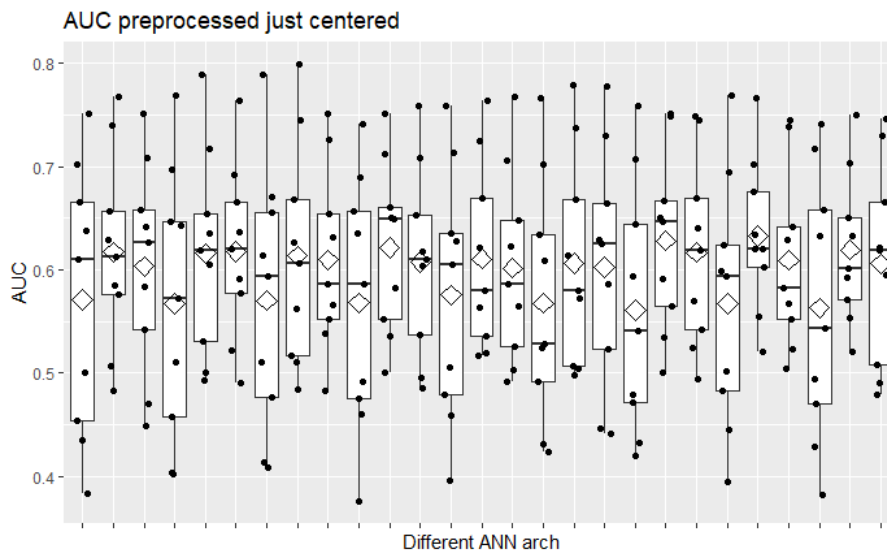


Figura 102. AUC de ANN de dos capas, sin reducción de la dimensión, centrado.

A64. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.620, con ningún grupo por debajo de 0.5.



Figura 103. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modo 2.

A65. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.631, con un grupo por debajo de 0.5.

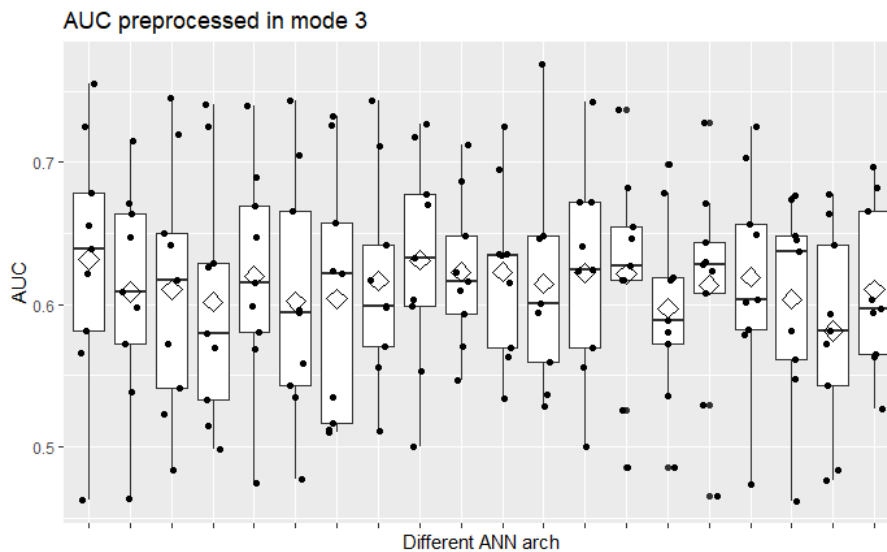


Figura 104. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modo 3.

A66. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.610, con un grupo por debajo de 0.5.

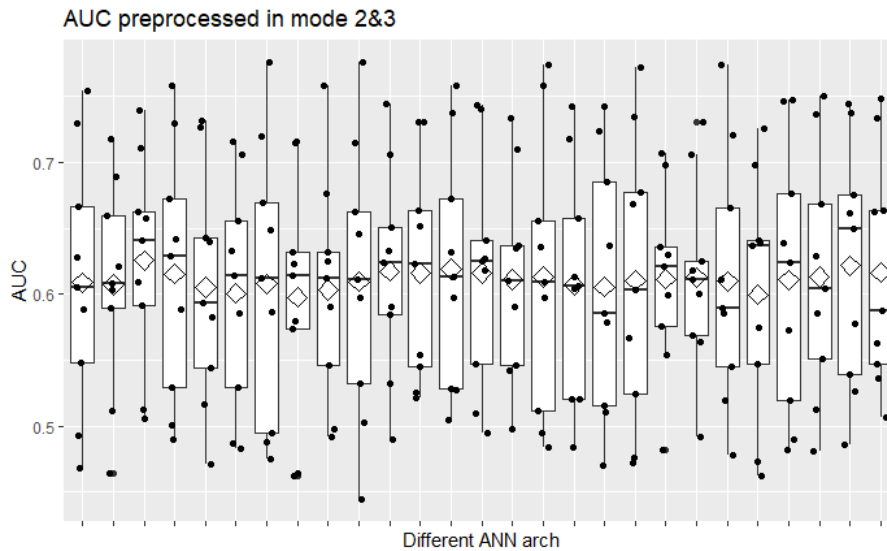


Figura 105. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modos 2&3.

A67. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.632, con dos grupos por debajo de 0.5.

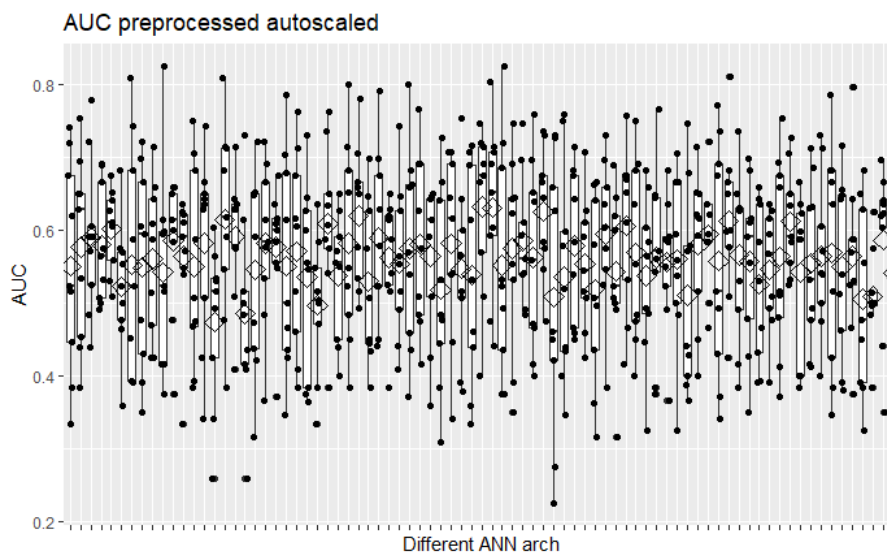


Figura 106. AUC de ANN de tres capas, sin reducción de la dimensión, autoescalado.

A68. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.678, con ningún grupo por debajo de 0.5.

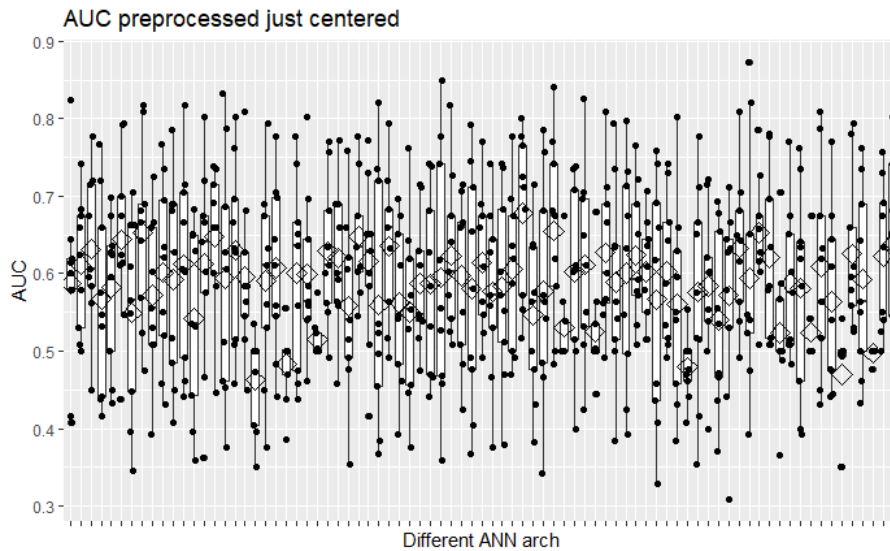


Figura 107. AUC de ANN de tres capas, sin reducción de la dimensión, centrado.

A69. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.654, con dos grupos por debajo de 0.5.

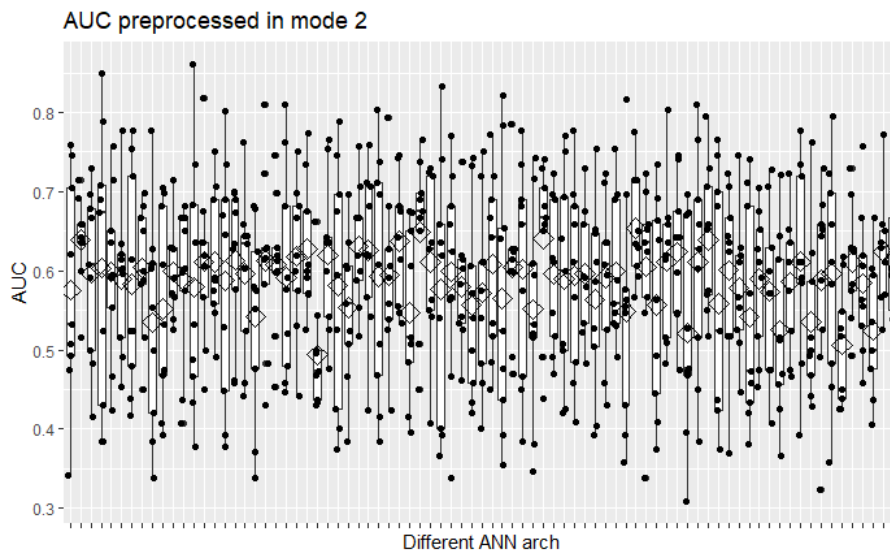


Figura 108. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modo 2.

A70. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.670, con un grupo por debajo de 0.5.

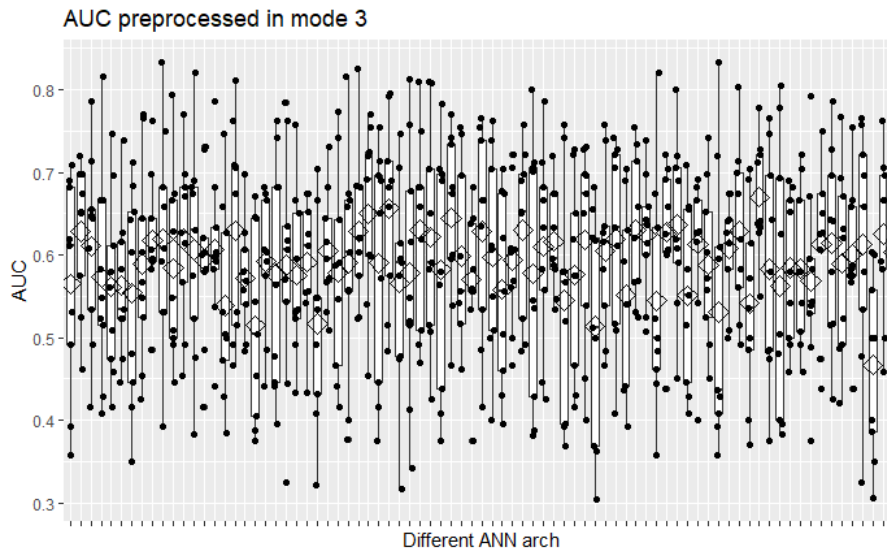


Figura 109. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modo 3.

A71. ANN SIN REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, con ningún grupo por debajo de 0.5.

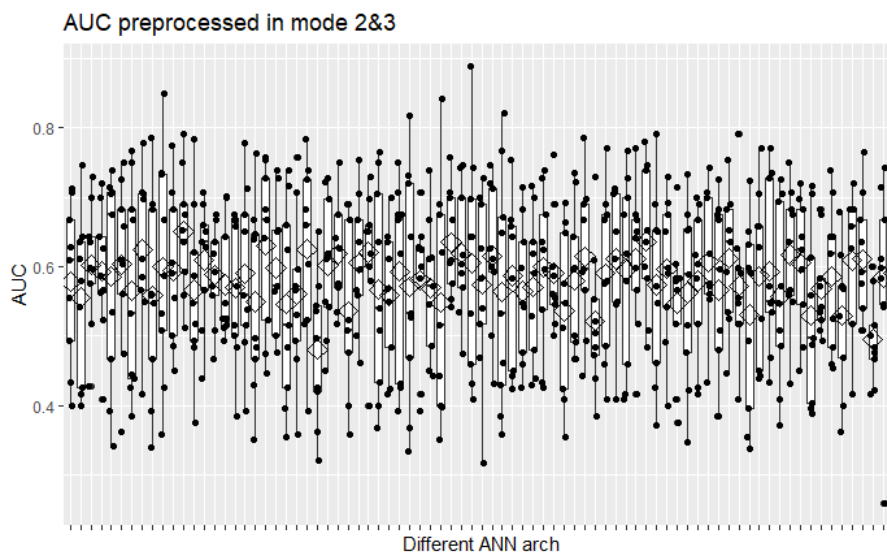


Figura 110. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modos 2&3.

A72. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.587, con algunos grupos por debajo de 0.5.

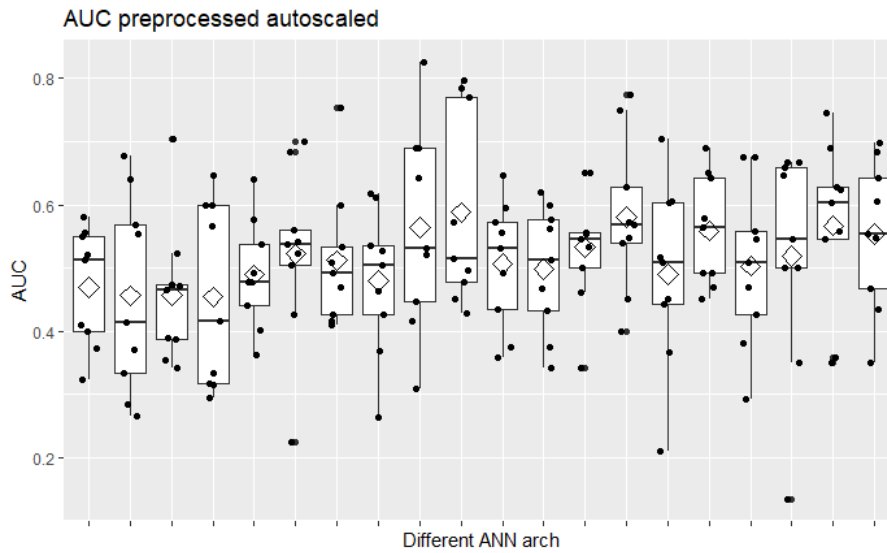


Figura 111. AUC de ANN de una capa, con reducción de la dimensión, autoescalado.

A73. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.650, con ningún grupo por debajo de 0.5.

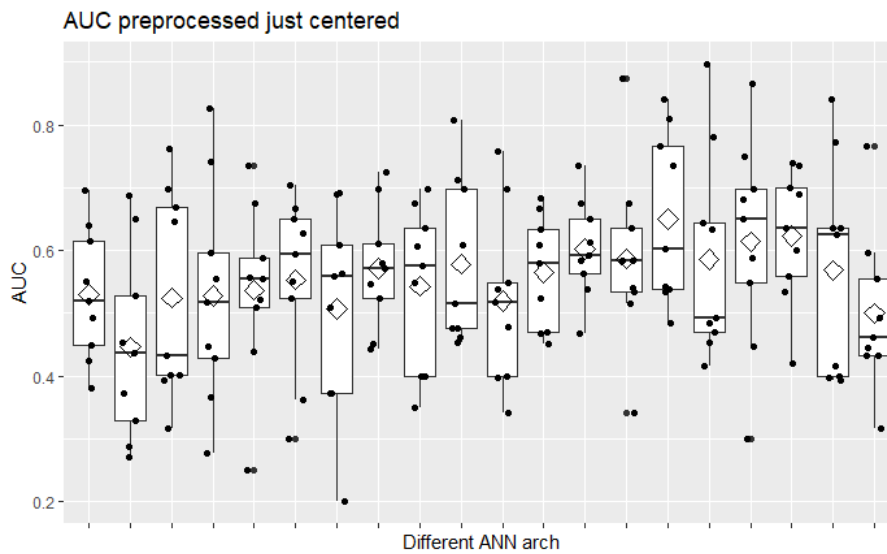


Figura 112. AUC de ANN de una capa, con reducción de la dimensión, centrado.

A74. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.628, con un grupo por debajo de 0.5.



Figura 113. AUC de ANN de una capa, con reducción de la dimensión, escalado modo 2.

A75. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.586, con un grupo por debajo de 0.5.

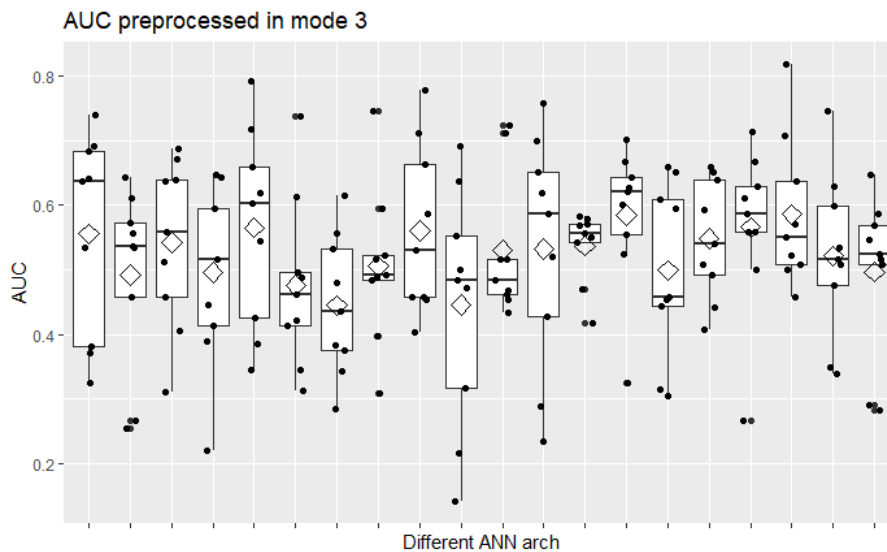


Figura 114. AUC de ANN de una capa, con reducción de la dimensión, escalado modo 3.

A76. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON UNA CAPA ESCALADO MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.583, con dos grupos por debajo de 0.5.

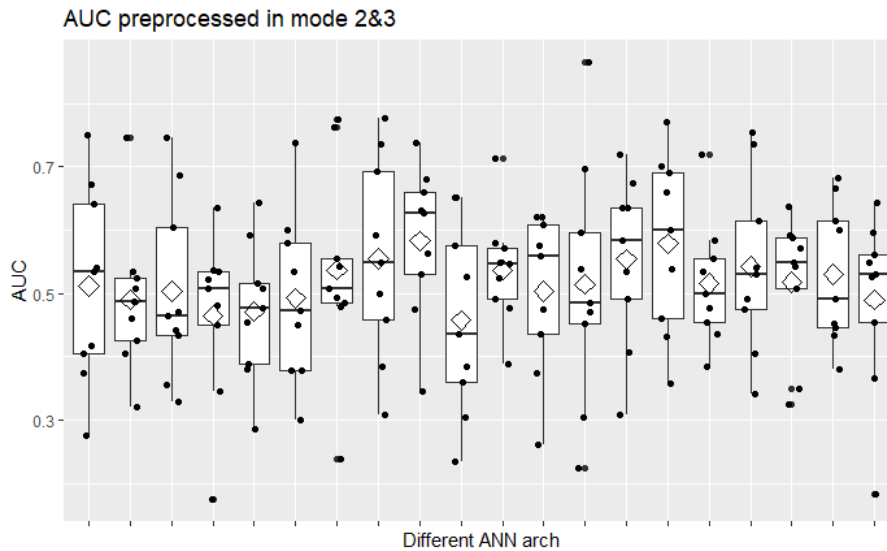


Figura 115. AUC de ANN de una capa, con reducción de la dimensión, escalado modos 2&3.

A77. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.640, con un grupo por debajo de 0.5.

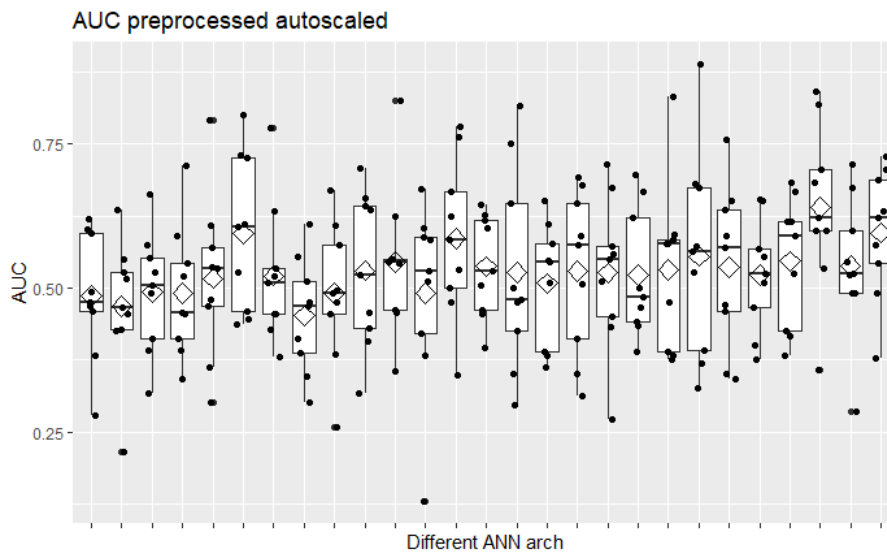


Figura 116. AUC de ANN de dos capas, con reducción de la dimensión, autoescalado.

A78. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.665, con un grupo en torno a 0.5.

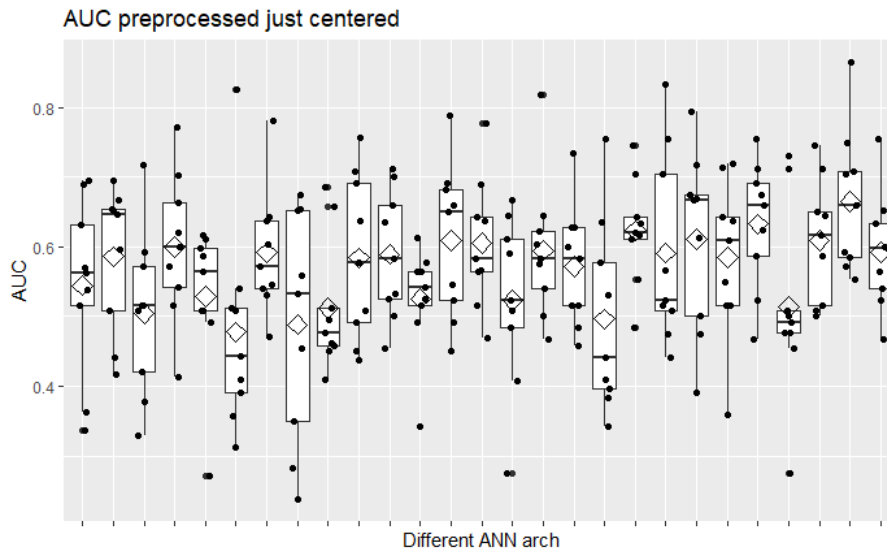


Figura 117. AUC de ANN de dos capas, con reducción de la dimensión, centrado.

A79. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.591, con ningún grupo por debajo de 0.5.

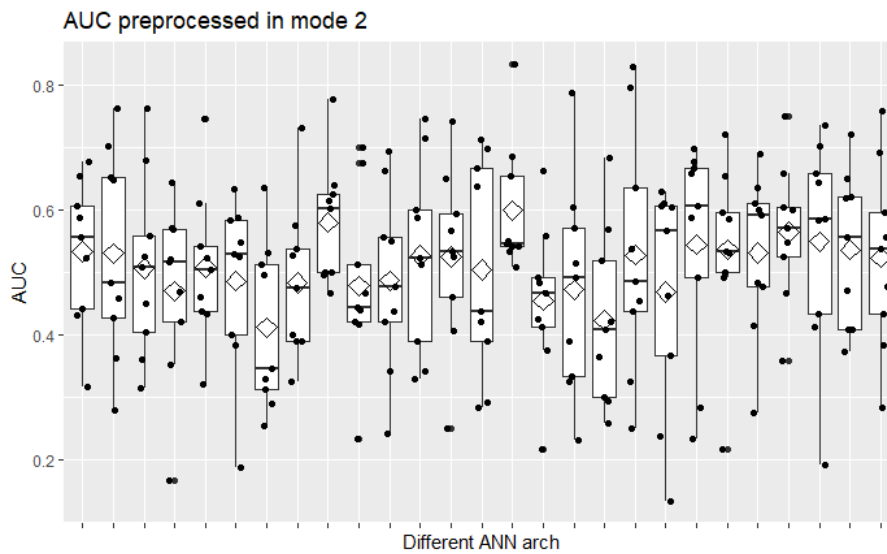


Figura 118. AUC de ANN de dos capas, con reducción de la dimensión, escalado modo 2.

A80. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.585, con algunos grupos por debajo de 0.5.

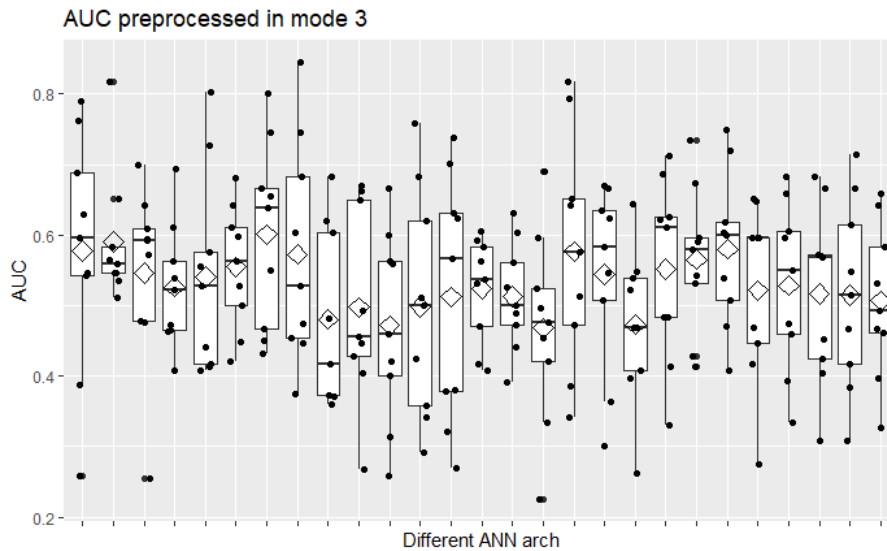


Figura 119. AUC de ANN de dos capas, con reducción de la dimensión, escalado modo 3.

A81. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON DOS CAPAS ESCALADO MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.587, con algunos grupos por debajo de 0.5.

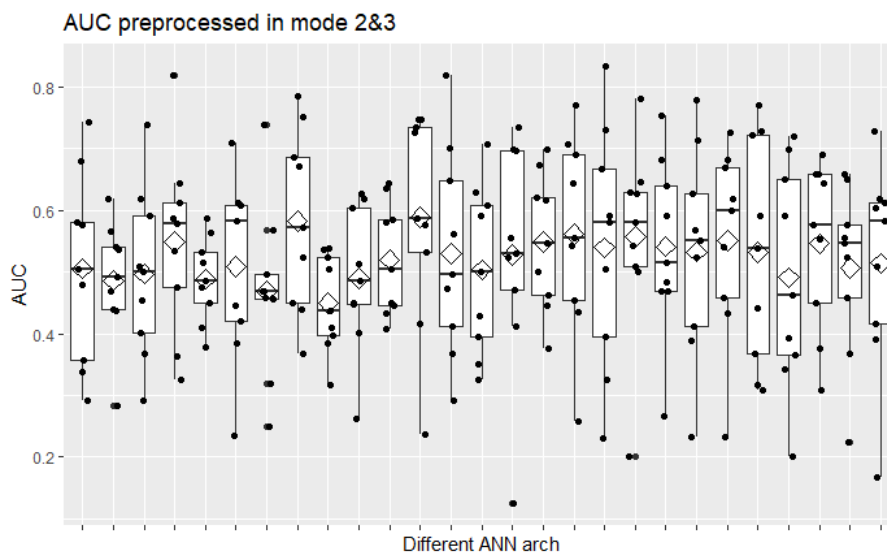


Figura 120. AUC de ANN de dos capas, con reducción de la dimensión, escalado modos 2 y 3.

A82. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.645, con un grupo por debajo de 0.5.

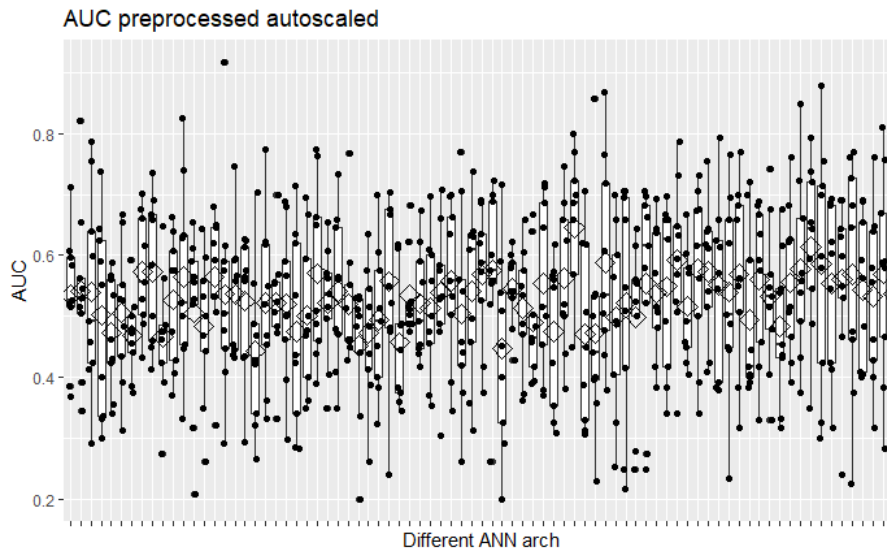


Figura 121. AUC de ANN de tres capas, con reducción de la dimensión, autoescalado.

A83. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.693, con ningún grupo por debajo de 0.5.

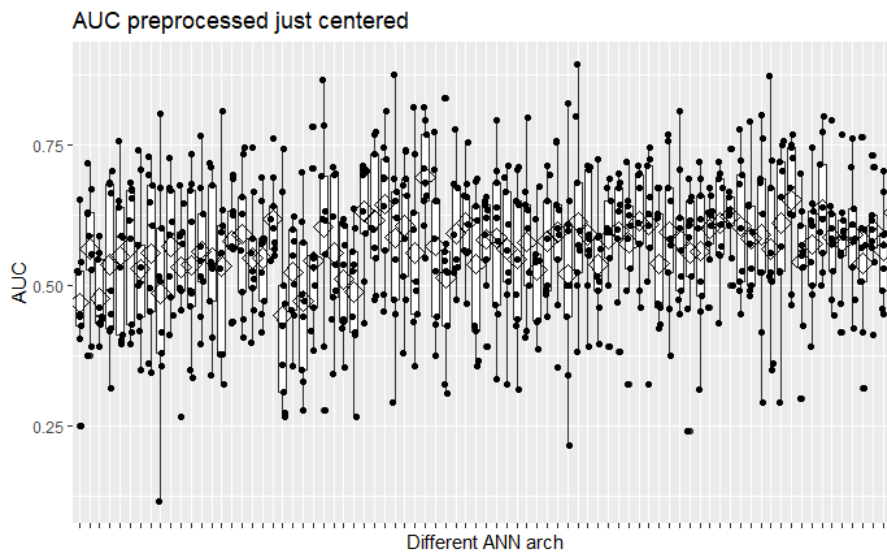


Figura 122. AUC de ANN de tres capas, con reducción de la dimensión, centrado.

A84. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.614, con algunos grupos por debajo de 0.5.



Figura 123. AUC de ANN de tres capas, con reducción de la dimensión, escalado modo 2.

A85. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de: 0.647, con un grupo por debajo de 0.5.

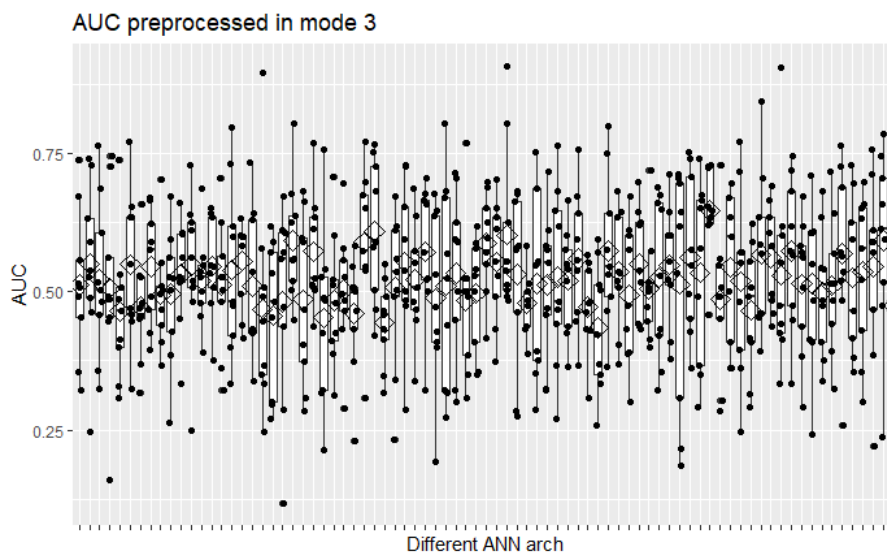


Figura 124. AUC de ANN de tres capas, con reducción de la dimensión, escalado modo 3.

A86. ANN CON REDUCCIÓN DE LA DIMENSIÓN CON TRES CAPAS ESCALADO MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de: 0.630, con ningún grupo por debajo de 0.5.

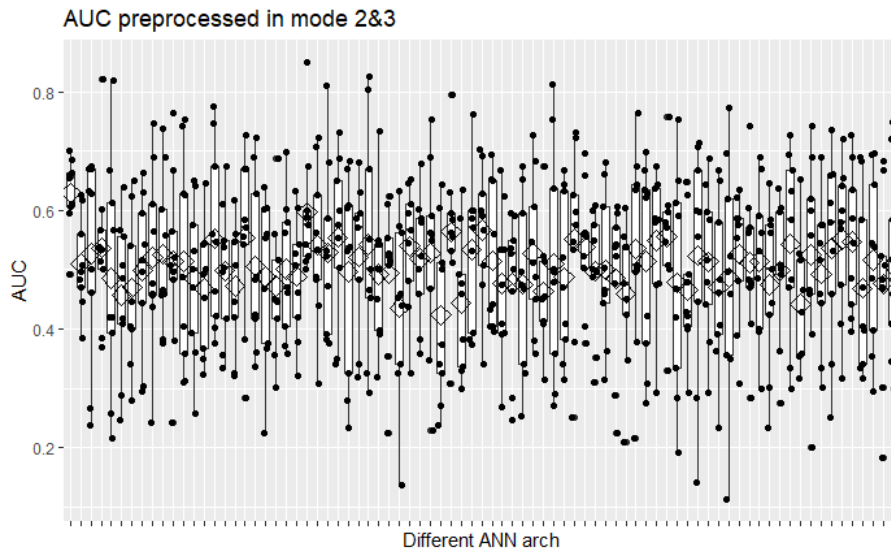


Figura 125. AUC de ANN de tres capas, con reducción de la dimensión, escalado modos 2&3.

ANEXO: CON VARIABLES ADICIONALES

Los anexos hacen referencia a la selección de número de componentes principales, factores, u otros parámetros de los modelos. Para ello, se atiende al AUC medio mediante validación cruzada (CV) a nueve grupos balanceados, comentándose adicionalmente si alguno de ellos, puntualmente, se encuentra por debajo de 0.5 (es decir, arroja peores resultados que un clasificador aleatorio).

A87. PCR AUTOESCALADO

Con 33 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.638, con uno de ellos en torno a 0.3.

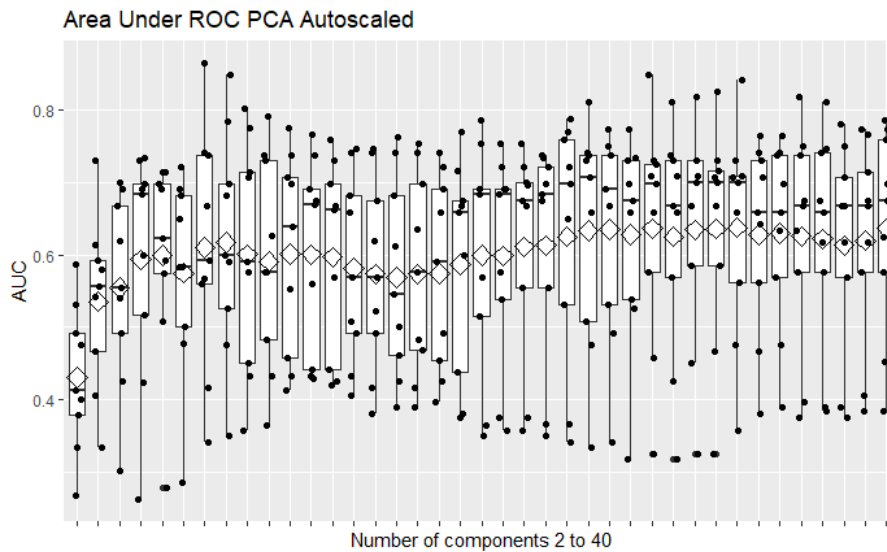


Figura 126. PCR Autoescalado.

A88. PFR POLINOMIAL DE ORDEN 2 AUTOESCALADO

Con 7 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.598, dos de ellos en torno a 0.45 y otro en torno a 0.3.

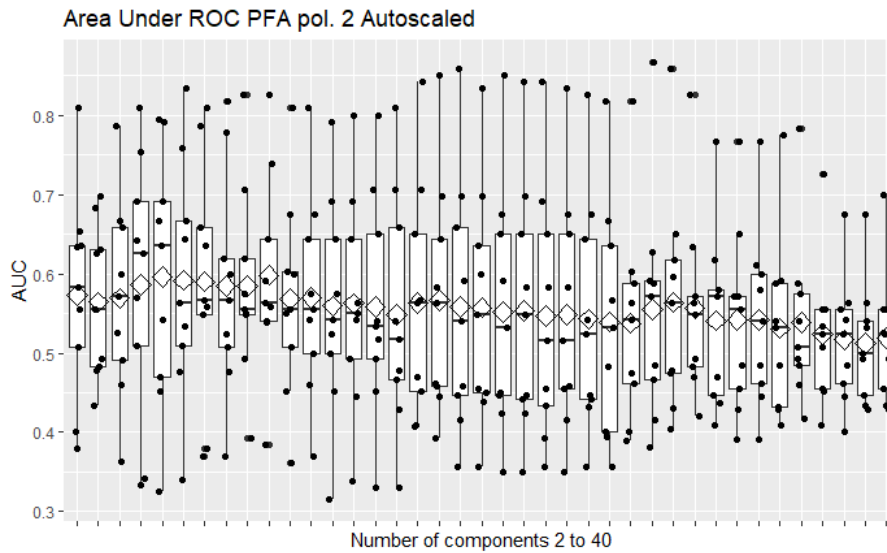


Figura 127. PFR pol. 2 autoescalado.

A89. PFR POLINOMIAL DE ORDEN 3 AUTOESCALADO

Con 6 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.598, dos de ellos en torno a 0.4.

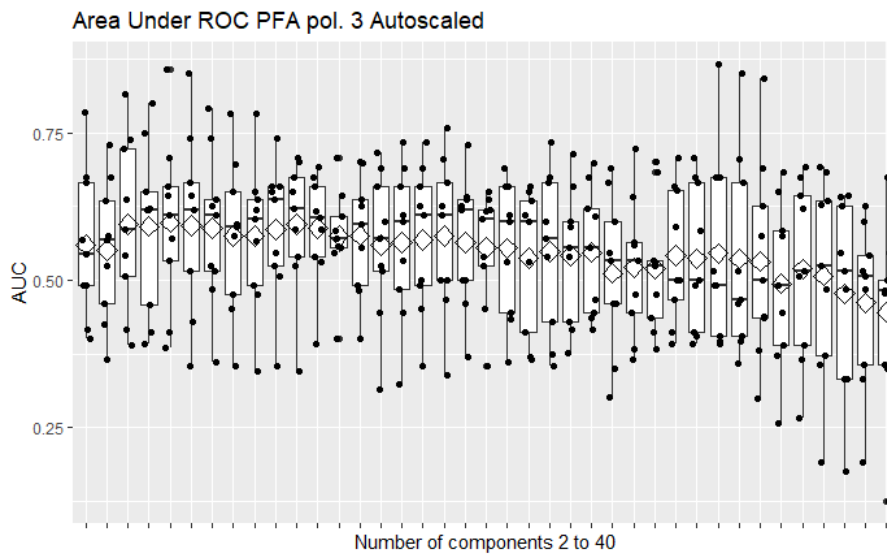


Figura 128. PFR pol. 3 autoescalado.

A90. PFR POLINOMIAL DE ORDEN 4 AUTOESCALADO

Con 5 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.581, uno de ellos en torno a 0.3, otro en torno a 0.4 y otro en torno a 0.45.

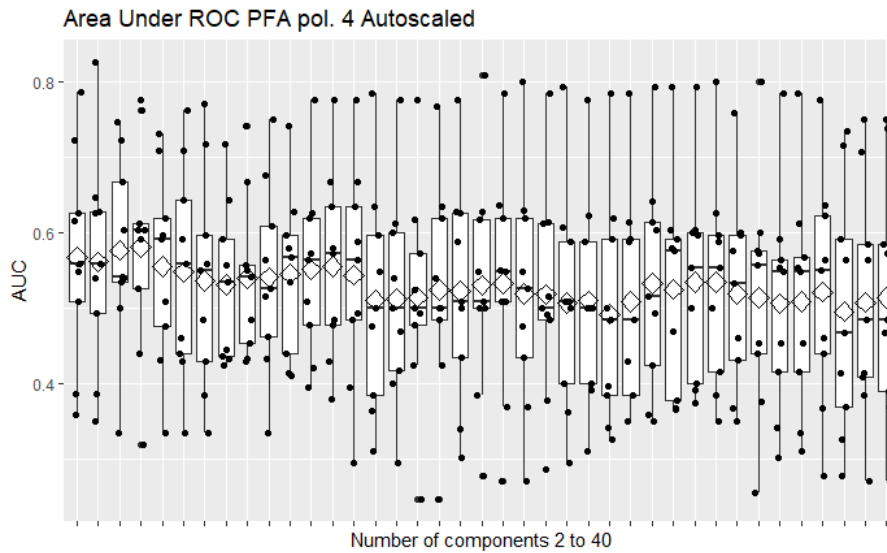


Figura 129. PFR pol. 4 autoescalado.

A91. PFR ANOVA AUTOESCALADO

Con 8 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.632, con dos de ellos en torno a 0.45.

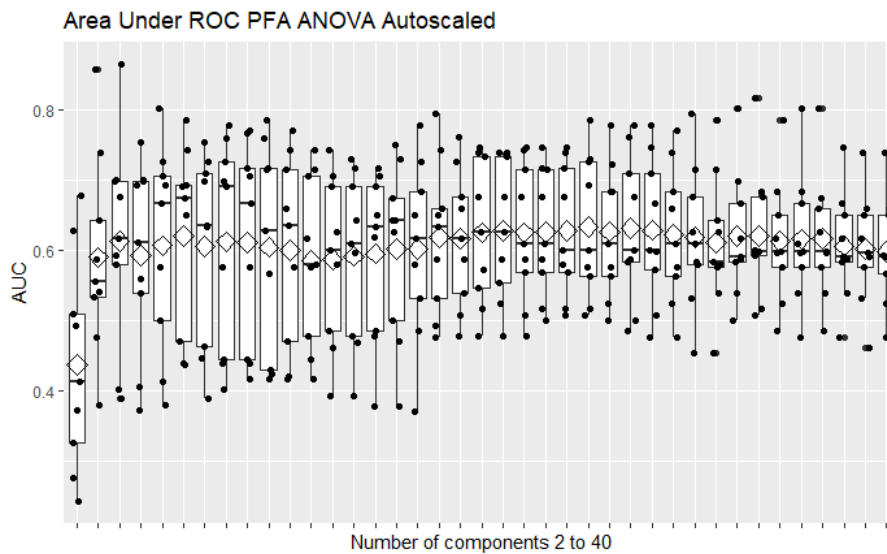


Figura 130. PFR ANOVA autoescalado.

A92. PCR SÓLO CENTRADO

Con 33 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.655, con un grupo por debajo de 0.4.

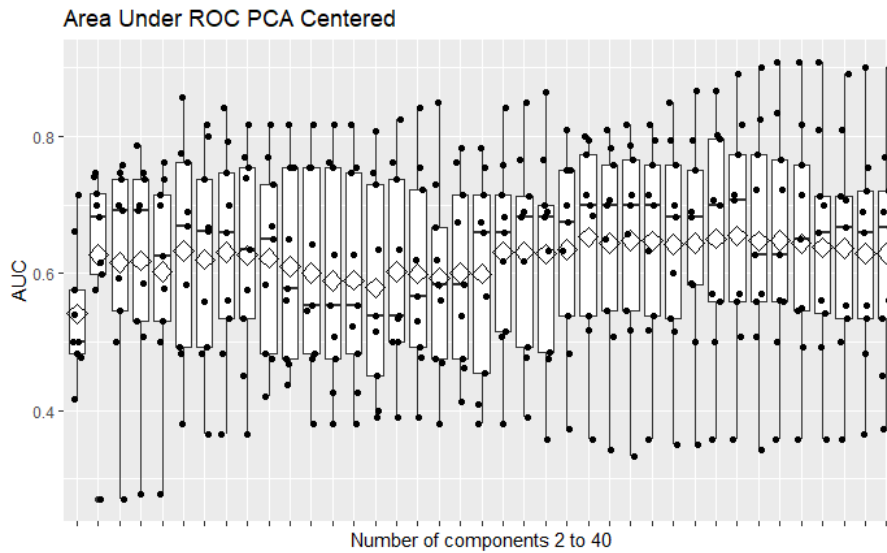


Figura 131. PCR centrado.

A93. PFR POLINOMIAL DE ORDEN 2 SÓLO CENTRADO

Con 3 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.570, con un grupo alrededor de 0.45 y otros dos sobre 0.4.

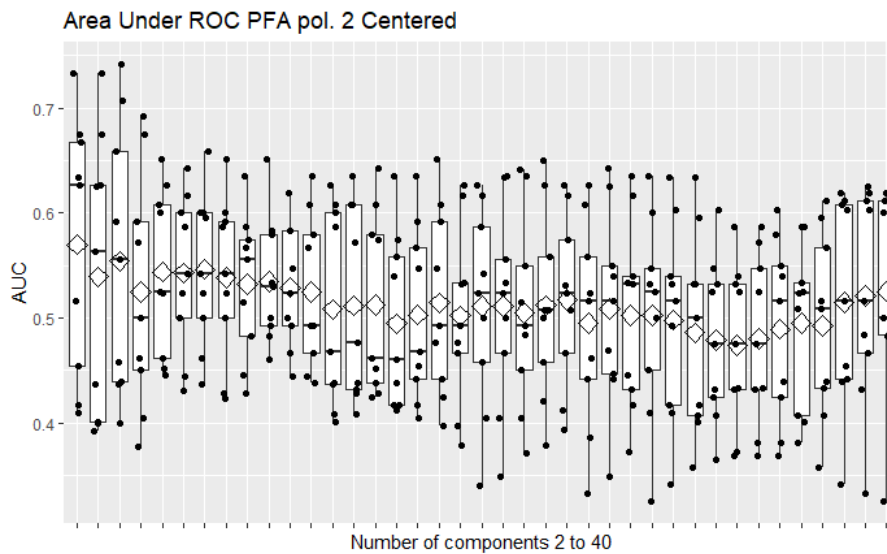


Figura 132. PFR pol. 2 centrado.

A94. PFR POLINOMIAL DE ORDEN 3 SÓLO CENTRADO

Con 37 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.604, tres de ellos en torno a 0.45.

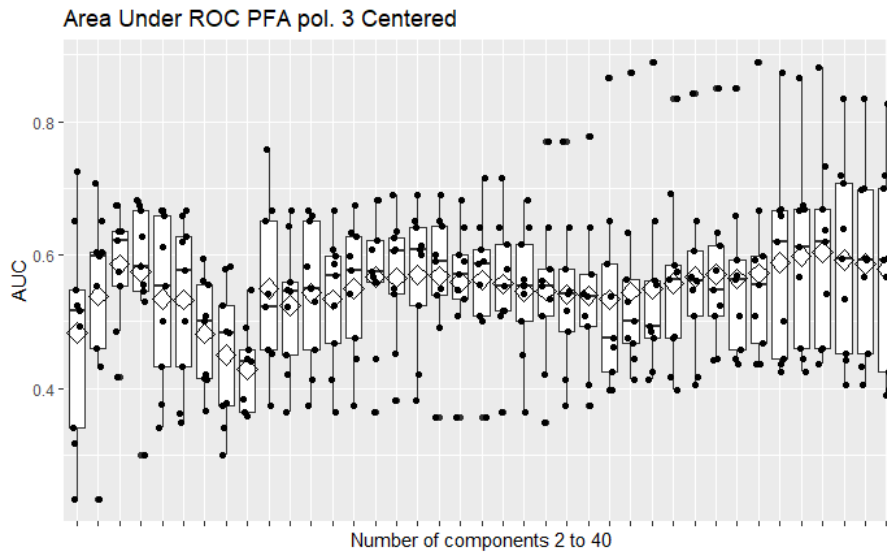


Figura 133. PFR pol. 3 centrado.

A95. PFR POLINOMIAL DE ORDEN 4 SÓLO CENTRADO

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.595, uno de ellos en torno a 0.3 y otro en torno a 0.9.

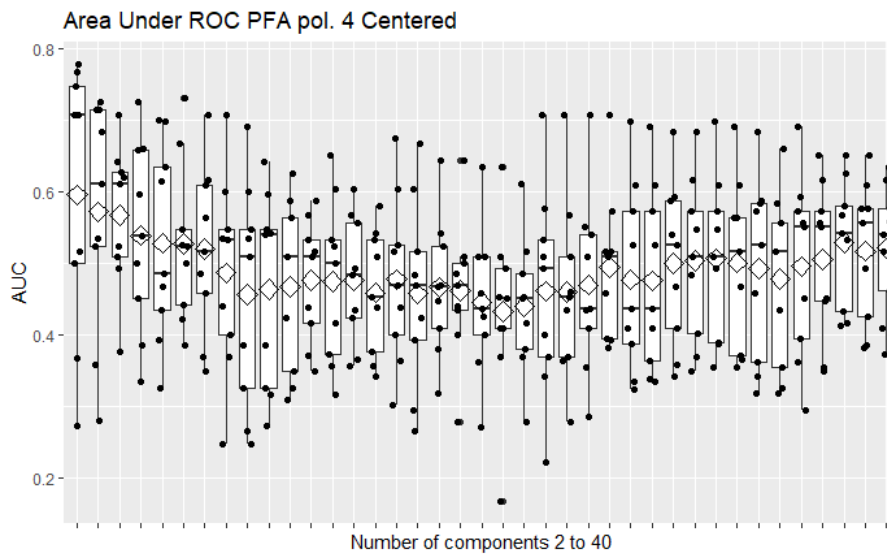


Figura 134. PFR pol. 4 centrado.

A96. PFR ANOVA SÓLO CENTRADO

Con 32 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, con un grupo en torno a 0.4.

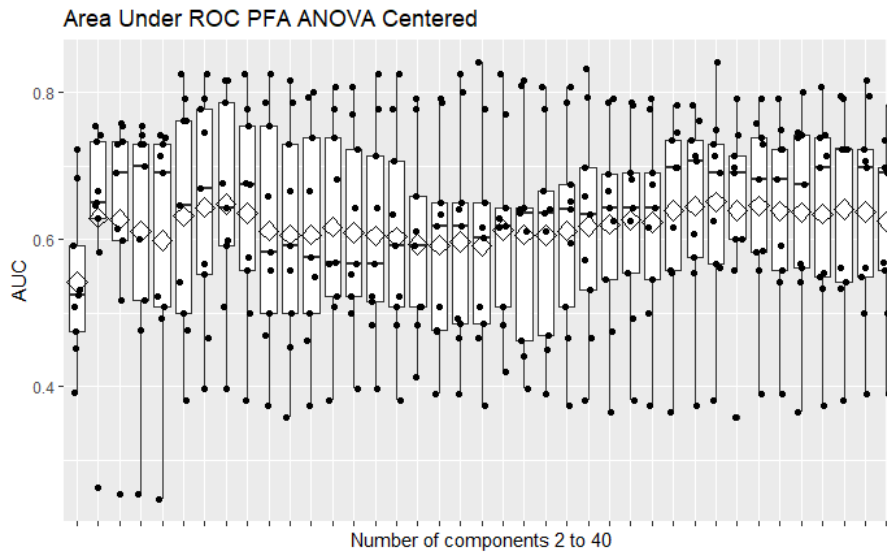


Figura 135. PFR ANOVA centrado.

A97. PCR ESCALADO EN EL MODO 2

Con 33 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.655, con un grupo en torno a 0.4.

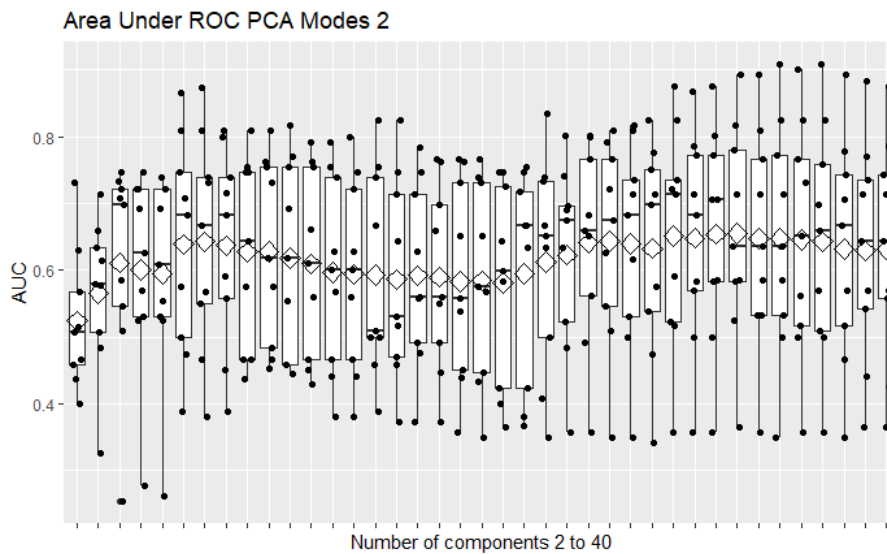


Figura 136. PCR escalado modo 2.

A98. PFR POLINOMIAL DE ORDEN 2 ESCALADO EN EL MODO 2

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.571, con tres de ellos en torno a 0.4.

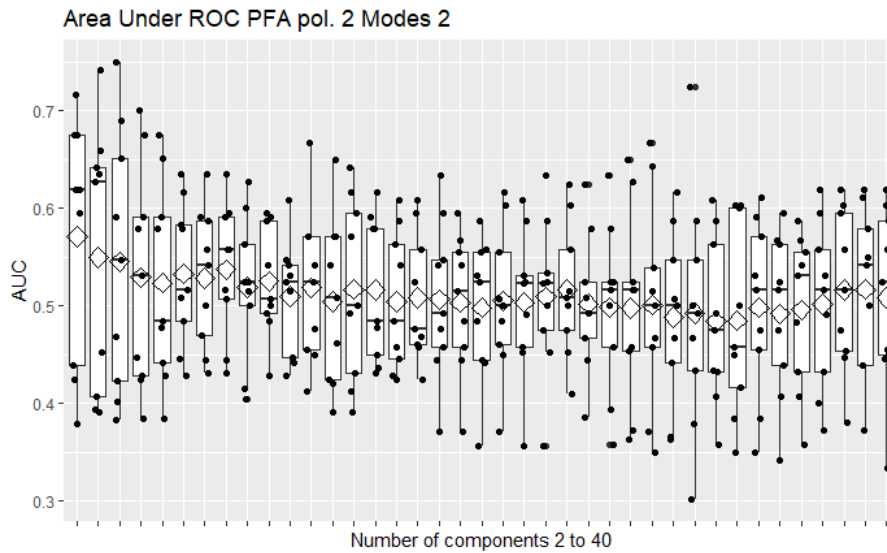


Figura 137. PFR pol. 2 escalado modo 2.

A99. PFR POLINOMIAL DE ORDEN 3 ESCALADO EN EL MODO 2

Con 38 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.612, con tres grupos ligeramente por debajo de 0.5.

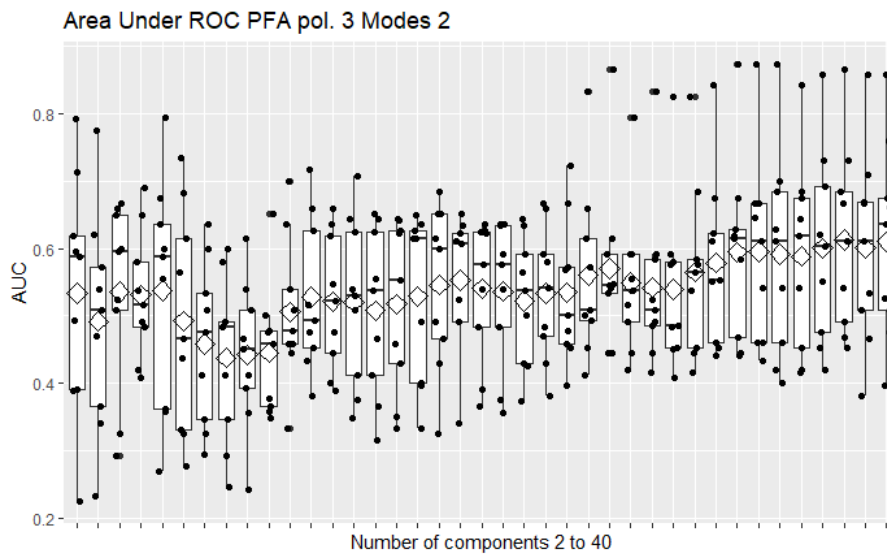


Figura 138. PFR pol. 3 escalado modo 2.

A100. PFR POLINOMIAL DE ORDEN 4 ESCALADO EN EL MODO 2

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.563, con dos grupos en torno a 0.4.

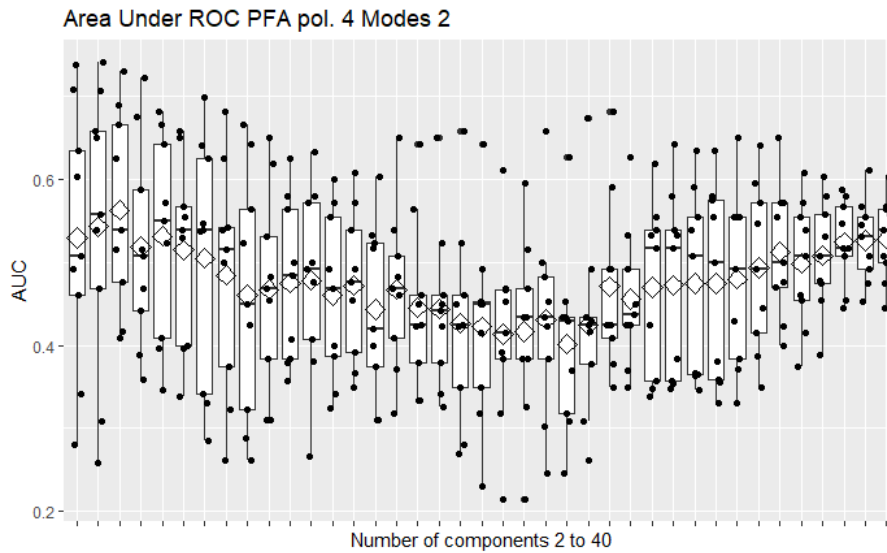


Figura 139. PFR pol. 4 escalado modo 2

A101. PFR ANOVA ESCALADO EN EL MODO 2

Con 26 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.644, con un grupo en torno a 0.4.

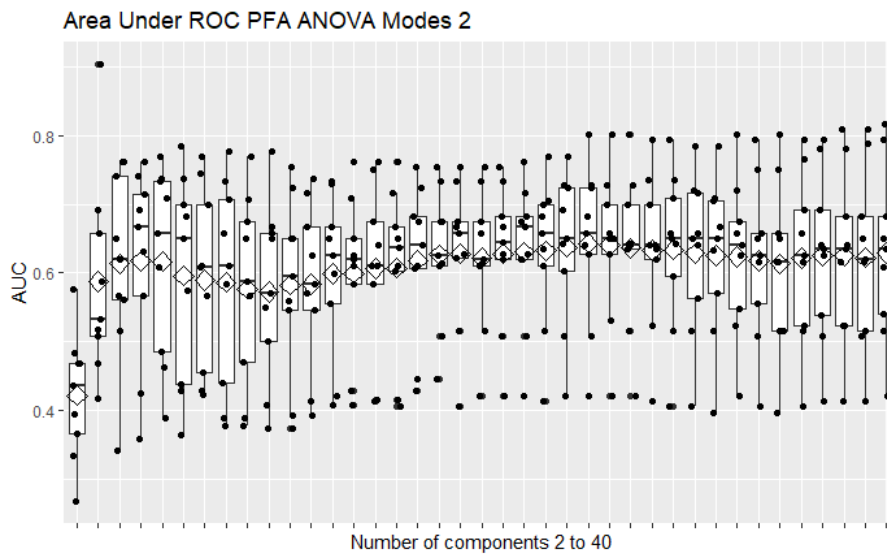


Figura 140. PFR ANOVA escalado modo 2.

A102. PCR ESCALADO EN EL MODO 3

Con 28 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.665, con un grupo en torno a 0.4.

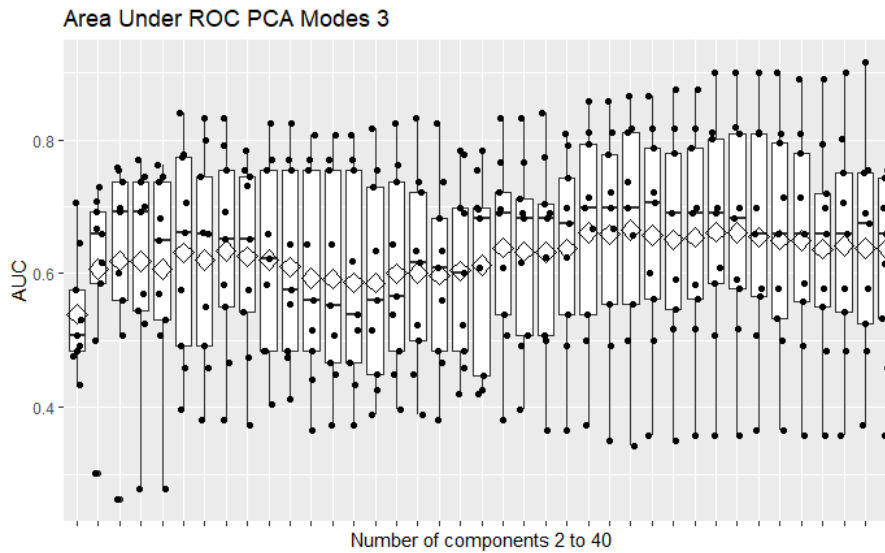


Figura 141. PCR escalado modo 2.

A103. PFR POLINOMIAL DE ORDEN 2 ESCALADO EN EL MODO 3

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.556, dos de ellos en torno a 0.4.

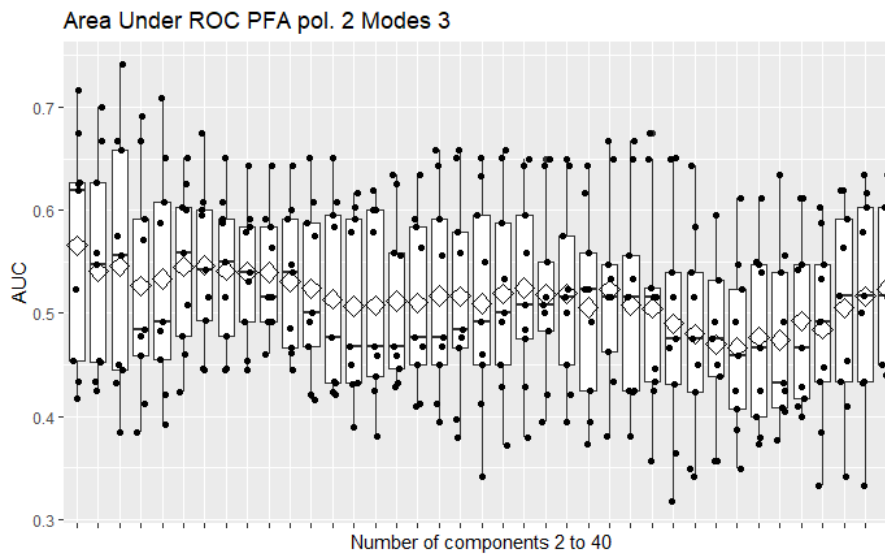


Figura 142. PFA pol. 2 escalado modo 3.

A104. PFR POLINOMIAL DE ORDEN 3 ESCALADO EN EL MODO 3

Con 39 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.607, tres de ellos en torno a 0.45.

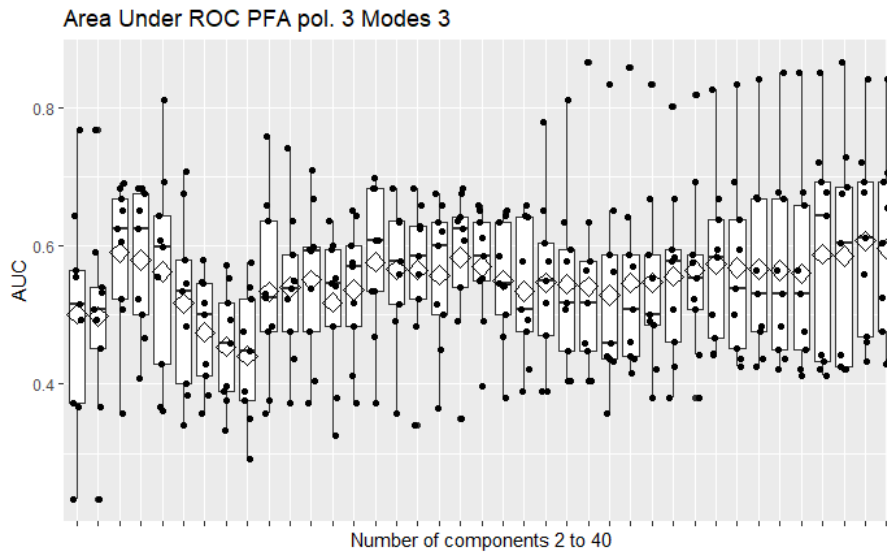


Figura 143. PFA pol. 3 escalado modo 3.

A105. PFR POLINOMIAL DE ORDEN 4 ESCALADO EN EL MODO 3

Con 5 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.584, uno de ellos en torno a 0.4.

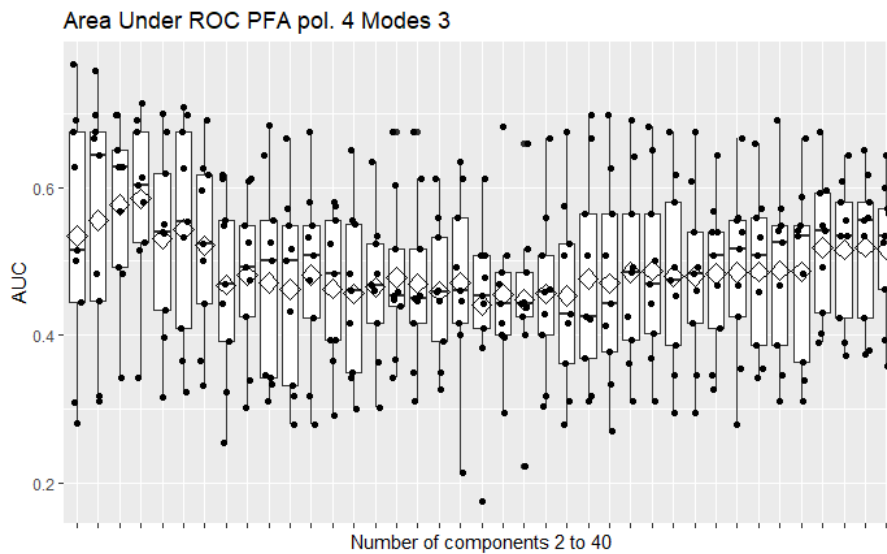


Figura 144. PFA pol. 4 escalado modo 3.

A106. PFR ANOVA ESCALADO EN EL MODO 3

Con 26 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.645, uno de los cuales alrededor de 0.4.

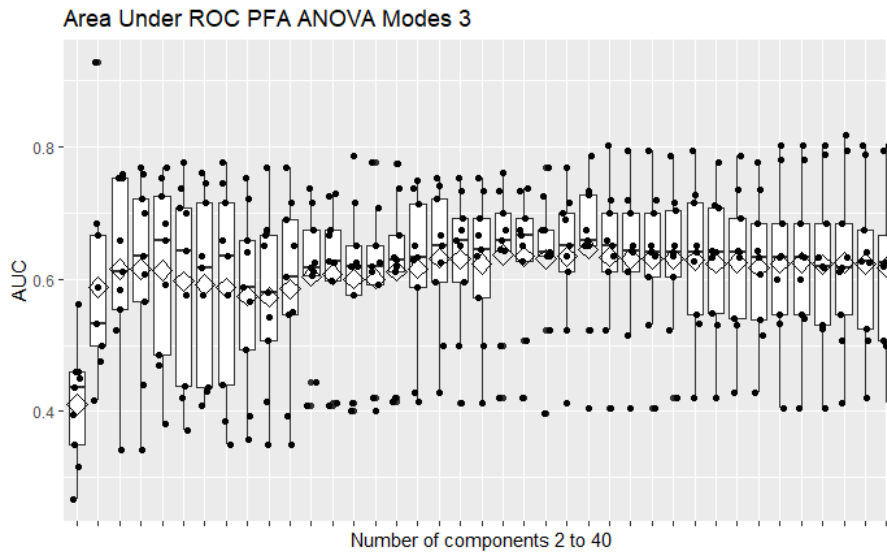


Figura 145. PFA ANOVA escalado modo 3.

A107. PCR ESCALADO EN LOS MODOS 2 Y 3

Con 34 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.661, con un grupo en torno a 0.4.

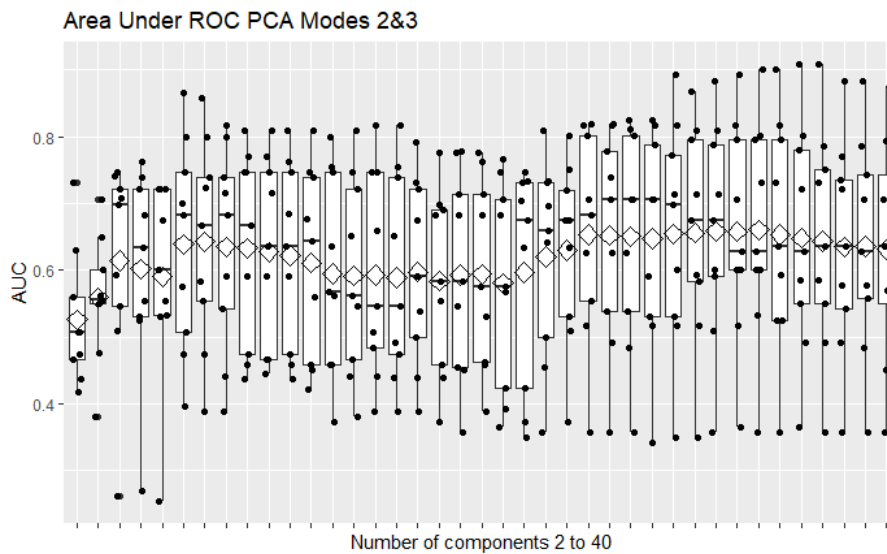


Figura 146. PCR escalado modos 2 y 3.

A108. PFR POLINOMIAL DE ORDEN 2 EN LOS MODOS 2 Y 3

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.567, tres de ellos en torno a 0.4.

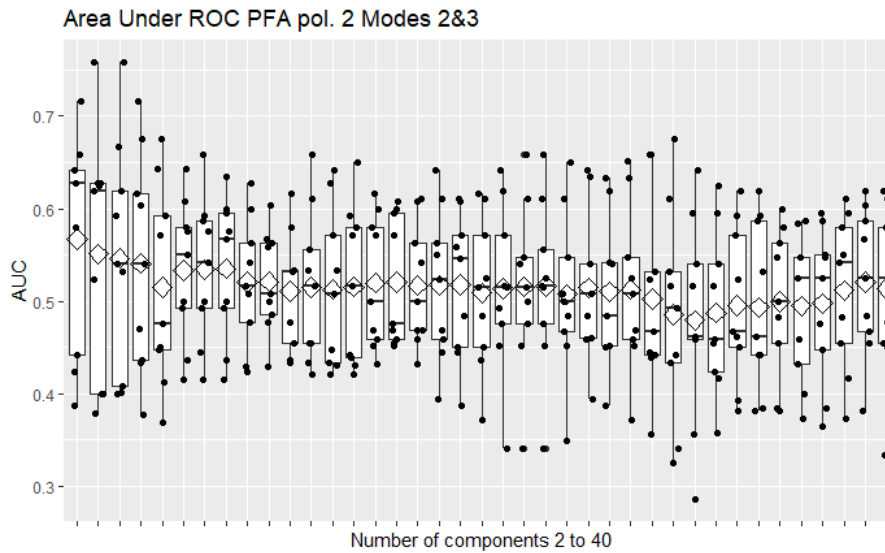


Figura 147. PFR pol. 2 escalado modos 2 y 3.

A109. PFR POLINOMIAL DE ORDEN 3 EN LOS MODOS 2 Y 3

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.609, dos de ellos en torno a 0.45.

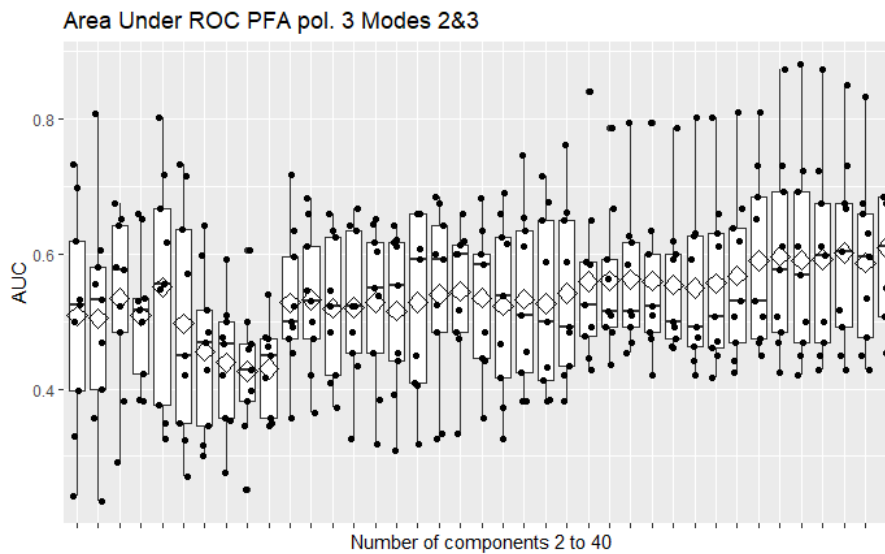


Figura 148. PFR pol. 3 escalado modos 2 y 3.

A110. PFR POLINOMIAL DE ORDEN 4 EN LOS MODOS 2 Y 3

Con 6 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.554, tres de ellos en torno a 0.4.

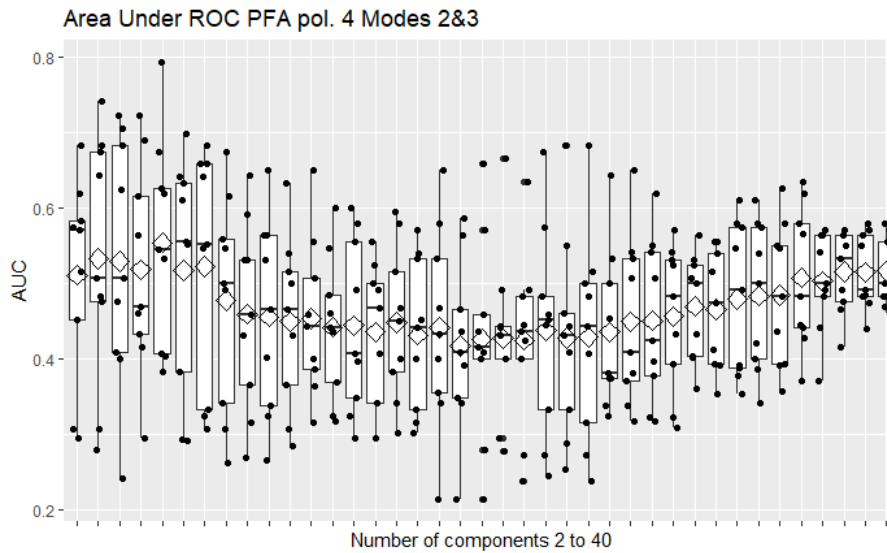


Figura 149. PFR pol. 4 escalado modos 2 y 3.

A111. PFR ANOVA EN LOS MODOS 2 Y 3

Con 26 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.643, un grupo en torno a 0.4.

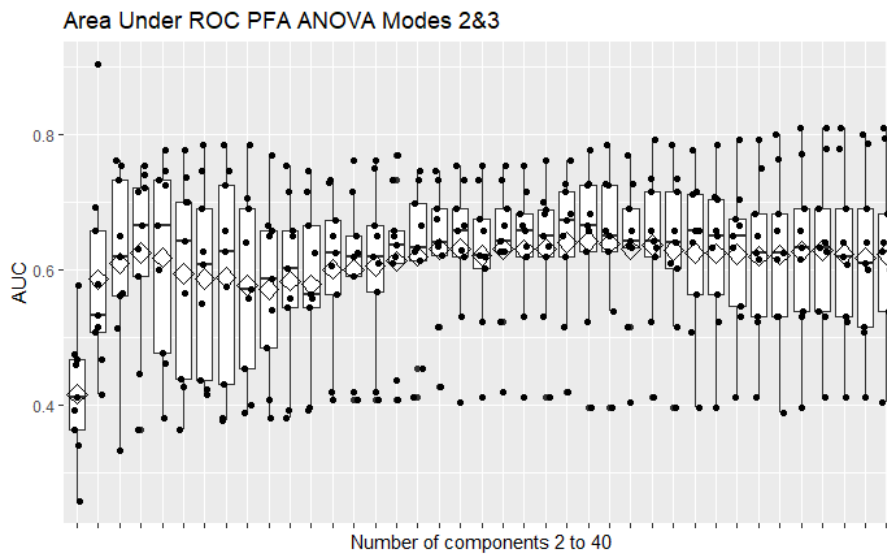


Figura 150. PFR ANOVA escalado modos 2 y 3.

A112. SVM SIN REDUCCIÓN DE LA DIMENSIÓN

Se puede observar que todos los modelos tienen AUC puntuales cercanos o por debajo a 0.4.

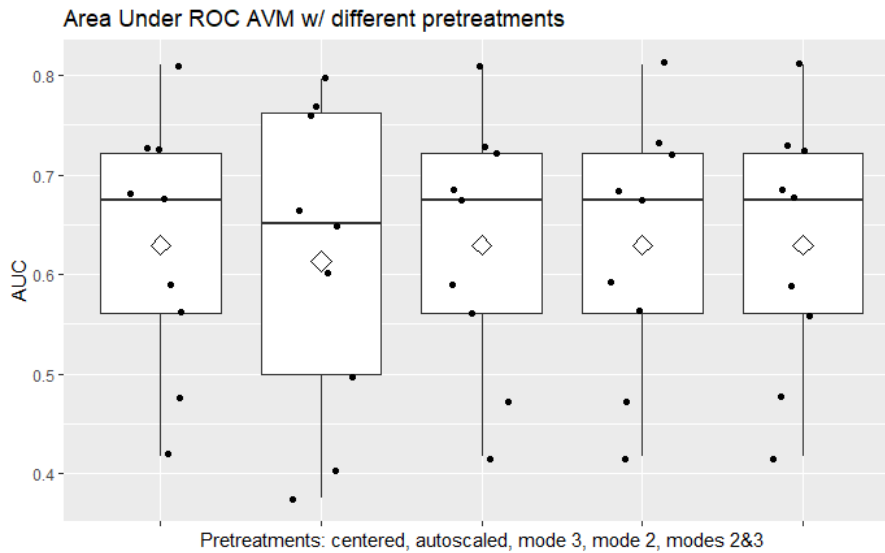


Figura 151. SVM con distintos pretratamientos, sin reducción de la dimensión.

A113. SVM PCA AUTOESCALADO

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.596, con dos grupos en torno a 0.4.

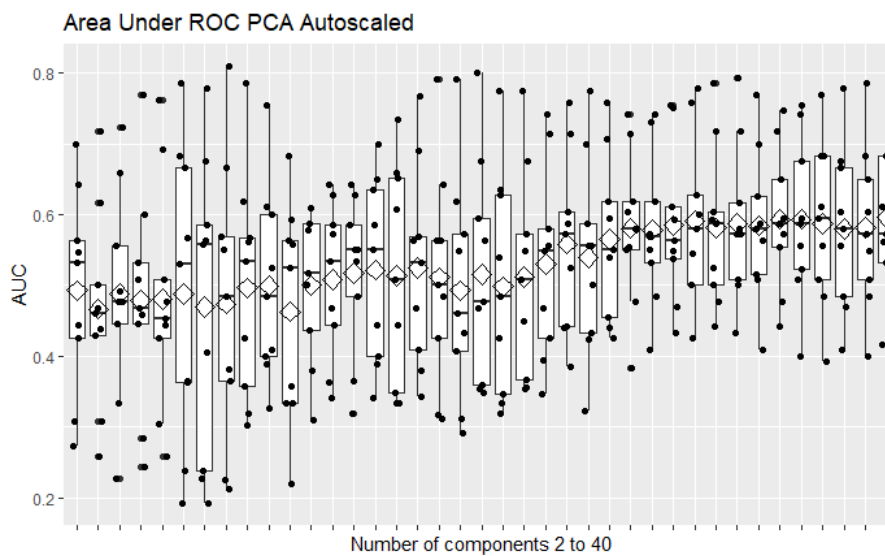


Figura 152. SVM PCA autoescalado.

A114. SVM PFA POL. 2ºG AUTOESCALADO

Con 31 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.595, con un grupo ligeramente por debajo de 0.5.

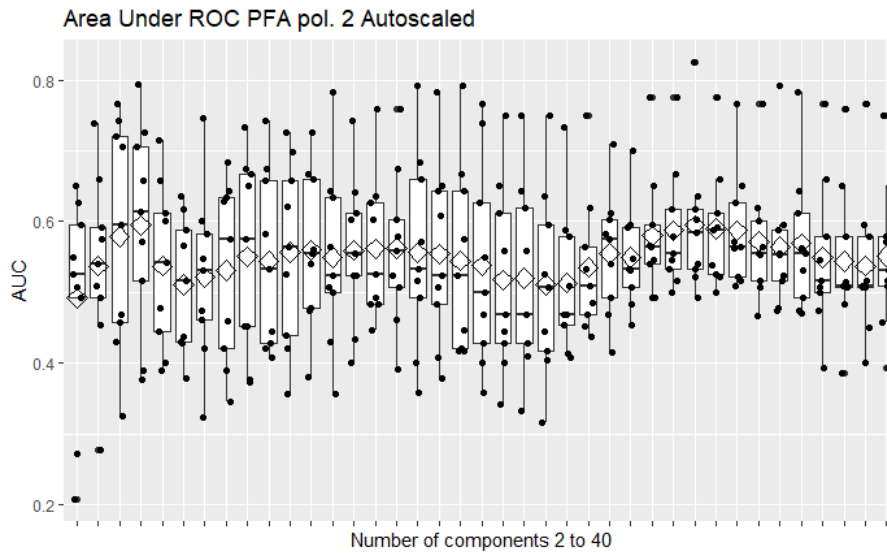


Figura 153. SVM PFA pol. 2 autoescalado.

A115. SVM PFA POL. 3ºG AUTOESCALADO

Con 5 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.608, con un grupo en torno a 0.45 y otro en torno a 0.4.

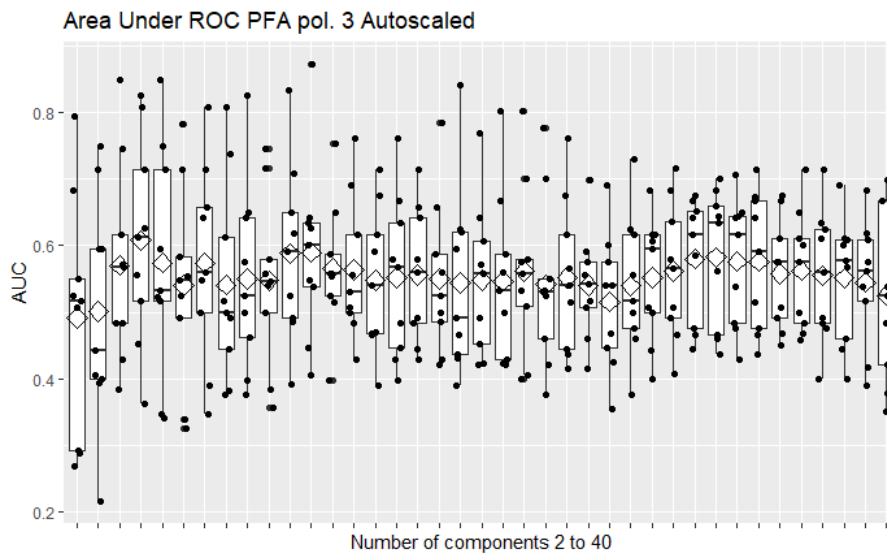


Figura 154. SVM PFA pol. 3 autoescalado.

A116. SVM PFA POL. 4ºG AUTOESCALADO

Con 5 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.606, con un grupo en torno a 0.4 y otro en torno a 0.3.

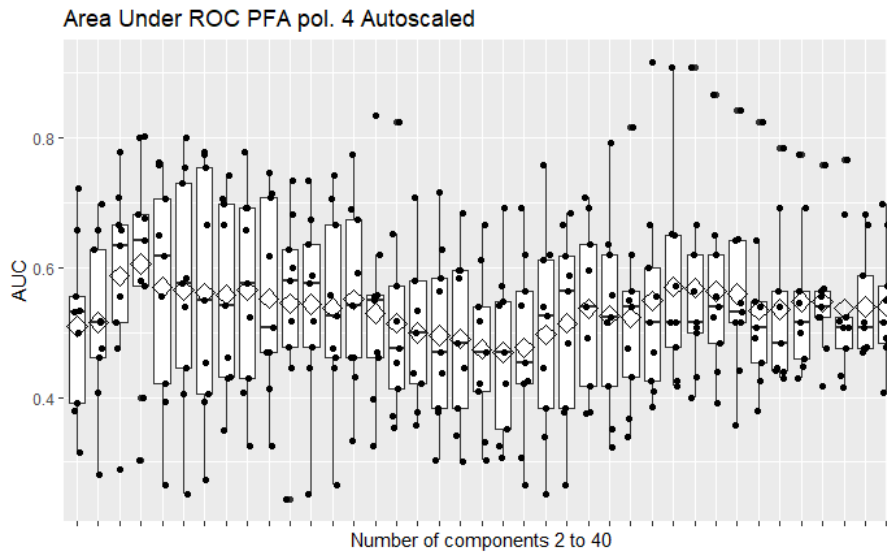


Figura 155. SVM PFA pol. 4 autoescalado.

A117. SVM PFA ANOVA AUTOESCALADO

Con 26 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.614, con un grupo en torno a 0.4.

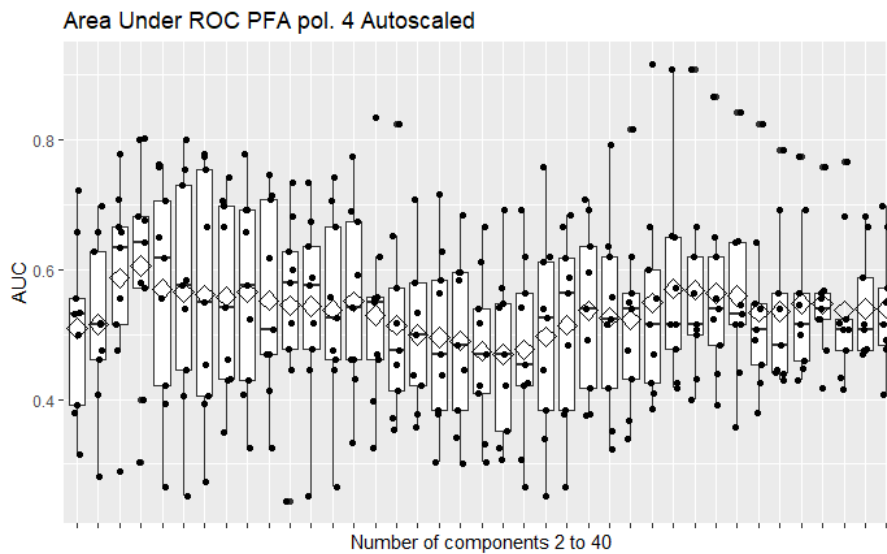


Figura 156. SVM PFA ANOVA autoescalado.

A118. SVM PCA CENTRADO

Con 32 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.638, con un grupo en torno a 0.4 y otro en torno a 0.45.

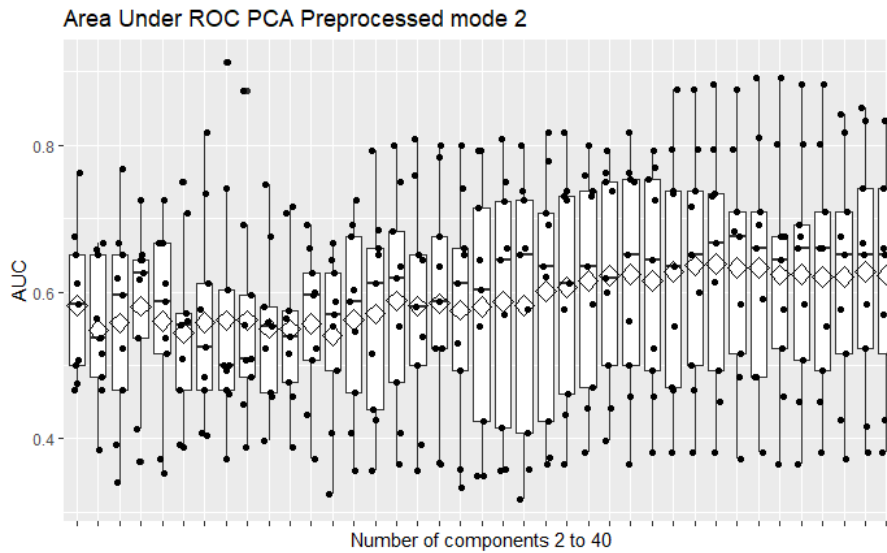


Figura 157. SVM PCA centrado.

A119. SVM PFA POL. 2ºG CENTRADO

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.605, con un grupo en torno a 0.4.

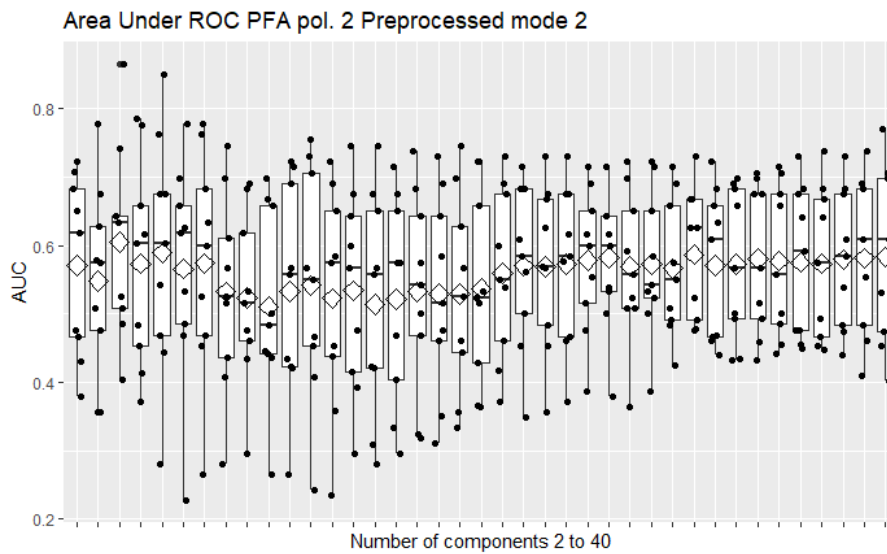


Figura 158. SVM PFA pol. 2 centrado.

A120. SVM PFA POL. 3ºG CENTRADO

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.596, con un grupo en torno a 0.3 y dos en torno a 0.4.

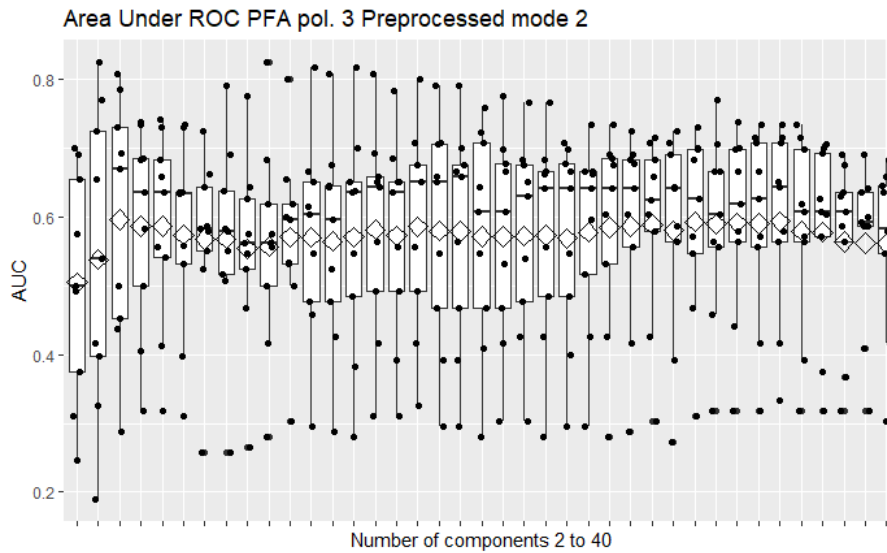


Figura 159. SVM PFA pol. 3 centrado.

A121. SVM PFA POL. 4ºG CENTRADO

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.591, con un grupo en torno a 0.2 y otro en torno a 0.35.

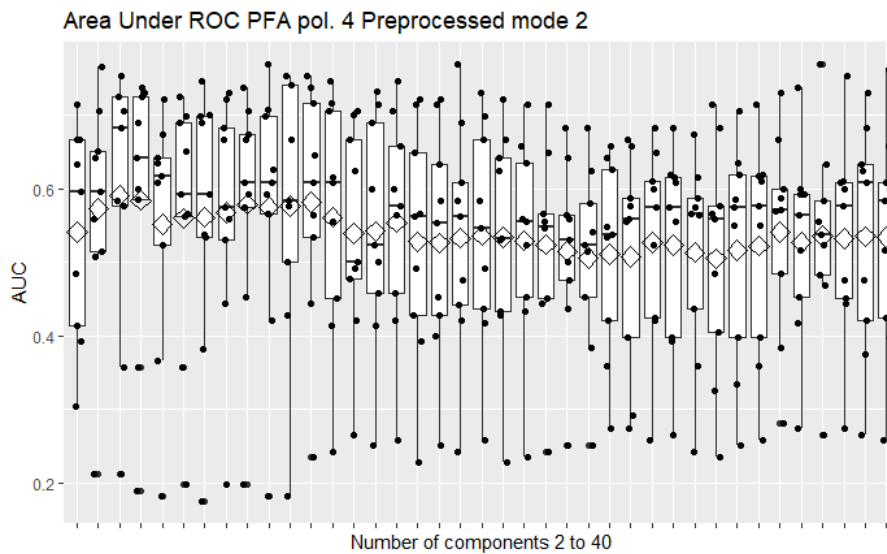


Figura 160. SVM PFA pol. 4 centrado.

A122. SVM PFA ANOVA CENTRADO

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.624, con dos grupos en torno a 0.4.

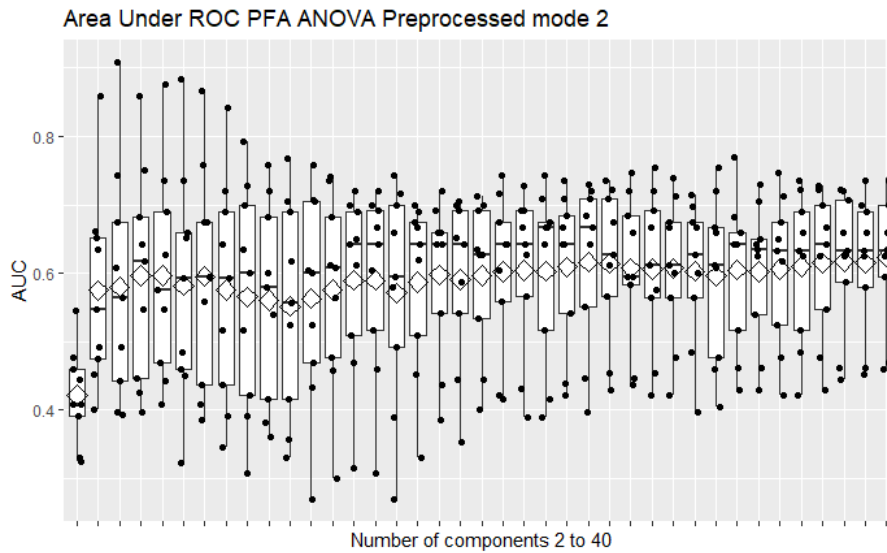


Figura 161. SVM PFA ANOVA centrado.

A123. SVM PCA ESCALADO MODO 2

Con 32 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.638, con un grupo en torno a 0.4 y otro en torno a 0.45.

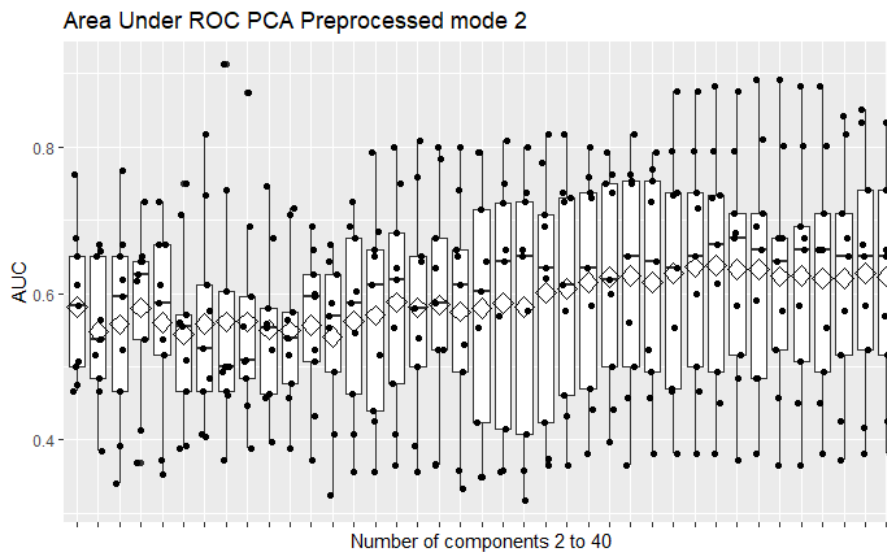


Figura 162. SVM PCA escalado modo 2.

A124. SVM PFA POL. 2ºG ESCALADO MODO 2

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de: 0.605, con un grupo en torno a 0.4.

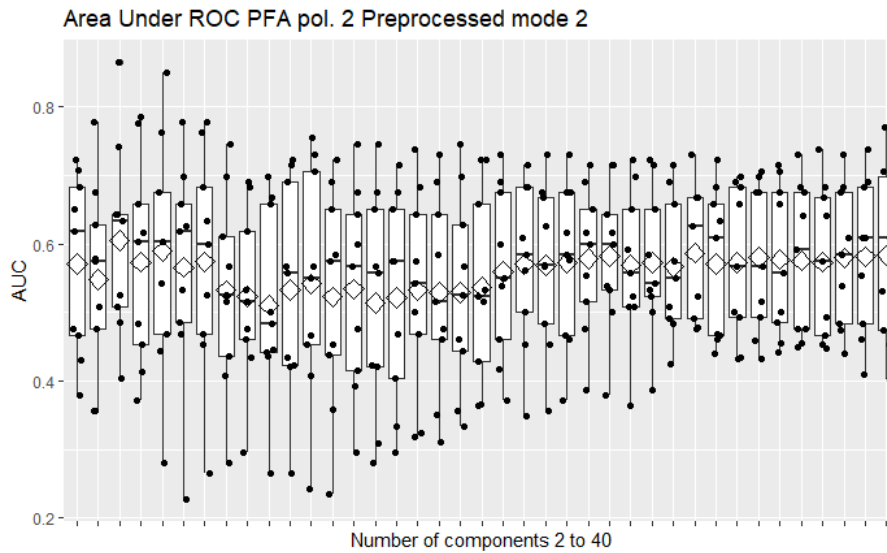


Figura 163. SVM PFA pol. 2 escalado modo 2.

A125. SVM PFA POL. 3ºG ESCALADO MODO 2

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.596, con dos grupos en torno a 0.45 y otro en torno a 0.3.

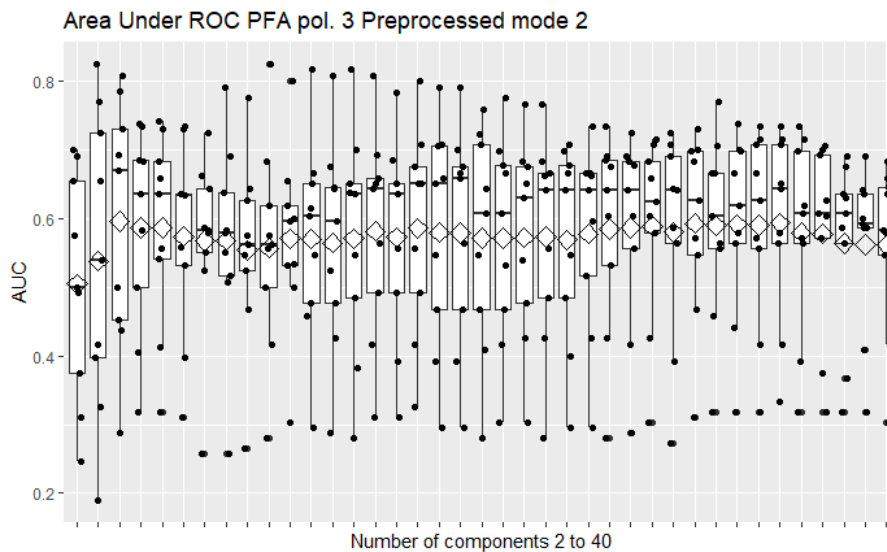


Figura 164. SVM PFA pol. 3 escalado modo 2.

A126. SVM PFA POL. 4ºG ESCALADO MODO 2

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.591, con un grupo en torno a 0.2 y otro en torno a 0.35.

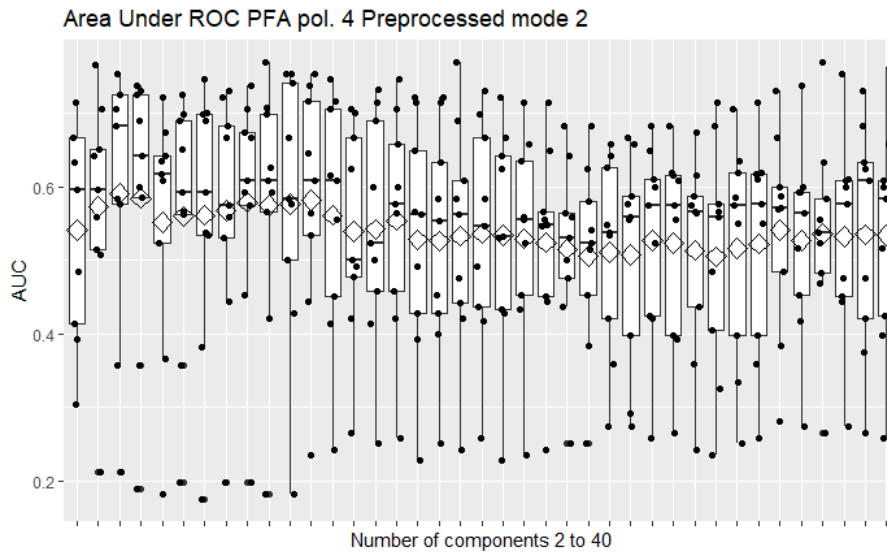


Figura 165. SVM PFA pol. 4 escalado modo 2.

A127. SVM PFA ANOVA ESCALADO MODO 2

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.623, con dos grupos en torno a 0.45.

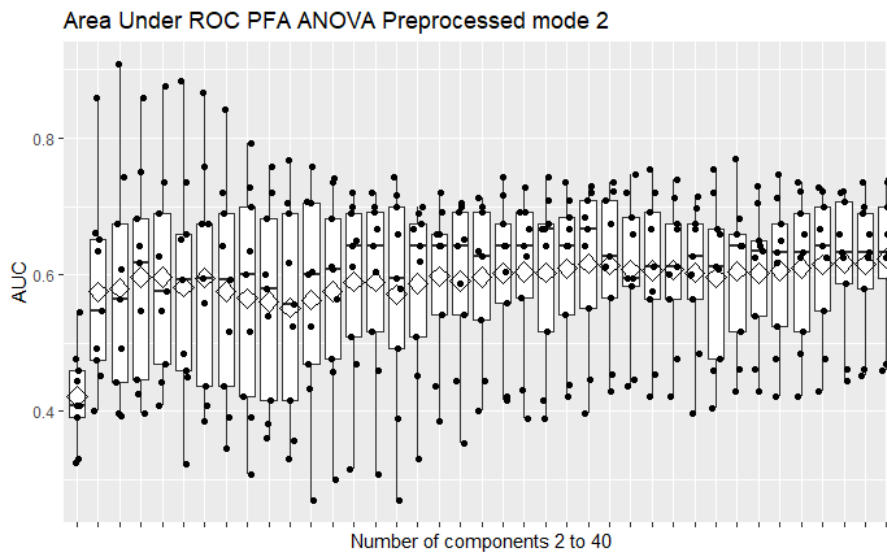


Figura 166. SVM PFA ANOVA escalado modo 2.

A128. SVM PCA ESCALADO MODO 3

Con 33 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.626, con dos de los grupos en torno a 0.4.

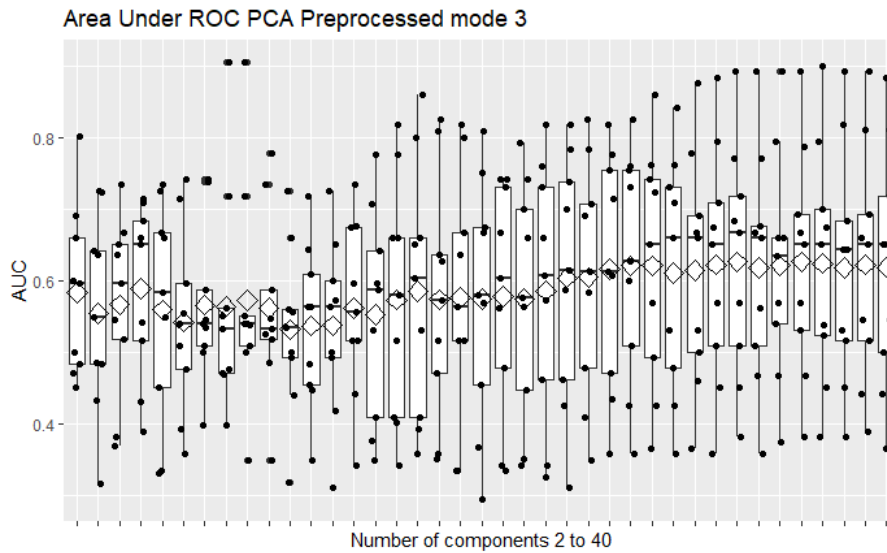


Figura 167. SVM PCA escalado modo 3.

A129. SVM PFA POL. 2ºG ESCALADO MODO 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.589, con un grupo en torno a 0.4 y otro en torno a 0.45.

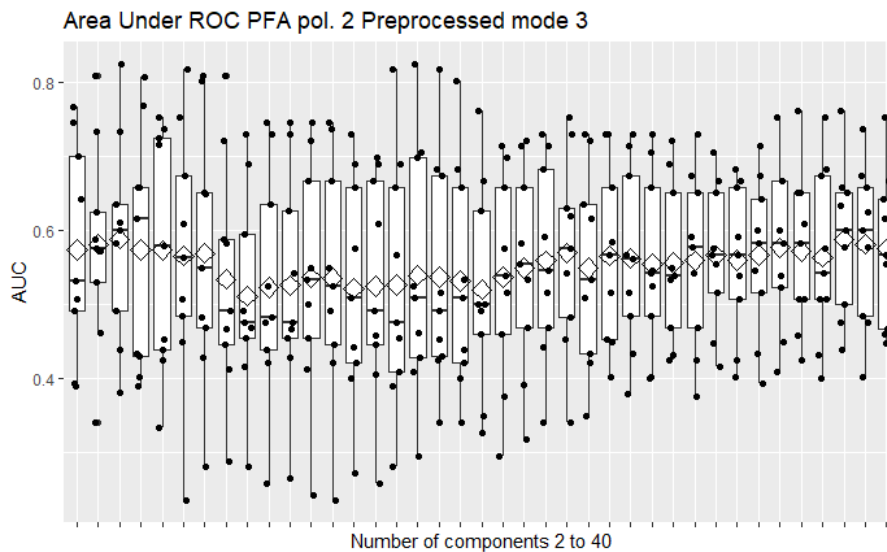


Figura 168. SVM PFA pol. 2 escalado modo 3.

A130. SVM PFA POL. 3ºG ESCALADO MODO 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.591, con un grupo en torno a 0.4.

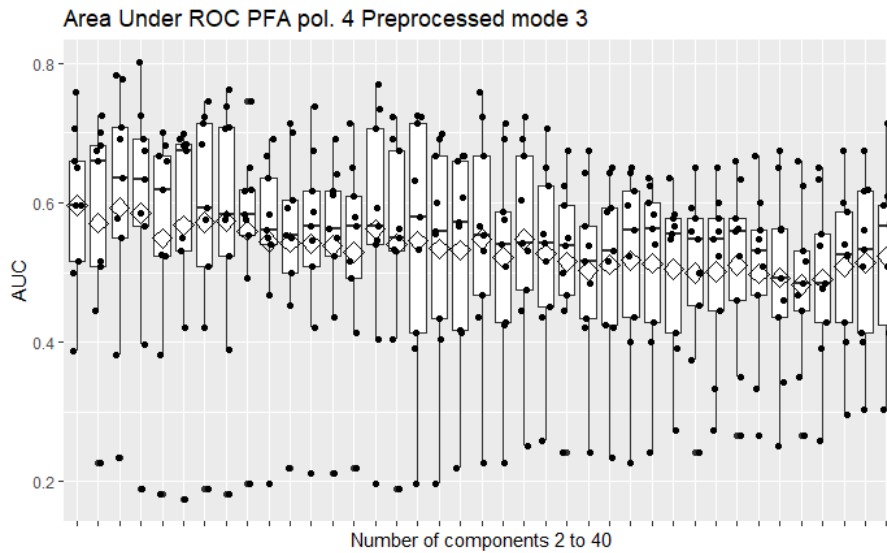


Figura 169. SVM PFA pol. 3 escalado modo 3.

A131. SVM PFA POL. 4ºG ESCALADO MODO 3

Con 2 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.596, con dos grupos en torno a 0.3.

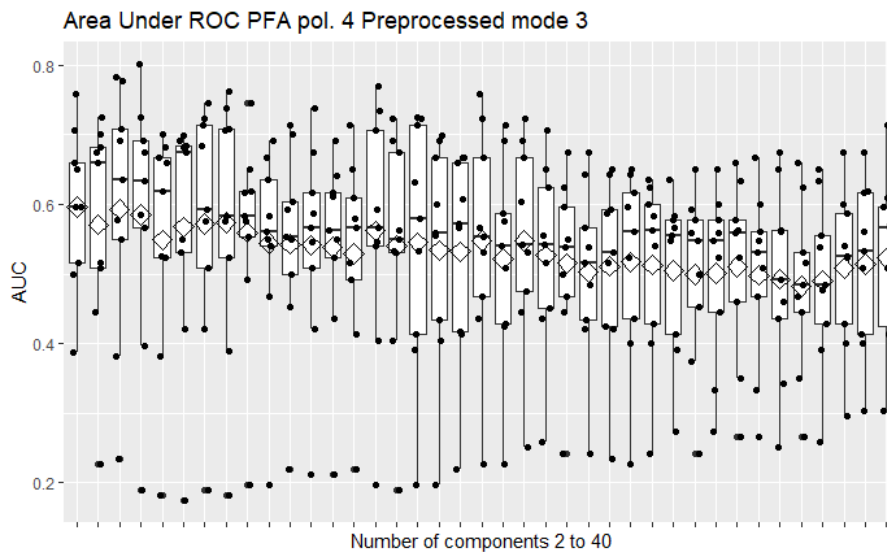


Figura 170. SVM PFA pol. 4 escalado modo 3.

A132. SVM PFA ANOVA ESCALADO MODO 3

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.628, con dos grupos en torno a 0.45.

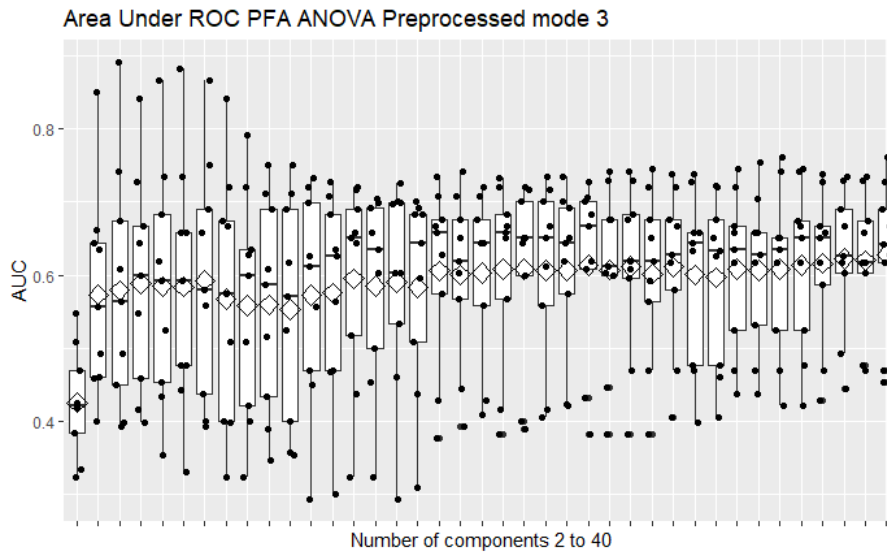


Figura 171. SVM PFA ANOVA escalado modo 3.

A133. SVM PCA ESCALADO MODOS 2 Y 3

Con 34 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.630, con un grupo en torno a 0.4 y otro en torno a 0.45.

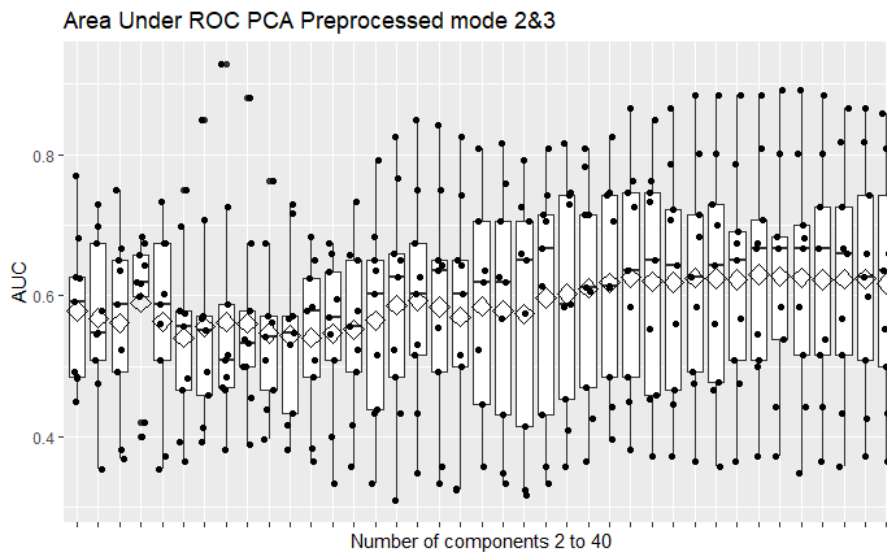


Figura 172. SVM PCA escalado modos 2 y 3.

A134. SVM PFA POL. 2ºG ESCALADO MODOS 2 Y 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.601, con dos grupos en torno a 0.4.

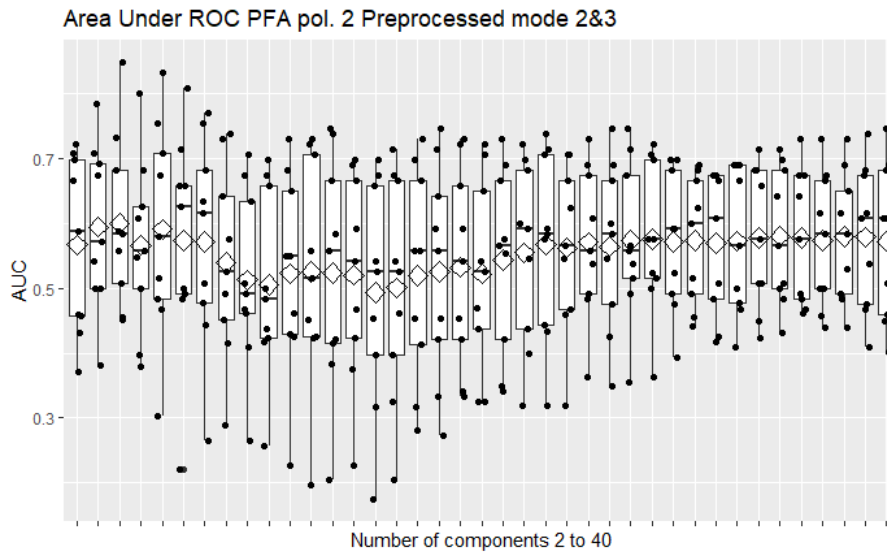


Figura 173. SVM PFA pol. 2 escalado modos 2 y 3.

A135. SVM PFA POL. 3ºG ESCALADO MODOS 2 Y 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.600, con un grupo en torno a 0.45 y otro en torno a 0.35.

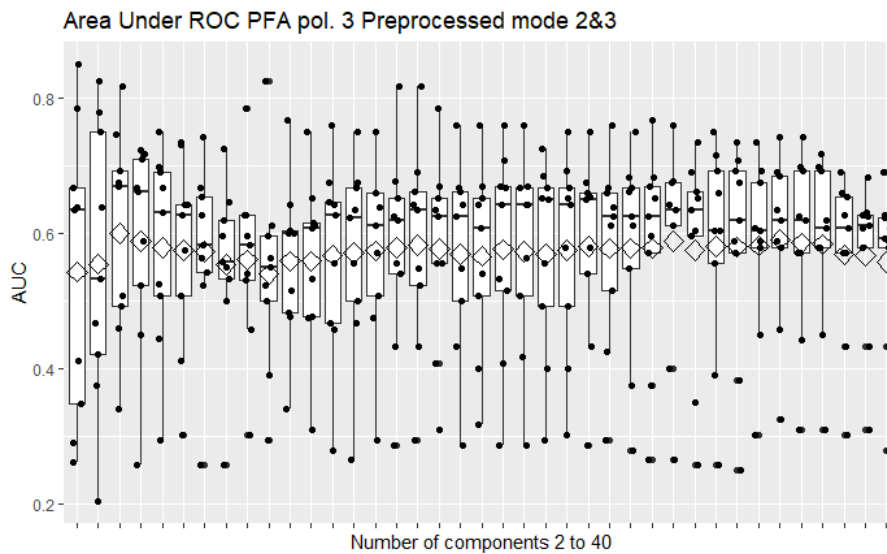


Figura 174. SVM PFA pol. 3 escalado modos 2 y 3.

A136. SVM PFA POL. 4ºG ESCALADO MODOS 2 Y 3

Con 4 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.596, con un grupo en torno a 0.25 y otro en torno a 0.35.

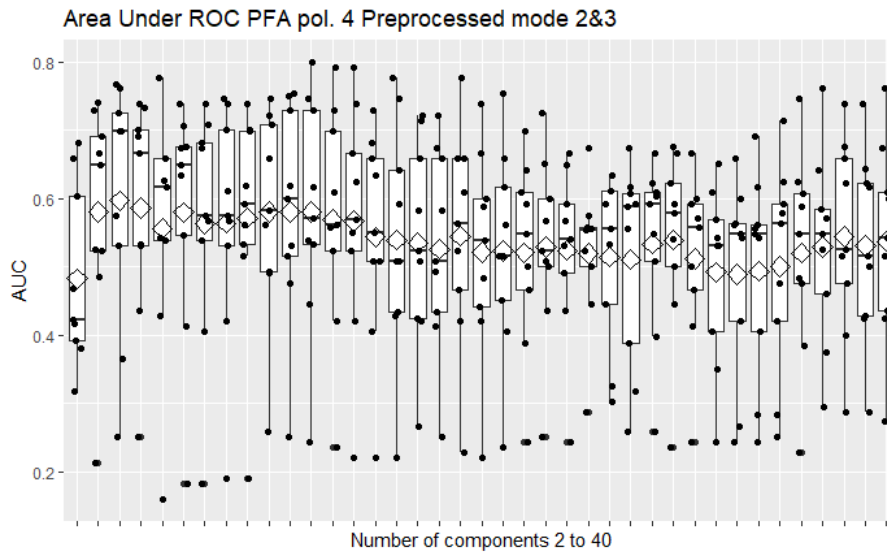


Figura 175. SVM PFA pol. 4 escalado modos 2 y 3.

A137. SVM PFA ANOVA ESCALADO MODOS 2 Y 3

Con 40 variables latentes obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.624, con dos de los grupos ligeramente por debajo de 0.5.

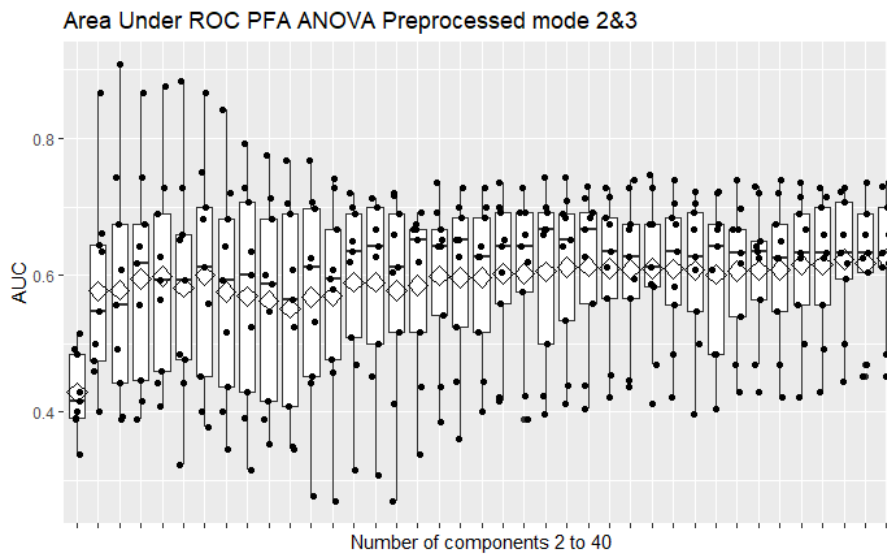


Figura 176. SVM PFA ANOVA escalado modos 2 y 3.

A138. SNPLSDA AUTOESCALADO

Con 100 puntos de corriente y 5 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.626, con algunos grupos por debajo de 0.5.

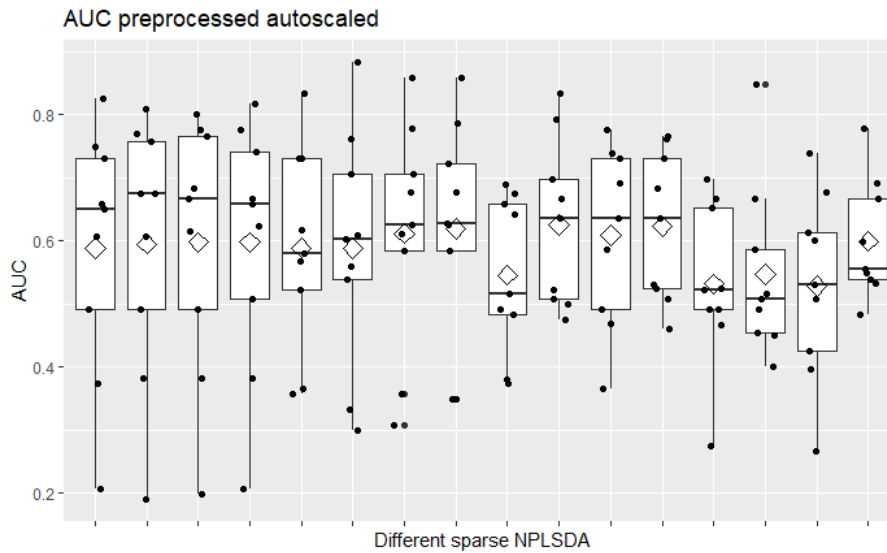


Figura 177. sNPLSDA 50 puntos de corriente y dos sensores, autoescalado.

A139. SNPLSDA CENTRADO

Con 50 puntos de corriente y 3 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, con dos grupos en torno a 0.4.

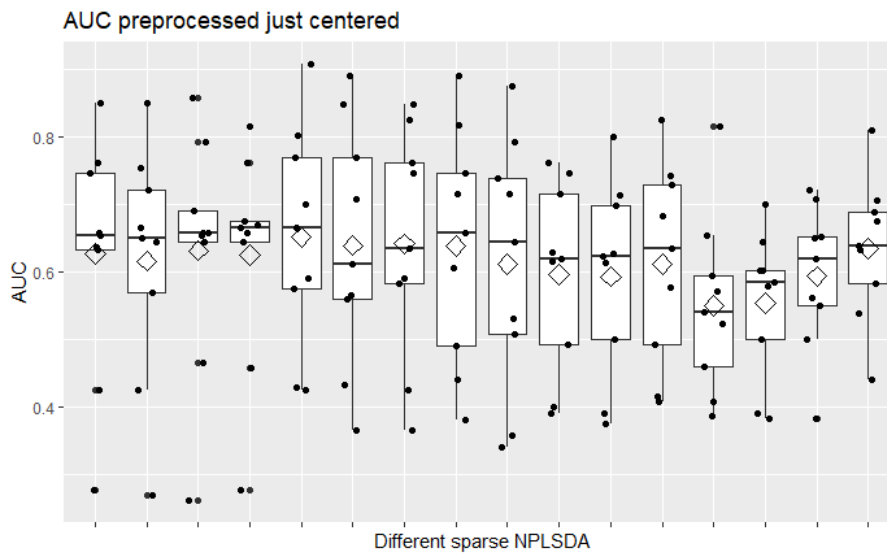


Figura 178. sNPLSDA 50 puntos de corriente y cinco sensores, centrado.

A140. SNPLSDA ESCALADO EN EL MODO 2

Con 300 puntos de corriente y 8 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.624, con un grupo en torno a 0.45.

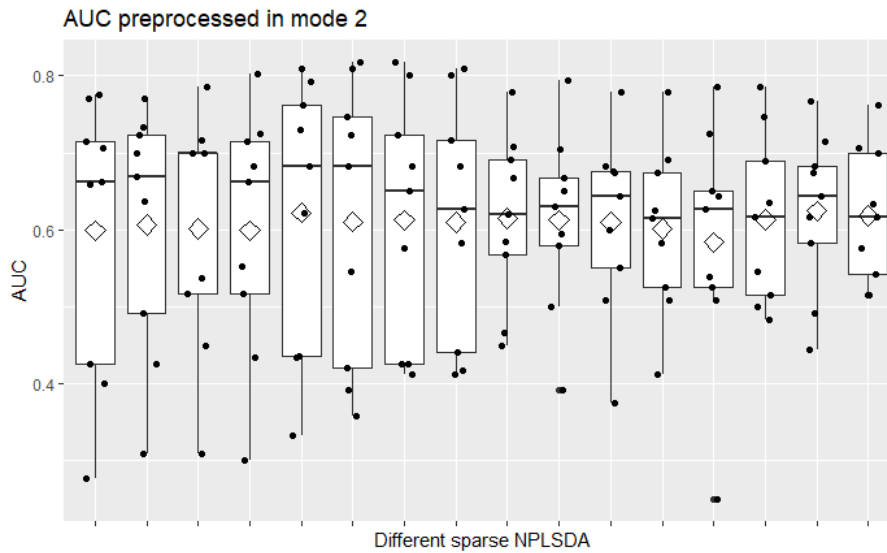


Figura 179. sNPLSDA 50 puntos de corriente y tres sensores, escalado modo 2.

A141. SNPLSDA ESCALADO EN EL MODO 3

Con 50 puntos de corriente y 3 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.675, con un grupo en torno a 0.4.

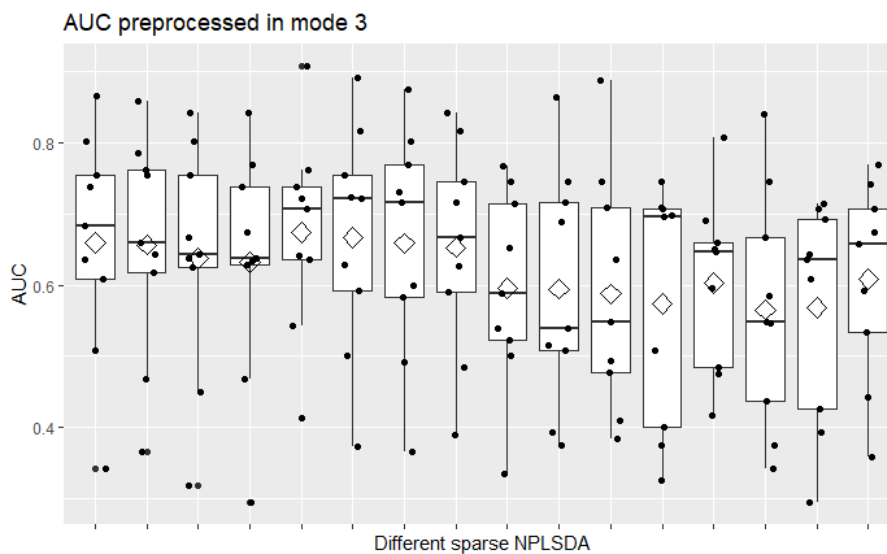


Figura 180. sNPLSDA 50 puntos de corriente y tres sensores, escalado modo 3.

A142. SNPLSDA ESCALADO EN LOS MODOS 2 Y 3

Con 100 puntos de corriente y 8 sensores obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.649, con un grupo en torno a 0.4

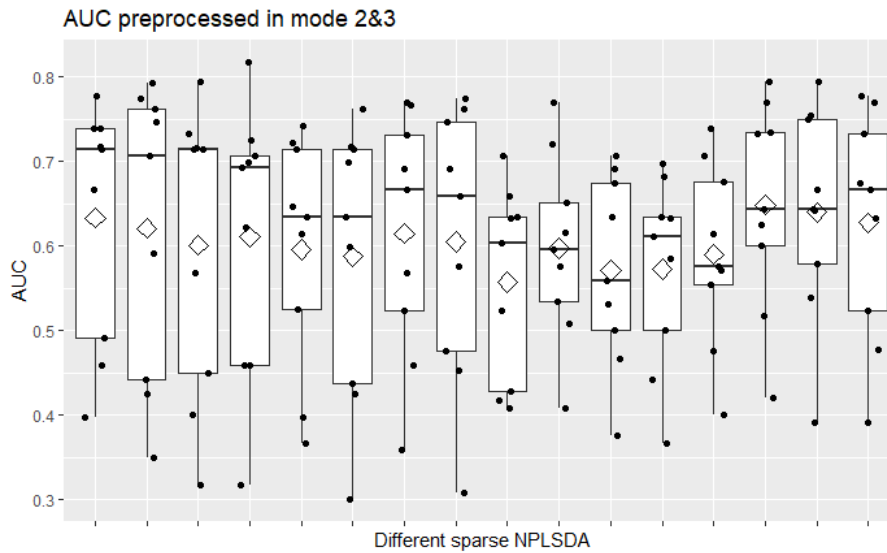


Figura 181. sNPLSDA 50 puntos de corriente y tres sensores, escalado modos 2 y 3.

A143. ANN SIN REDUCCIÓN DE LA DIMENSIÓN UNA CAPA AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.629, con dos grupos por debajo de 0.5.

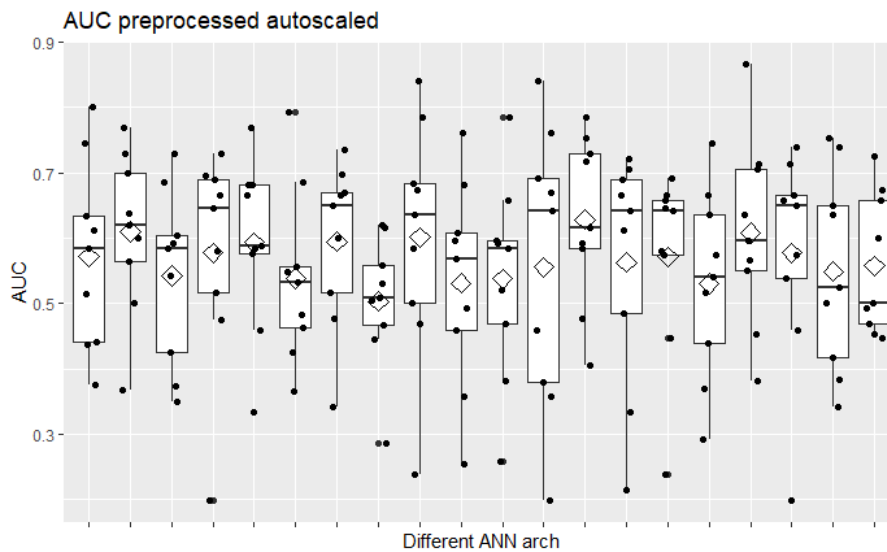


Figura 182. AUC de ANN de una capa, sin reducción de la dimensión, autoescalado.

A143. ANN SIN REDUCCIÓN DE LA DIMENSIÓN UNA CAPA CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.644, con un grupo en torno a 0.3.

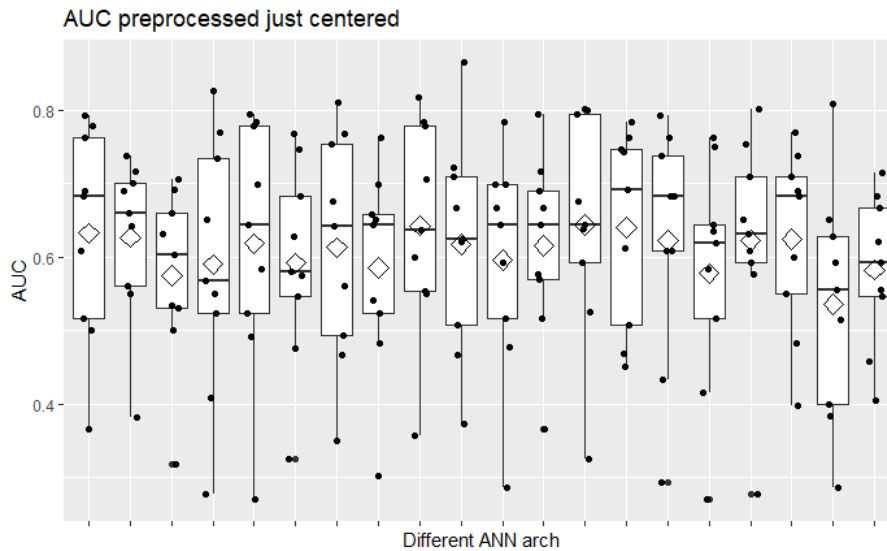


Figura 183. AUC de ANN de una capa, sin reducción de la dimensión, centrado.

A144. ANN SIN REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.648, con un grupo en torno a 0.4.

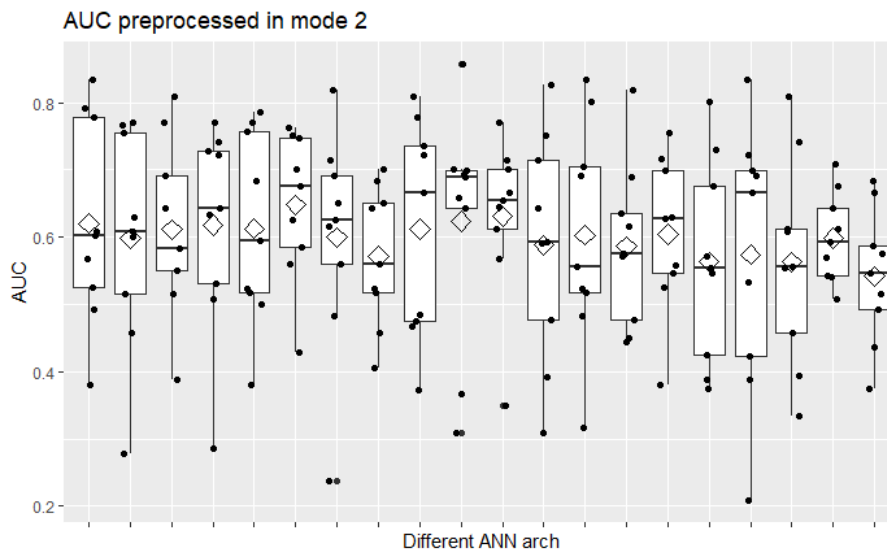


Figura 184. AUC de ANN de una capa, sin reducción de la dimensión, escalado modo 2.

A145. ANN SIN REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.639 con un grupo ligeramente por debajo de 0.5 y otro en torno a 0.4.

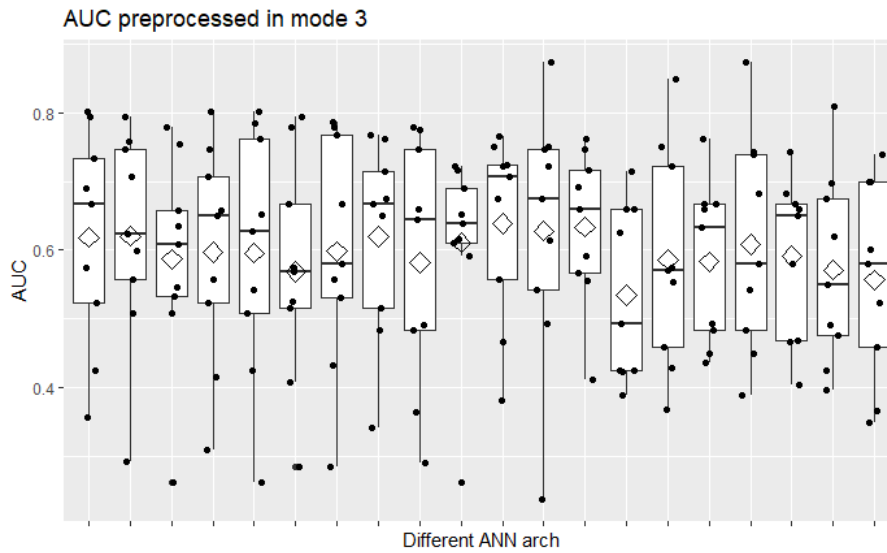


Figura 185. AUC de ANN de una capa, sin reducción de la dimensión, escalado modo 3.

A146. ANN SIN REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.618, con un grupo ligeramente por debajo de 0.5 y otro en torno a 0.4.

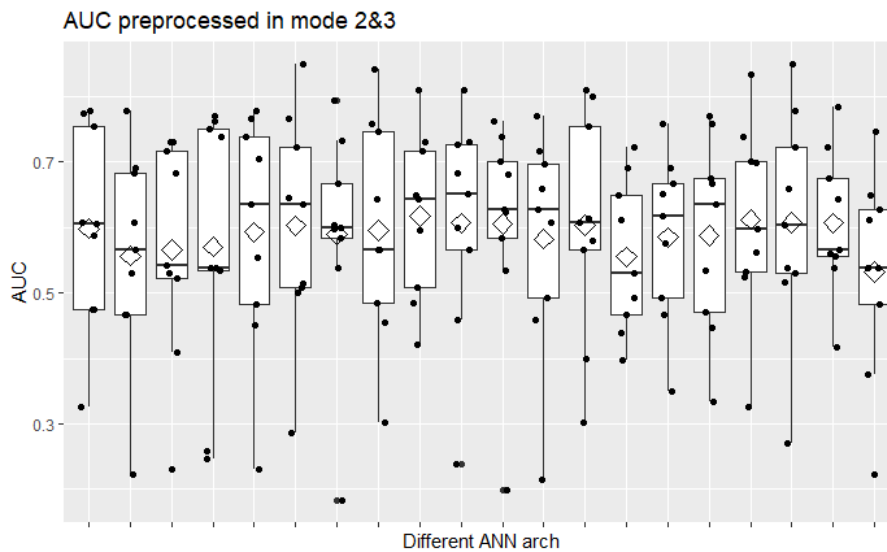


Figura 186. AUC de ANN de una capa, sin reducción de la dimensión, escalado modos 2&3.

A147. ANN SIN REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.639, con un grupo en torno a 0.4.

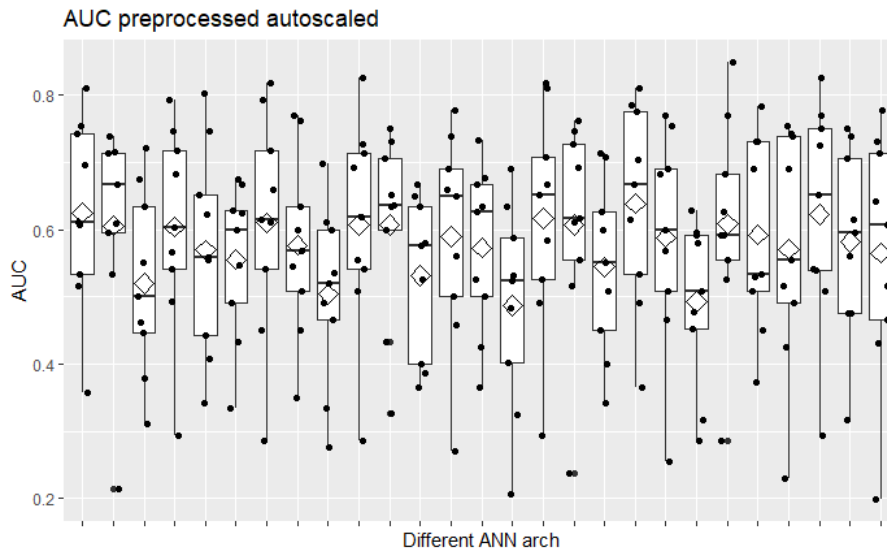


Figura 187. AUC de ANN de dos capas, sin reducción de la dimensión, autoescalado.

A148. ANN SIN REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.641, con dos grupos en torno a 0.4.

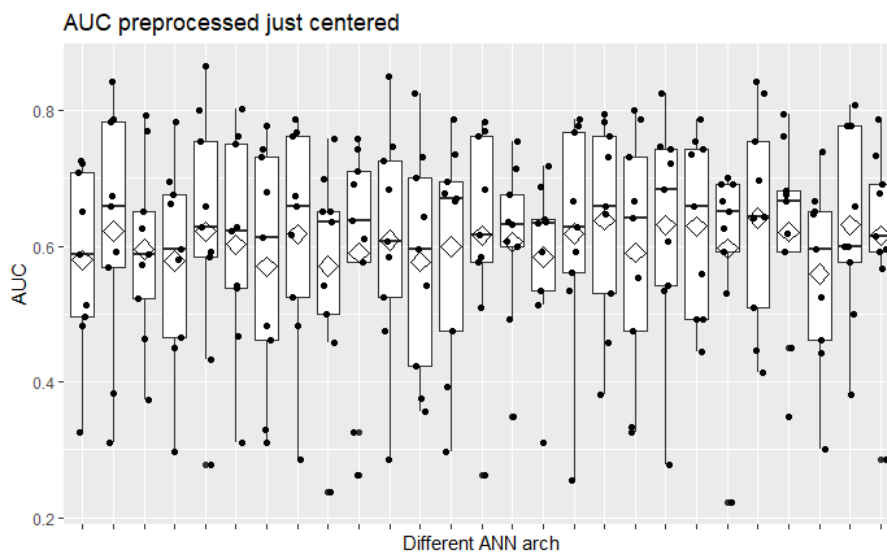


Figura 188. AUC de ANN de dos capas, sin reducción de la dimensión, centrado.

A149. ANN SIN REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.651, con un grupo en torno a 0.4.

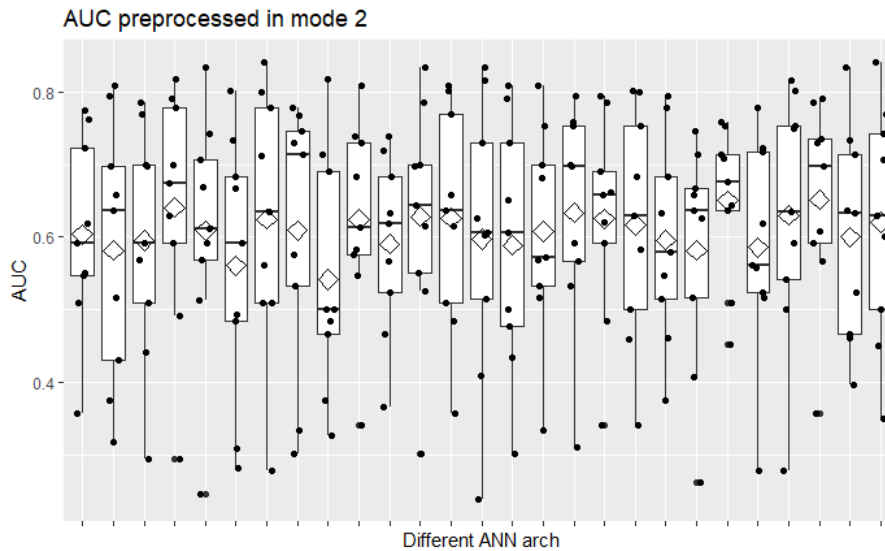


Figura 189. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modo 2.

A150. ANN SIN REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.660, con un grupo por debajo de 0.5.

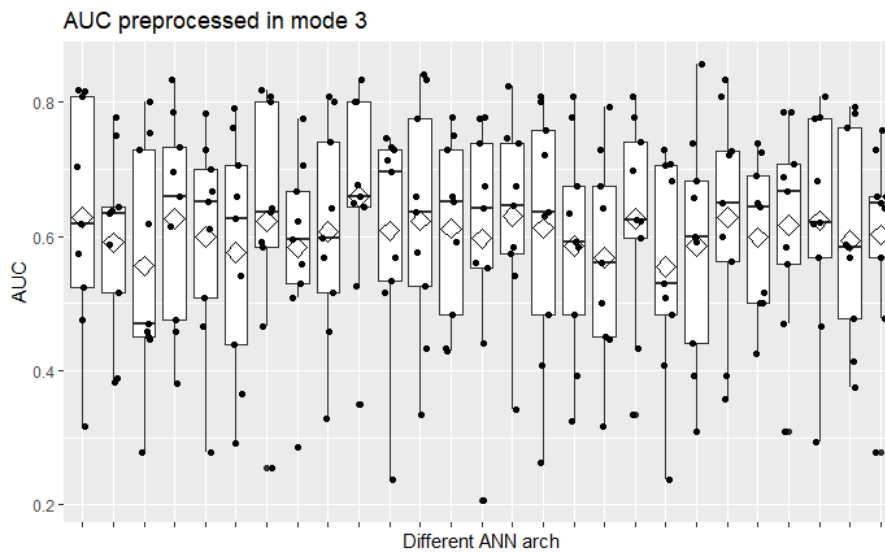


Figura 190. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modo 3.

A151. ANN SIN REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.650, con un grupo por debajo de 0.35.

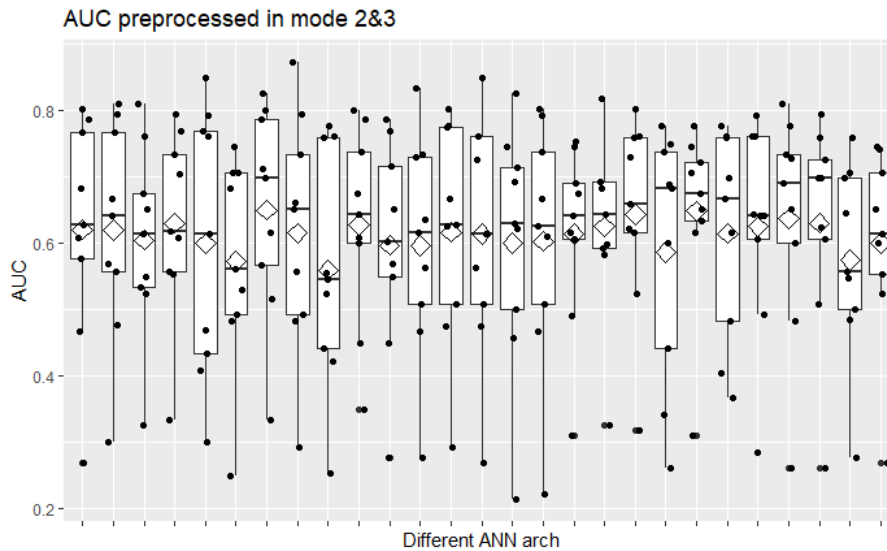


Figura 191. AUC de ANN de dos capas, sin reducción de la dimensión, escalado modos 2&3.

A152. ANN SIN REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.641, con un grupo en torno a 0.35.

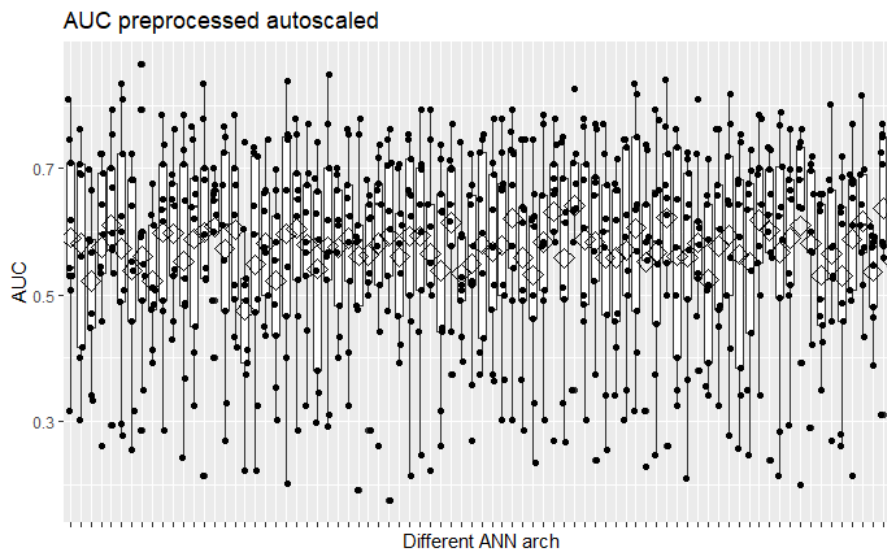


Figura 192. AUC de ANN de tres capas, sin reducción de la dimensión, autoescalado.

A153. ANN SIN REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.644, con ninun grupo en torno a 0.35.

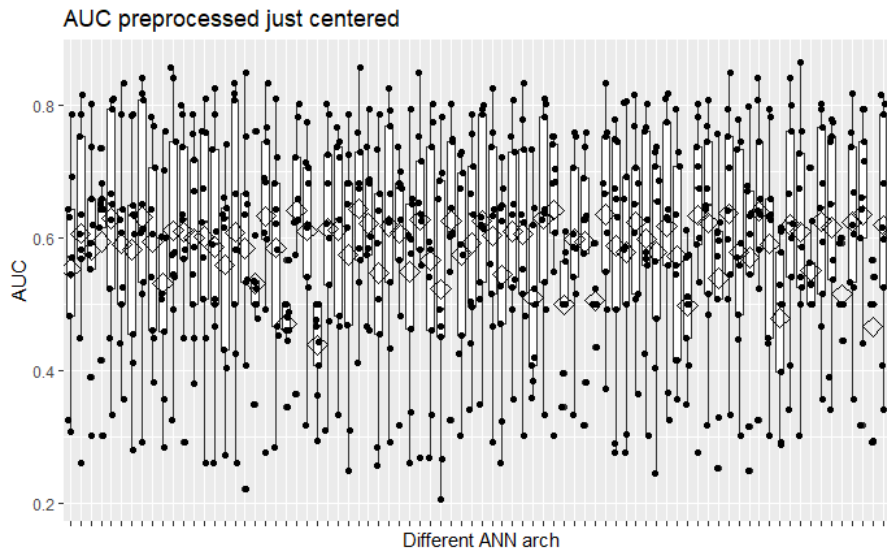


Figura 193. AUC de ANN de tres capas, sin reducción de la dimensión, centrado.

A154. ANN SIN REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.633, con un grupo en torno a 0.3.

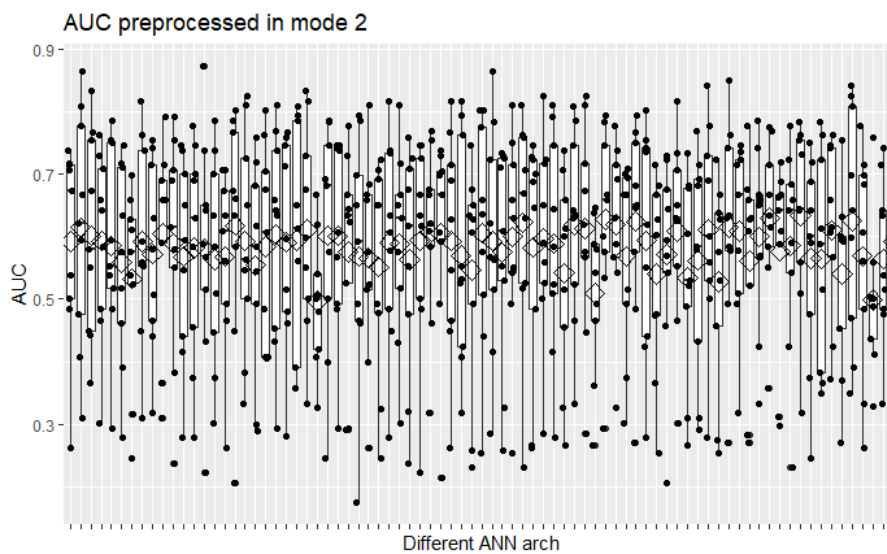


Figura 194. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modo 2.

A155. ANN SIN REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.658, con un grupo por debajo de 0.35.

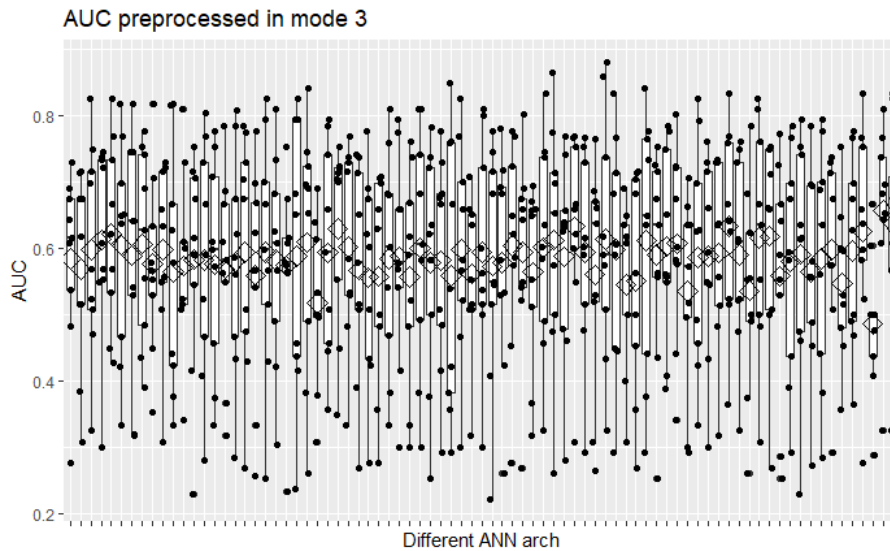


Figura 195. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modo 3.

A156. ANN SIN REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.635, con ningún grupo en torno a 0.35.

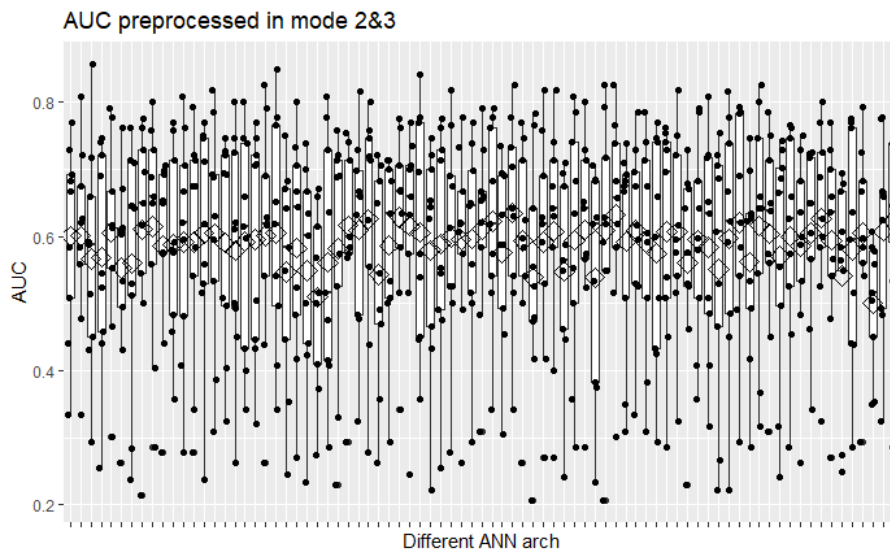


Figura 196. AUC de ANN de tres capas, sin reducción de la dimensión, escalado modos 2&3.

A157. ANN CON REDUCCIÓN DE LA DIMENSIÓN UNA CAPA AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.603, con ninguno de los grupos por debajo de 0.5.

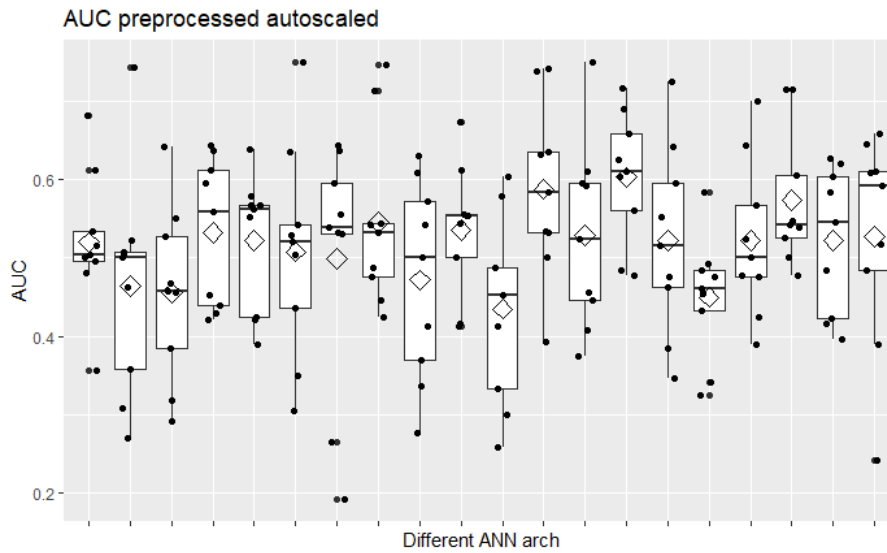


Figura 197. AUC de ANN de una capa, con reducción de la dimensión, autoescalado.

A158. ANN CON REDUCCIÓN DE LA DIMENSIÓN UNA CAPA CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.658, con ningún grupo por debajo de 0.5.

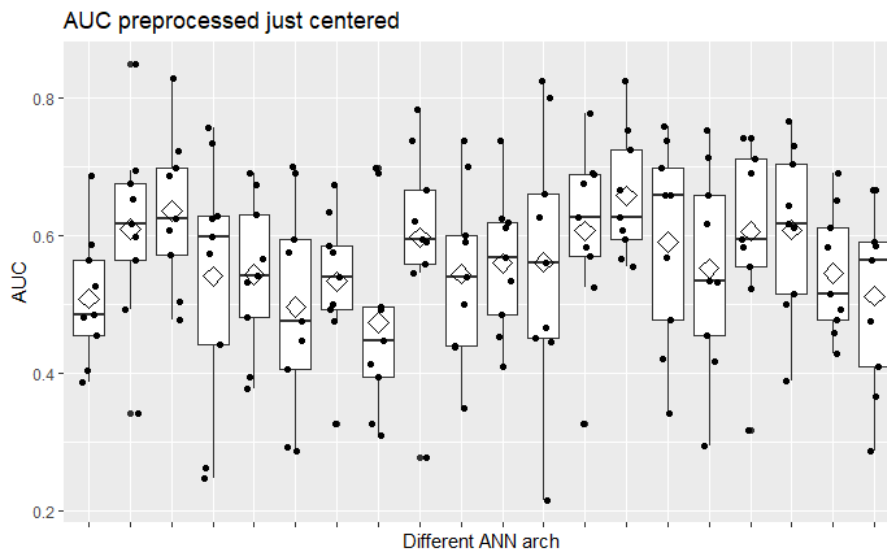


Figura 198. AUC de ANN de una capa, con reducción de la dimensión, centrado.

A159. ANN CON REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.620, con ningún grupo por debajo de 0.5.

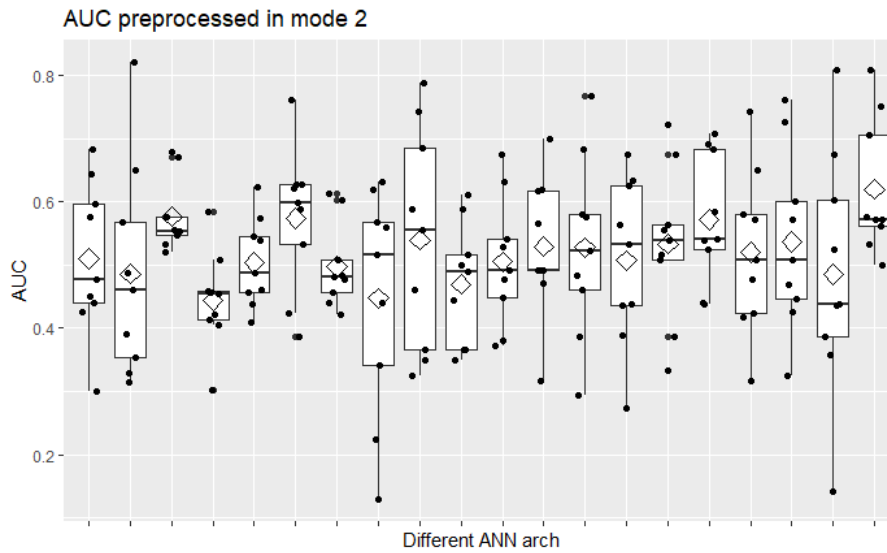


Figura 199. AUC de ANN de una capa, con reducción de la dimensión, escalado modo 2.

A160. ANN CON REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.593, con un grupo por debajo de 0.5.

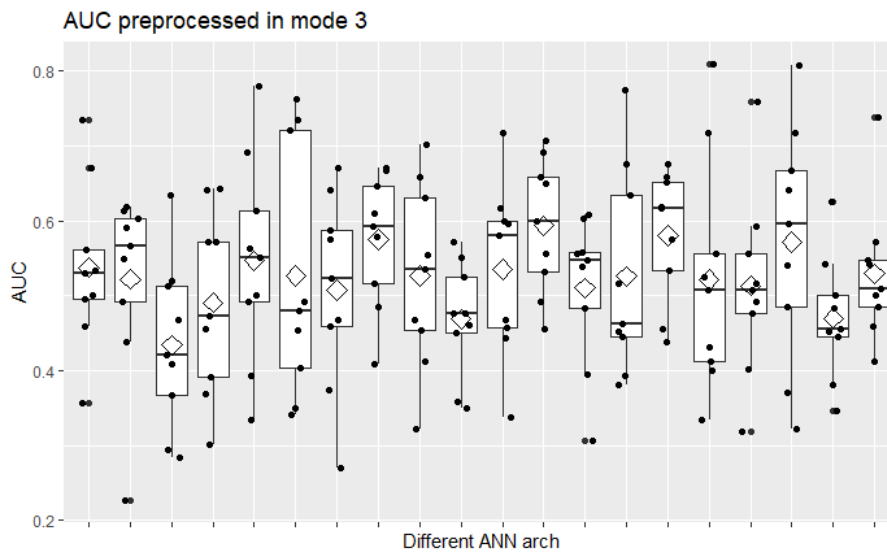


Figura 200. AUC de ANN de una capa, con reducción de la dimensión, escalado modo 3.

A161. ANN CON REDUCCIÓN DE LA DIMENSIÓN UNA CAPA MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.581, con dos grupos alrededor de 0.4.

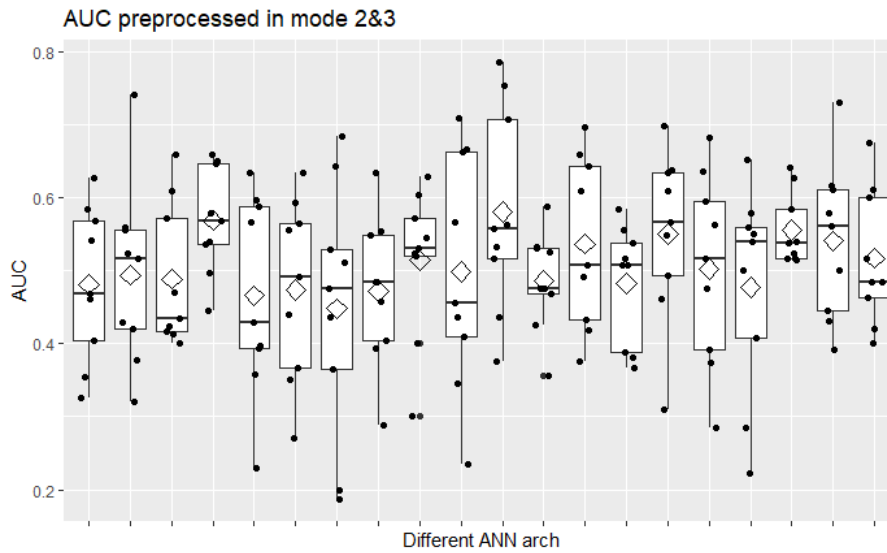


Figura 201. AUC de ANN de una capa, con reducción de la dimensión, escalado modos 2&3.

A162. ANN CON REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.594, con dos grupos por debajo de 0.5.

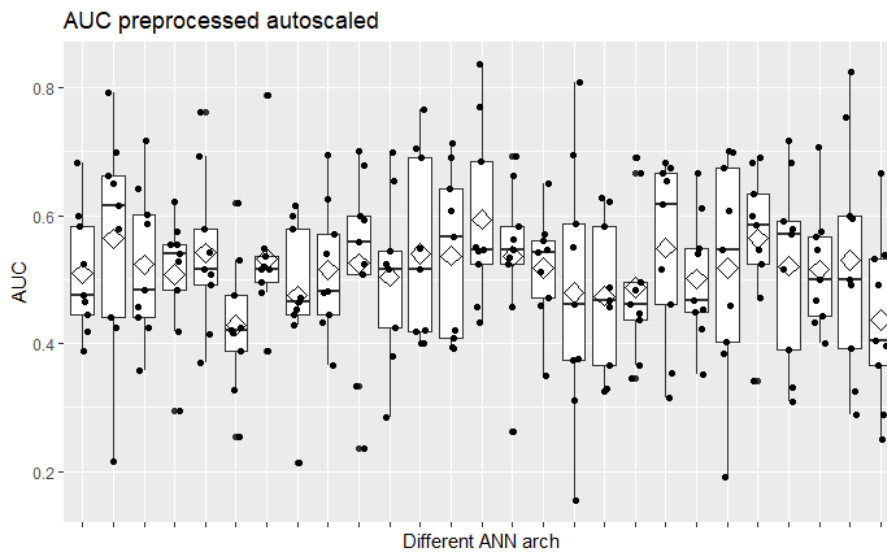


Figura 202. AUC de ANN de dos capas, con reducción de la dimensión, autoescalado.

A163. ANN CON REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.650, con un grupo en torno a 0.25.

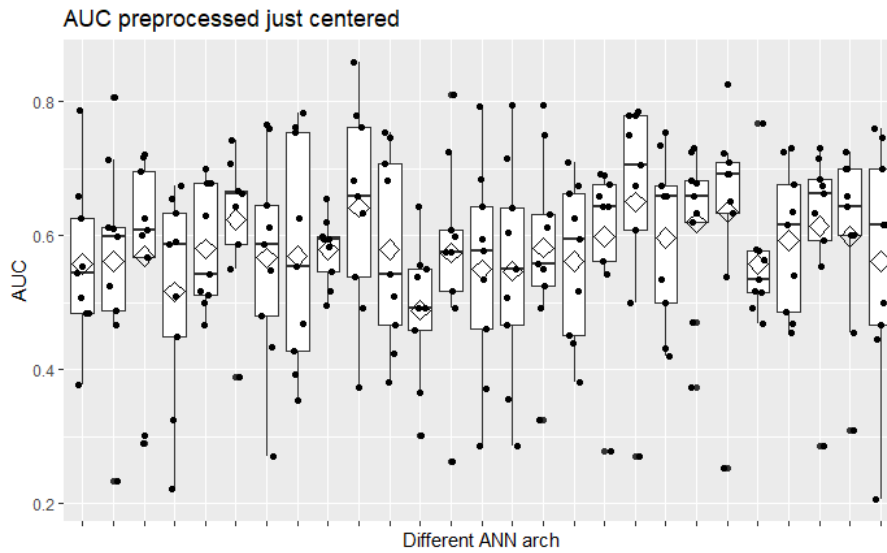


Figura 203. AUC de ANN de dos capas, con reducción de la dimensión, centrado.

A164. ANN CON REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.576, con tres grupos en torno a 0.4.



Figura 204. AUC de ANN de dos capas, con reducción de la dimensión, escalado modo 2.

A165. ANN CON REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.605, con un grupo en torno a 0.4.

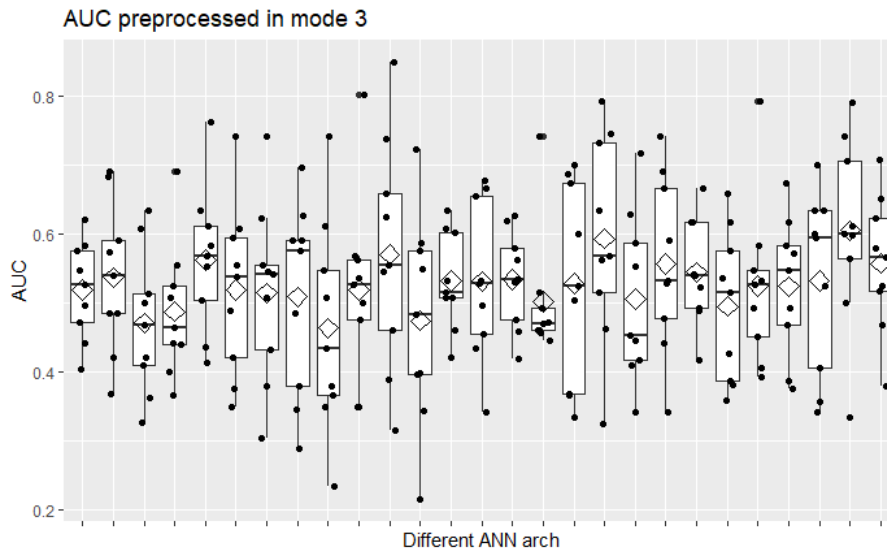


Figura 205. AUC de ANN de dos capas, con reducción de la dimensión, escalado modo 3.

A166. ANN CON REDUCCIÓN DE LA DIMENSIÓN DOS CAPAS MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.564, con un grupo en torno a 0.35.

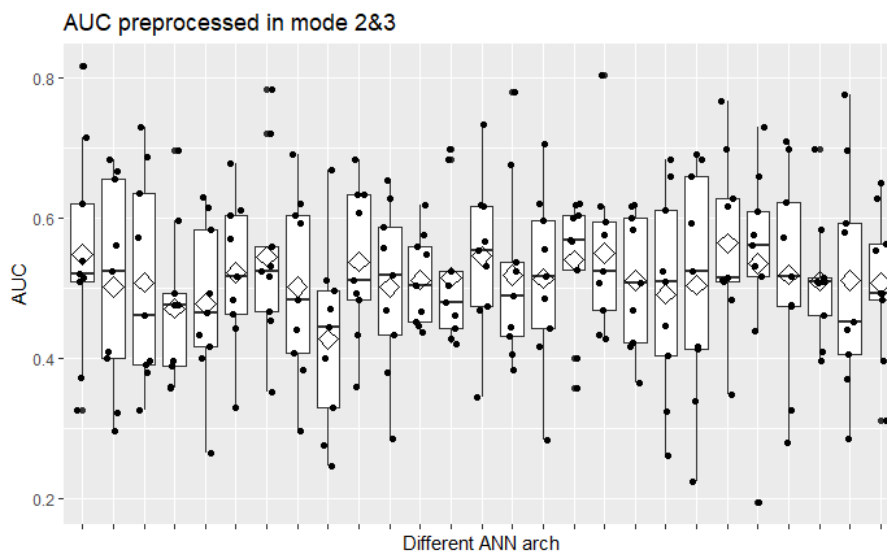


Figura 206. AUC de ANN de dos capas, con reducción de la dimensión, escalado modos 2 y 3.

A167. ANN CON REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS AUTOESCALADO

Con autoescalado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.601, con un grupo en torno a 0.4.

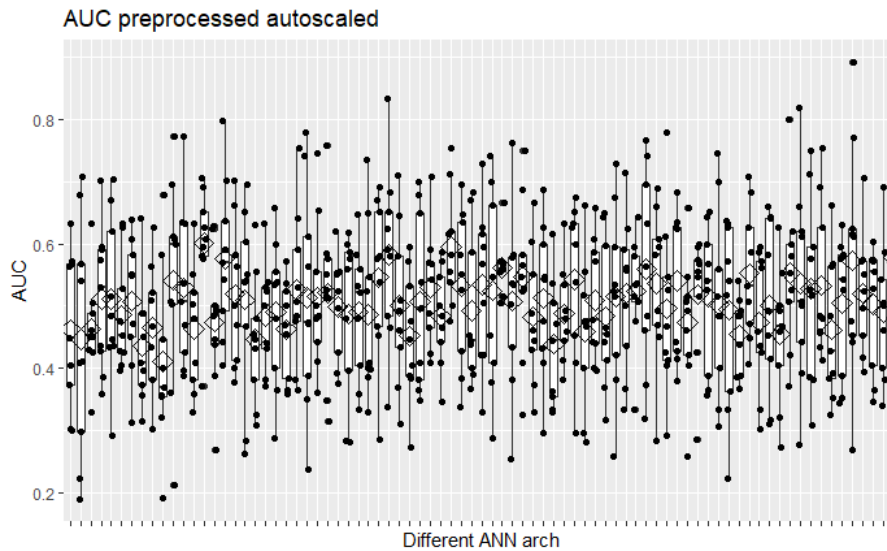


Figura 207. AUC de ANN de tres capas, con reducción de la dimensión, autoescalado.

A168. ANN CON REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS CENTRADO

Con sólo centrado, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.619, con ningún grupo por debajo de 0.5.

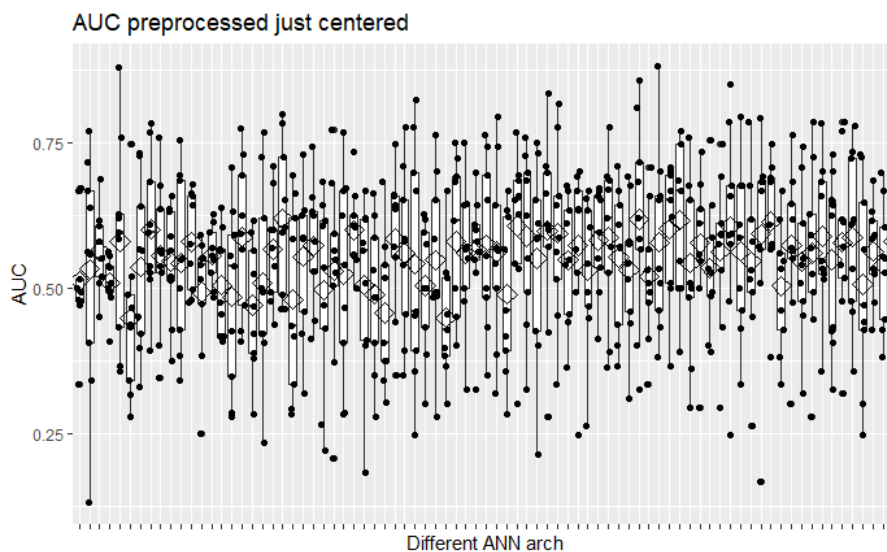


Figura 208. AUC de ANN de tres capas, con reducción de la dimensión, centrado.

A169. ANN CON REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODO 2

Con escalado en el modo 2, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.610, con un grupo ligeramente por debajo de 0.5.

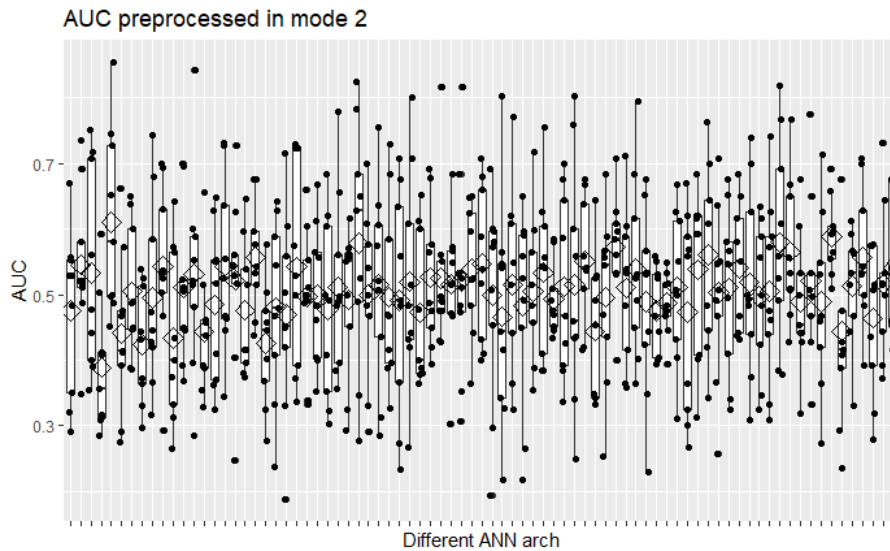


Figura 209. AUC de ANN de tres capas, con reducción de la dimensión, escalado modo 2.

A170. ANN CON REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODO 3

Con escalado en el modo 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.584, con un grupo en torno a 0.4.



Figura 210. AUC de ANN de tres capas, con reducción de la dimensión, escalado modo 3.

A171. ANN CON REDUCCIÓN DE LA DIMENSIÓN TRES CAPAS MODOS 2 Y 3

Con escalado en los modos 2 y 3, el mejor modelo obtiene un AUC medio mediante CV a 9 grupos balanceados de 0.628, con ningún grupo por debajo de 0.5.

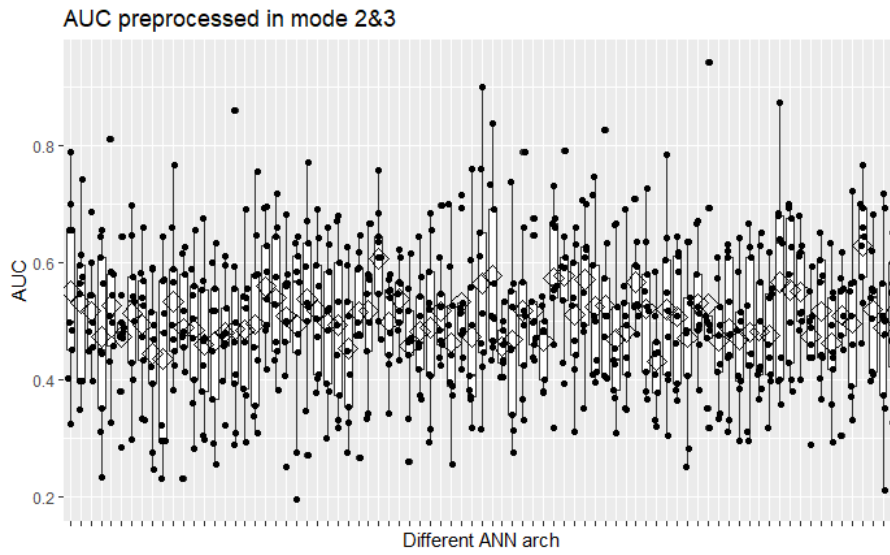


Figura 211. AUC de ANN de tres capas, con reducción de la dimensión, escalado modos 2&3.