UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

TRABAJO DE FIN DE GRADO

USO DEL ANÁLISIS DISCRIMINANTE CUADRÁTICO EN LA MINERÍA DE DATOS CREACIÓN DE UNA APLICACIÓN WEB CON R PARA CLASIFICAR O PREDECIR DATOS REALES.

Tutor: MIGUEL RODRÍGUEZ ROSA

Autora: JUDIT SAN SEGUNDO LOZANO

Grado en Estadística

Curso académico 2023-24

UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

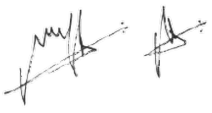
VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO DE FIN DE GRADO

USO DEL ANÁLISIS DISCRIMINANTE CUADRÁTICO EN LA MINERÍA DE DATOS CREACIÓN DE UNA APLICACIÓN WEB CON R PARA CLASIFICAR O PREDECIR DATOS REALES.

Firmado por:



Autora: Judit San Segundo Lozano

Tutor: Miguel Rodríguez Rosa

Grado en Estadística

Curso académico 2023-24

Certificado del tutor TFG Grado en Estadística

D. Miguel Rodríguez Rosa, profesor del Departamento de Estadística de la Universidad de Salamanca,

HACE CONSTAR:

Que el trabajo titulado "*Uso del análisis discriminante cuadrático en la minería de datos. Creación de una aplicación web con R para clasificar o predecir datos reales*", que se presenta, ha sido realizado por D.ª Judit San Segundo Lozano, con DNI 71042782Z y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, a 4 de julio de 2024.

RODRIGUEZ ROSA MIGUEL - 70821913Z

Firmado digitalmente por RODRIGUEZ ROSA MIGUEL - 70821913Z
Fecha: 2024.07.04 09:42:19 +02'00'

Fdo.: Miguel Rodríguez Rosa

Índice general

Summary **9**

Resumen **11**

Introducción **13**

Objetivos **19**

1. Desarrollo Estadístico **21**
1.1. Análisis Discriminante 21
1.1.1. Discriminante Lineal 22
1.1.2. Discriminante cuadrático 26
1.1.3. Comparación LDA y QDA 28
1.2. Método biplot. 28

2. Datos Reales **31**
2.1. Obtención de la base de datos 31
2.2. Descripción de la base de datos 32

3. Software **35**
3.1. Exploración del entorno y lenguaje de R 35
3.2. Desarrollo del software 36
3.3. Shiny 37
3.3.1. Funcionamiento de Shiny 37
3.3.2. Manual de la interfaz de la App 40
3.4. Código de R 43
3.4.1. Funcion QDA 44
3.4.2. Predicción 46
3.4.3. Interfaz 46
3.5. Algoritmo 47
3.6. Representación 48
3.6.1. Representación QDA 48
3.6.2. Representación Biplot 48

4. Resultados **51**

5. Conclusiones **61**

Bibliografía **63**

Anexo **65**

Summary

In today's digital era, the ability to generate and store large amounts of data has transformed numerous fields, including medicine. This work focuses on the development of a web application using RStudio, aimed at predicting and diagnosing ichthyosis through quadratic discriminant analysis (QDA). The main motivation for this project is to provide a useful tool for clinicians to facilitate the diagnostic process of this genetic skin disease.

Quadratic discriminant analysis, an advanced statistical technique that handles classes with different covariance matrices, is described in detail. This is crucial for obtaining accurate predictions in complex data, such as the microbiome data associated with ichthyosis. Unlike linear discriminant analysis (LDA), QDA better captures the variability and intrinsic characteristics of biomedical data.

The web application was developed with Shiny, an R package that facilitates the creation of interactive applications. This application allows users to load new datasets and obtain immediate predictions about the presence of ichthyosis, with an intuitive interface designed for accessibility to all types of users.

Various QDA models were implemented and evaluated on a database of 34 patients under 18 years old, with samples from healthy and affected skin regions. Although the results did not achieve optimal accuracy in all tests, they demonstrated the effectiveness of QDA in classifying different diagnoses based on microbiome variables. These results validate the feasibility of using machine learning techniques and advanced statistical analysis in the diagnosis of rare diseases.

In conclusion, this work has successfully integrated a practical and robust tool for the diagnosis of ichthyosis, significantly contributing to the fields of bioinformatics and medicine. The developed web application not only improves the accuracy of clinical diagnosis but also provides a framework for future research and applications in biomedical data analysis. This project highlights the importance of emerging technologies in improving medical care and disease diagnosis, opening new possibilities for the use of advanced statistical techniques in daily clinical practice.

Resumen

En la era digital actual, la capacidad de generar y almacenar grandes cantidades de datos ha transformado numerosos campos, incluida la medicina. Este trabajo se centra en el desarrollo de una aplicación web utilizando RStudio, destinada a predecir y diagnosticar la ictiosis mediante el análisis discriminante cuadrático (QDA). La motivación principal de este proyecto radica en proporcionar una herramienta útil para los clínicos, que facilite el proceso de diagnóstico de esta enfermedad genética de la piel.

Se ha descrito detalladamente el análisis discriminante cuadrático, una técnica estadística avanzada que permite manejar clases con diferentes matrices de covarianza, crucial para obtener predicciones precisas en datos complejos como los del microbioma asociado a la ictiosis. A diferencia del análisis discriminante lineal (LDA), el QDA captura mejor la variabilidad y las características intrínsecas de los datos biomédicos.

El desarrollo de la aplicación web se llevó a cabo con Shiny, un paquete de R que facilita la creación de aplicaciones interactivas. Esta aplicación permite a los usuarios cargar nuevos conjuntos de datos y obtener predicciones inmediatas sobre la presencia de ictiosis, con una interfaz intuitiva diseñada para la accesibilidad a todo tipo de usuarios.

Se implementaron y evaluaron diferentes modelos de QDA aplicados a una base de datos de 34 pacientes menores de 18 años, con muestras de regiones sanas y afectadas de la piel. Aunque los resultados no alcanzaron una precisión óptima en todas las pruebas, demostraron la eficacia del QDA para clasificar diferentes diagnósticos basados en variables del microbioma. Estos resultados validan la viabilidad de utilizar técnicas de aprendizaje automático y análisis estadístico avanzado en el diagnóstico de enfermedades raras.

En conclusión, este trabajo ha integrado exitosamente una herramienta práctica y robusta para el diagnóstico de ictiosis, contribuyendo significativamente a los campos de la bioinformática y la medicina. La aplicación web desarrollada no solo mejora la precisión del diagnóstico clínico, sino que también proporciona un marco para futuras investigaciones y aplicaciones en el análisis de datos biomédicos. Este proyecto destaca la importancia de las tecnologías emergentes en la mejora de la atención médica y el diagnóstico de enfermedades, abriendo nuevas posibilidades para la utilización de técnicas estadísticas avanzadas en la práctica clínica diaria.

Introducción

La era digital ha transformado la manera en que se procesa, se almacena, se distribuye y se transmite la información digitalizada. Gracias al constante avance en informática y de las tecnologías, se genera y almacena una vasta cantidad de datos. El desafío radica en descubrir y aprovechar el conocimiento contenido en este enorme volumen de datos. Avances en la tecnología de gestión de bases de datos permiten ahora la integración fluida de diferentes tipos de datos, como imágenes, videos, textos y datos numéricos, facilitando así el procesamiento integral de información multimedia.

El término de Machine Learning es bastante nuevo, sin embargo, una gran parte de los conceptos que se conocen hoy en día fueron desarrollados hace bastante tiempo, entre estos conceptos tenemos por ejemplo, el método de mínimos cuadrados que se desarrolló a principios del siglo XIX (Cutro, 2010). Este se implementó de tal forma que hoy en día se puede conocer como regresión lineal. La primera vez que se aplicó este método fue en el ámbito de la astronomía.

La regresión lineal se aplica para predecir variables cualitativas, como, por ejemplo, para predecir si un paciente puede sobrevivir o no a dicha enfermedad o si en el mercado de la bolsa aumentan o disminuyen los valores.

En el año 1936 se propuso el método del discriminante lineal, y, en 1940, diferentes autores plantearon la técnica de la regresión logística (James et al., 2017). A principios de la década de 1970, se popularizó la regresión lineal, y de esta manera se desarrollaron diferentes métodos del aprendizaje estadístico. Todos estos métodos se basan en la regresión lineal y en la regresión logística, siendo estos los métodos del aprendizaje estadístico por excelencia.

A finales de esta década, ya se habían creado diversas técnicas para poder analizar diferentes tipos de datos, sin embargo, una gran parte de estos métodos eran lineales, ya que, computacionalmente, era complicado en esa época hacer métodos no lineales.

A mediados de la década de los 80 ya había más facilidades computacionalmente hablando, por lo tanto, ya se podían realizar otros métodos que no fuesen lineales. Durante estos años se desarrollaron algunos métodos tales como los árboles de decisión, la regresión, además de modelos generativos generalizados. En esta década, ganó importancia el método de redes neuronales. En los 90 se llevó a cabo el método estadístico de máquinas de vector soporte. A partir de entonces, el aprendizaje estadístico se considera como un subcampo de la estadística. Este se centra en la predicción de modelos supervisados y no supervisados. Este subcampo ha crecido exponencialmente en los últimos años, gracias a la disponibilidad de diversos programas como por ejemplo R. Actualmente, existen una gran cantidad de métodos para poder predecir diferentes tipos de datos, además de diversos programas que permiten desarrollar los algoritmos de estos métodos de Machine Learning.

En cuanto a efectos prácticos, este análisis es utilizado por las empresas para analizar conjuntos de datos, lo que permite a dichas empresas desarrollar soluciones competitivas además de innovadoras, que afectan a la publicidad, a las predicciones y a otros muchos aspectos estratégicos, los

cuales los consumidores lo verifican.

También se utiliza para identificar las tendencias de mercado con anterioridad, es decir, cuáles serán aquellos productos más novedosos que demandarán los consumidores.

En recursos humanos, se utiliza este método para realizar una selección de los individuos que han aplicado a un trabajo determinado. También se aplica en el ámbito médico para las pruebas diagnósticas para poder distinguir pacientes patológicos y normales, a partir de los resultados de los parámetros hemodinámicos.

En este caso el estudio se centrará en el ámbito de la medicina, concretamente para diagnosticar a pacientes con ictiosis, el tipo concreto que tiene, y para ello se va a usar el método del análisis discriminate cuadrático. Esta enfermedad, la ictiosis, se define como un conjunto de desórdenes dermatológicos caracterizados por tener piel escamosa y eritematosa (roja) (Gutiérrez-Cerrajero et al., 2023). Esta afección suele venir dada por una alteración funcional de la barrera epidérmica de la piel, la cual se localiza en la parte más superficial de la epidermis. La función de esta zona es proteger al organismo frente a agresiones químicas y entrada de patógenos, además de restringir la absorción de líquidos y solutos y restringir la pérdida de líquidos mediante esta zona.

La barrera epidérmica está compuesta de corneocitos, estos se generan por división de los queratinocitos en la capa más profunda de la epidermis. Estas células van experimentando un proceso de diferenciación, de esta manera van adquiriendo distintos fenotipos según van migrando a la superficie de la piel. Estos cambios del fenotipo son los que generan las diferentes capas epidérmicas. Estas son el estrato basal, el estrato espinoso, estrato granuloso, estrato lúcido y estrato córneo.

En el estrato córneo se encuentran células sin núcleo, producto de la diferenciación terminal de queratinocitos a corneocitos. El grosor de esta capa varía dependiendo del lugar anatómico. El estrato lúcido se encuentra en pieles engrosadas, como pueden ser las plantas de los pies y palmas de las manos. El estrato granuloso está constituido por células romboidales que contienen gránulos de queratohialina, como su propio nombre indica. El estrato espinoso está formado por células poligonales en el que se forman unas estructuras que forman nexos entre este tipo de células, estas uniones se denominan uniones intercelulares. Las células de la última capa denominada estrato basal, está constituida por células madre que tienen forma cilíndrica.

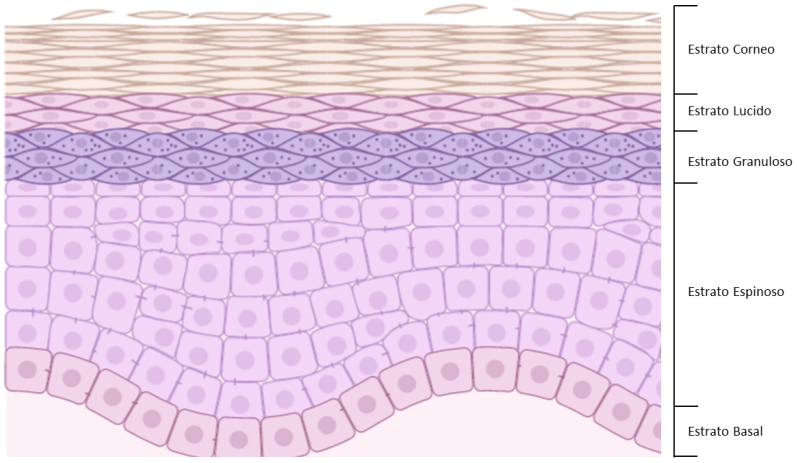


Figura 1: Capas de la piel

Según la causa de la alteración epidérmica, las ictiosis pueden clasificarse en adquiridas o hereditarias. Las ictiosis adquiridas se suelen caracterizar por una aparición tardía ya que estas aparecen mediante factores ambientales, como pueden ser medicamentos, infecciones o tumores, mientras que las ictiosis hereditarias están causadas por mutaciones durante el desarrollo embrionario. Este último tipo se divide en más de 50 subtipos de ictiosis, que, a su vez, se subdividen en ictiosis comunes, ictiosis congénitas autosómicas recesivas, ictiosis queratinopáticas e ictiosis sindrómicas.

Ictiosis comunes, las de mayor incidencia en la población, son:

- Ictiosis vulgar: se caracteriza por la aparición a los pocos meses o años de vida y por las escamas marrones claras. Las zonas más frecuentemente afectadas son las piernas y las palmas hiperlineales, soliendo no afectar a las regiones antecubital, poplíteax y a la cara. *FLG* es el gen más comunmente alterado, sin embargo, hay otras formas de ictiosis autosómicas recesivas y raras que afectan a la expresión de filagrina. *CASP14*, es un gen que causa escamas blancas y finas sin membrana colodión al nacer. *ASPRV1* se parece a la ictiosis laminar (ver debajo), pero este tipo tiene las líneas de palmas de las manos muy marcadas, además de no tener membrana colodión, la cual es una membrana brillante que suelen tener los bebés con algunos tipos de ictiosis cuando nacen, y se desprende a las pocas semanas de vida.
- Ictiosis recesiva ligada al X: Estos pacientes presentan escamas poligonales marrones oscuras o grises, las cuales están firmemente adheridas en la piel. No se suelen presentar en zonas como por ejemplo la cara o las palmas de las manos o plantas de los pies, al contrario que el cuello y el cuero cabelludo ya que estas suelen ser las zonas más afectadas. Este tipo de ictiosis afecta prácticamente exclusivamente a los varones.

Ictiosis congénitas autosómicas recesivas (ICAR): presentan síntomas desde el nacimiento, a menudo como bebé colodión, para después transicionar a un fenotipo más específico, todas tienen un patrón de herencia autosómico recesivo

- Ictiosis laminar (IL): Se caracteriza por presentar grandes escamas poligonales. El color de este tipo varía a distintas tonalidades de marrón. Afecta a las zonas como la parte interior de las rodillas o de los codos. Los genes más comunmente alterados son *TGM1*, *ALOX12B*, y *CYP4F22*
- Eritrodermia ictiosiforme congénita (EIC): Los pacientes suelen nacer con una afectación menos severa, sin embargo, a medida que crecen se transicionan a síntomas como escamas blancas generalizadas, además de tener la piel roja e inflamada. Suele tener una mayor afectación, al igual que el caso anterior, en las zonas flexurales del cuerpo. Los genes más comunmente alterados son *TGM1*, *ALOX12B*, *PNPLA1* y *CYP4F22*
- Ictiosis Arlequín (IA): Los individuos nacen ya con la piel rígida, ya que la tiene más gruesa de lo normal, esto conlleva una limitación del movimiento, en esto se incluye respirar, chupar, y además presentan fisuras cutáneas profundas. Se caracteriza por presentar eversión de los parpados y de los labios.Esta enfermedad puede ser letal y los pacientes que sobreviven suelen desarrollar una forma grave de Eritrodermia ictiosiforme congénita (EIC). Asociado a mitaciones en *ABCA12*

Ictiosis queratinopáticas (KPI): mutaciones afectan a queratinas (*KRT1* y *KRT10*)

- Ictiosis epidermolítica autosómica dominante (IE): se caracteriza por tener ampollas al nacer, esto hace que se confunda con epidermólisis bullosa, con el paso del tiempo, los pacientes desarrollan hiperqueratosis con abundantes escamas en las articulaciones además de tener eritrodermia, la cual varía la intensidad dependiendo del paciente. Las causadas por alteraciones en *KRT10* no suelen afectar a la piel palmoplantar por compensación del producto génico de esta con el de *KRT9*. Las causadas por *KRT1* suele presentar con queratodermia palmoplantar (condición en la que la piel de las palmas y plantas es más gruesa y áspera de

lo normal) grave.

- Ictiosis con confeti (IWC): destaca por la eritrodermia y el engrosamiento cutáneo, además de tener la aparición de puntos de piel sana la cual aumentan la cantidad y el tamaño con la edad. Estos puntos en la piel son causados por recombinación mitótica durante la división celular que causa reversión del alelo mutado. Este se suele equivocar con EIC, hasta que aparecen las reversiones.

Ictiosis sindrómicas: la mutación causa síntomas fuera de la piel

- Síndrome de Netherton (SN): se caracteriza por eritrodermia congénita, escamas y deshidratación prematura. En las edades más avanzadas, el síntoma más común de los pacientes es la eritrodermia con un patrón característico de escamas llamado ictiosis lineal circunfleja. En este, las lesiones se encuentran rodeadas por un borde de escamas, también se caracteriza por tener pelo en bambú. Causado por variantes en *SPINK5*
- Síndrome de queratitis-ictiosis-sordera (KID): el síntoma principal es la inflamación corneal, hiperqueratosis espinosa (a nivel microscópico) con queratodermia plamoplantar (engrosamiento de la piel, especialmente en manos y pies) y perdida de la audición. *GJB2* es el principal gen asociado a lesiones queratóticas.
- Síndrome de dermatitis severa-alergias múltiples y desgaste metabólico (SAM): los síntomas más comunes son eritrodermia con descamación superficial, hipotricosis, infecciones recurrentes y múltiples alergias alimenticias que pueden causar retraso en el desarrollo del paciente. Está causado por mutaciones en *DSG1* y *DSP*

En cuanto al tratamiento de la ictiosis, no existe tratamiento curativo, dado que se trata de una enfermedad genética, sin embargo, hay terapias disponibles, las cuáles se necesitarán de por vida y solo ofrece alivio sintomático. Esta terapia usa productos con aceites para poder hidratar la piel, ya que esta enfermedad se caracteriza por perder agua a través de esta. La terapia requiere mucho tiempo y su efectividad varía mucho entre pacientes. Existe otro método, utilizado sobre todo en las ictiosis que producen hiperqueratosis pronunciada (mayor engrosamiento), se trata de la terapia sistémica con retinoides orales, pero al igual que todos los medicamentos tienen efectos adversos, como puede ser un aumento de la fragilidad de la piel, además de otros efectos secundarios.También se emplean terapias dirigidas que ayudan a reducir la inflamación, y se están desarrollando terapias de reemplazo de lípidos y proteínas, así como estrategias de terapia génica. Estas técnicas son prometedoras, aunque todavía requieren validación. Muchos pacientes necesitan tratamiento diario y crónico, por lo que los tratamientos deben ser tanto seguros como bien tolerados por el paciente. Actualmente, existen pocos estudios sobre los perfiles de seguridad a largo plazo de estos tratamientos.

Dejando de lado el tipo de ictiosis, una gran parte de los pacientes sufren muchos efectos secundarios de dicha enfermedad. Estos efectos pueden ser prurito, infecciones recurrentes, hipohidrosis con intolerancia al calor y diversas complicaciones oculares, auditivas y nutricionales, las cuáles se deben controlar de forma continua.

- El prurito se manifiesta en el 93 % de los individuos que padecen ictiosis. Esta afección se refiere a la picazón que sufren los pacientes en las zonas que tienen afectadas debido a la enfermedad. Se ve afectado en entornos más cálidos o en situaciones de estrés, ya que esto puede intensificar el prurito. Esta afección puede causar otras, tales como pueden ser el rascado, el insomnio o el mal humor. El prurito en ictiosis no se ha estudiado sistemáticamente, pero se cree que puede estar relacionado con la inflamación cutánea. La terapia con biológicos como dupilumab, ustekimumab y secukinumab, parece eficaces frente al prurito.
- La hipohidrosis afecta incluso a los pacientes más leves, ya que el no poder refrigerar el cuerpo mediante la sudoración, aumenta el riesgo de sufrir golpes de calor, por lo que deben

tener un mayor cuidado, especialmente en los meses más calurosos. Lo más recomendable es que los pacientes permanezcan en lugares frescos, con ventiladores o con cualquier otro sistema de refrigeración. A pesar de esto, existen algunos tipos de ictiosis que mejoran con el sol.

- Uno de los síntomas más comunes en todos los tipos de ictiosis son las complicaciones oculares, con más exactitud, la descamación y tirantez en los parpados, la conjuntivitis y las anomalías en las pestañas. Se recomienda tratar estos síntomas mediante la lubricación ocular profiláctica de manera frecuente. Las inyecciones de ácido hialurónico, también ayudan a tratar el ectropión (eversión de los párpados) de estos pacientes. Existe una técnica quirúrgica para el ectropión, pero no suele ser muy efectiva, y en caso de serlo, las reapariciones suelen ser bastante frecuentes.
- Las complicaciones auditivas también son bastante comunes, dado que la descamación en la piel es frecuente, el canal auditivo también se ve afectado, esto causa bloqueo en el oído, por lo que los pacientes pueden llegar a perder audición además de sufrir infecciones recurrentes del oído externo y medio, los cuales pueden llegar a dañar el tímpano permanentemente. La pérdida auditiva puede afectar al desarrollo cognitivo, por ello, es necesaria la temprana intervención, además de un seguimiento regular por un otorrinolaringólogo.
- Las infecciones cutáneas recurrentes son debidas a los defectos en la barrera epidérmica, esto causa distintos tipos de infecciones tanto bacteriana, viral y fúngica. A pesar de que no hay muchos estudios que estudien el microbioma de la piel afectada (el conjunto de bacterias presentes en el tejido), se suele caracterizar por un olor especialmente característico, además suele ser bastante desagradable en los pacientes que padecen hiperqueratosis prominente. Algunos tipos de ictiosis son más propensos a sufrir infecciones, sin embargo, pueden llegar a pasar desapercibidas debido a la piel ictiótica, estos tipos son IE, SN y el síndrome KID. La sepsis también es especialmente frecuente en los bebes que padecen SN, y se debe a los defectos que presentan en la barrera cutánea, esto puede conllevar a poner en riesgo de muerte a los pacientes.

La calidad de vida de los individuos que padecen ictiosis varía dependiendo del tipo de ictiosis que se tenga. En términos generales, los tipos más graves poseen una peor calidad de vida, mientras que los más leves, desarrollan una mejor vida. Los síntomas que más empeoran la calidad de vida de los individuos son los siguientes; el dolor cutáneo, prurito y las escamas. Un estudio multivariante desveló que las puntuaciones obtenidas en el Índice de Calidad de Vida en Dermatología (DLQI) son más altas en las mujeres, por lo que estas tienen una peor calidad de vida, aunque no se han estudiado en profundidad las diferencias en el género, pueden deberse a los estándares de belleza de la sociedad.

Una vez definidos los términos clave de este proyecto, analizaremos su estructura en detalle. El trabajo se dividirá en varios capítulos, cada uno enfocado en aspectos específicos de la metodología y los resultados obtenidos. A continuación, se proporciona una descripción más detallada de cada capítulo.

Tras un primer apartado con los objetivos del trabajo, el primer capítulo se centrará en el desarrollo estadístico del análisis discriminante cuadrático.

En el segundo capítulo, se desarrollará cómo se han obtenido los datos reales, es decir las bases de datos utilizadas en el estudio.

En el siguiente capítulo, se abordará el software. Aquí se explicará todo el código realizado en RStudio.

En el capítulo final se presentará los resultados de los análisis descritos en los captitulos anteriores.

Por último, se ofrecerá la bibliografía utilizada para la realización de este trabajo, en esta encontraremos diversos libros, artículos, y páginas web.

Además, se ofrecerá una sección de anexos que contendrá el código con el que se ha realizado la app además del código de los gráficos para la visualización del análisis.

Objetivos

El principal objetivo de este trabajo es crear una aplicacion web con RStudio para poder predecir la base de datos de Ictiosis y ayudar a los clínicos a la hora de diagnosticar esta enfermedad.

En este trabajo tendremos los siguientes objetivos:

- Describir matemáticamente el análisis discriminante cuadrático.
- Clasificar la base de datos del microbioma mediante el método QDA.
- Aplicar los algoritmos de R a la base de datos de ictiosis para ayudar con el diagnóstico de esta enfermedad.

Capítulo 1

Desarrollo Estadístico

1.1. Análisis Discriminante

El objetivo del análisis discriminante consiste en optimizar de la mejor forma posible la información de las variables independientes de manera que se pueda crear una función que sea combinación lineal de X_1 y X_2 (Marin, 2016). Estas variables representan el espacio bivariante de los puntos correspondientes a dos grupos hipotéticos. Dicha función tiene que ser capaz de distinguir lo máximo posible a ambos grupos. Esta sería la siguiente:

$$D = b_1X_1 + b_2X_2$$

b_1 y b_2 se refieren a las ponderaciones de las variables independientes, estas son las que hacen que los individuos de cada grupo tengan los resultados más altos posibles en la ecuación.

La función discriminante se trata de una función unidimensional. Cuando $p > 2$ resulta más complicado su representación, y añade poca información a la interpretación de la ecuación. En la figura de abajo (Figura 1.1) se representa la función discriminante D extraída del espacio de las variables X_1 y X_2. La representación de esta función es la siguiente:

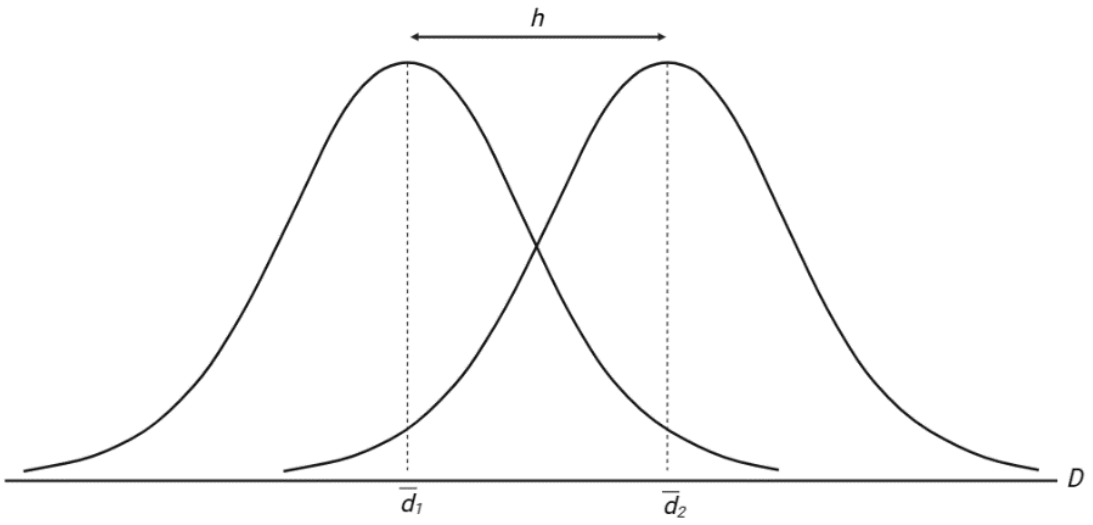


Figura 1.1: Función Discriminante

Tras obtener la función discriminante, se sustituye el valor de las medias del grupo 1 en las variables X_1 y X_2. Con estos cálculos se obtienen los centroides del grupo 1.

$$\overline{d}_1 = b_1\overline{x}_1^{(1)} + b_2\overline{x}_2^{(1)}$$

De la misma forma, se sustituye en el grupo 2, por lo tanto, se obtiene el centroide de este grupo.

$$\overline{d}_2 = b_1\overline{x}_1^{(2)} + b_2\overline{x}_2^{(2)}$$

La distancia entre los centroides tiene que máxima, de esta manera la separación entre los grupos sea lo más distante posible. La ecuación para expresar la distancia entre los centroides es la siguiente:

$$h = \overline{d}_1 - \overline{d}_2$$

donde $\overline{d}_1$ es la media del grupo 1, y $\overline{d}_2$ es la media del grupo 2.

El objetivo es reducir a una sola dimensión, la dimensionalidad de las variables independientes. En esta dimensión los grupos se tienen que diferenciar lo máximo posible, dado que, las puntuaciones obtenidas de esta nueva dimensión son las que se van a utilizar para llevar a cabo la clasificación de los individuos.

1.1.1. Discriminante Lineal

Se trata de un método supervisado para variables cualitativas, en este se tiene que conocer al menos dos de los grupos, y el resto de las variables se clasifican en medida a estos grupos según las características de estos individuos (Amat-Rodrigo, 2016).

En el análisis discriminante lineal, el número de variables debe ser el número de objetos menos dos ($p < (n - 2)$), donde n es el número de objetos. Ninguna de las variables discriminante puede ser combinación lineal de las otras.

Para poder realizar el análisis discriminante lineal, se tienen que cumplir los siguientes requisitos:

- En un modelo de clasificación, se asume que cada variable predictora sigue una distribución normal dentro de cada categoría de la variable objetivo. Cuando se tienen múltiples variables predictoras, se considera que las observaciones siguen una distribución normal multivariante en todas las categorías.
- La versatilidad de cada variable predictora tiene que ser la misma en todas las categorías de la variable objetivo. Cuando hay múltiples de estas predictoras, se asume que la matriz de covarianza, que describe cómo las variables se relacionan entre sí, es igual en todas las categorías.

Sin embargo, si estas suposiciones no se cumplen y la variabilidad y la matriz de covarianza difieren entre las categorías, entonces se recurre al Análisis Discriminante Cuadrático (QDA), para manejar esta mayor complejidad en la relación entre las variables predictoras y la variable objetivo.

Este método es una alternativa a la regresión logística cuando la variable cualitativa tiene más de dos categorías. Aunque existen otras opciones, el análisis discriminante lineal ofrece una serie de ventajas en comparación con la regresión logística.

- Si hay una distinción clara entre los grupos, los parámetros que han sido estimados de la regresión logística pueden ser inestables, sin embargo, el análisis discriminante lineal no

sufre este problema.

- Si no hay una gran cantidad de observaciones y los predictores se distribuyen de forma normal en cada uno de los grupos, el discriminante lineal es más estable que la regresión logística.

La dificultad de este método es realizar el cálculo de la matriz de covarianzas cuando hay más variables predictoras que datos, o cuando hay una fuerte correlación entre las variables predictoras. En estos casos, se suele realizar un análisis de componentes principales (PCA) para poder reducir la dimensión y obtener predictores incorrelacionados y de esta manera realizar el discriminante lineal. Sin embargo, este método se realiza sin saber la información que poseen cada una de las categorías, por lo que una alternativa es realizar el metodo de *partial least squares discriminant analysis* (PLSDA). Este se centra en realizar un análisis de regresión PLS donde las variables de interés actúan como la salida deseada. El propósito es disminuir la complejidad del modelo al mismo tiempo que se potencia la relación con las salidas esperadas. Este enfoque de reducción de dimensionalidad seguido de la clasificación basada en estas nuevas variables es común en diversas áreas de análisis de datos y aprendizaje automático, ya que puede simplificar la interpretación de los datos y mejorar la eficiencia computacional en el análisis y la predicción (Friedman, 1989).

l análisis discriminante lineal se basa en el teorema de Bayes, pero en 1936, Ronald Fisher propuso una técnica que reduce el espacio de características de dimensión p (el número de características originales) a un subespacio de menor dimensión. Este subespacio está compuesto por combinaciones lineales de las características que mejor explican la separación entre las clases. Una vez identificadas estas combinaciones, la clasificación se realiza en este subespacio reducido. Fisher definió el subespacio óptimo como aquel que maximiza la distancia entre grupos en términos de varianza.

Los términos "discriminante lineal de Fisherz "LDA"se utilizan frecuentemente de manera intercambiable para describir la misma idea. No obstante, el enfoque original de Fisher describe un discriminante ligeramente diferente, ya que no asume algunas de las suposiciones del LDA, como la distribución normal de las clases o la igualdad de covarianzas entre las clases. A pesar de estas diferencias, el trabajo de Fisher sentó una base esencial para el desarrollo posterior del LDA y otras técnicas relacionadas en el campo del análisis de datos y la estadística.

La aproximacion de Fisher se divide en dos partes:

- La reducción de dimensionalidad implica la transformación de un conjunto de p variables discriminantes en k combinaciones lineales de estas variables, con el propósito de explicar la separación entre grupos, pero en un espacio de menor dimensión ($k < p$)
- Este subespacio de menor dimensión, es usado para clasificar las observaciones. Esto implica asignar a cada observación un grupo o categoría específica basada en su posición en el espacio de las variables discriminantes.

Aunque Fisher fue el primero que lo describió, no fue hasta unos años mas tarde, 1939, cuando Welch lo enfocó utilizando el Teorema de Bayes, asumiendo que todas las variables tienen la misma matriz, de la cual de esta se obtienen las funciones discriminantes.

Los resultados de la categorización obtenidos mediante el método de Fisher son equivalentes a los del método de Bayes cuando se cumple la siguiente condición:

- En el método de Bayes, se asume que la matriz de covarianzas es constante en todos los grupos y se utiliza como estimación la matriz de covarianza combinada dentro de cada clase. (pooled within-class covariance matrix).
- Todos los discriminantes lineales se emplean para la categorización, según el método de Fisher. El número máximo de discriminantes obtenidos después de la reducción de la di-

mensionalidad es igual al número de grupos menos uno.

Bayes establece que dados dos eventos A y B, la probabilidad de que ocurra B, habiendo ocurrido A, es lo mismo que ocurra simultáneamente A y B entre la probabilidad de que ocurra A (Amat-Rodrigo, 2016).

$$P(B|A) = \frac{P(A,B)}{P(A)}$$

Si se quisiera clasificar una nueva observación en una de las K clases de la variable cualitativa Y, siendo $k \geq 2$, teniendo un solo predictor X, se definen los siguientes conceptos:

- La probabilidad de que una observación aleatoria pertenezca a la clase k se conoce como probabilidad previa (π_k), que también se define como la probabilidad global o probabilidad inicial.
- La función de densidad condicional de X perteneciente a la clase k es la siguiente:

  $$f_k(X) \equiv P(X = x|Y = k)$$

  A mayor $f_k(x)$ mayor es la posibilidad que se obtenga $X \approx x$, teniendo una observación de la clase k.
- La probabilidad de que una observación pertenezca a la clase k, dado un valor del predictor X, se conoce como probabilidad a posterior ($P(Y = k|X = x)$).

El teorema de Bayes aplicado a cada una de las clases sería:

$$P\left(\text{pertencer a la clase } k|\text{valor } x \text{ observado}\right) = \frac{P\left(\text{pertenecer a la clase } k \text{ y observar } x\right)}{P(\text{observar } x)}$$

Introduciendo los términos definidos anteriormente, se obtiene la siguiente ecuación (Greyrat, 2022):

$$P(Y = k|X = x) = \frac{\pi_k P(X = x|Y = k)}{\sum_{j=1}^{K} \pi_j P(X = x|Y = j)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$$

Se logra la clasificación con menor margen de error utilizando el enfoque de clasificación de Bayes, el cual consiste en asignar cada observación al grupo que maximice la probabilidad posterior. Debido a que el denominador, que es la suma de las probabilidades previas para todas las clases, es constante para todas ellas. La regla de clasificación se puede simplificar al asignar cada observación al grupo cuyo producto de la probabilidad previa (π_k) y la función de densidad ($f_k(x)$) sea el más grande.

Para realizar una clasificación basada en el enfoque de Bayes, es crucial disponer de las siguientes probabilidades poblacionales:

- Probabilidades a priori (π_k): Estas reflejan la probabilidad de que una observación aleatoria pertenezca a cada clase K
- Distribución condicional ($f_k(x)$): Esta indica la probabilidad poblacional de que una observación perteneciente a la clase k tenga un valor específico x en la variable predictora.

Sin embargo, en la práctica, rara vez se dispone de esta información de la población en su totalidad. Por lo tanto, es necesario estimar estos parámetros a partir de una muestra de datos. Como resultado, el clasificador del Análisis Discriminante Lineal (LDA) que se obtiene, se acerca al clasificador de Bayes en sus fundamentos teóricos, pero debido a que los parámetros se estiman a

partir de datos muestrales, el LDA resultante no es idéntico al otro clasificador.

Para que el análisis discriminante Lineal pueda clasificar de una manera más optima, depende directamente de π_k y de $f_k(x)$. Cuanto más cercanos estén estos valores al real, más se acercará el análisis discriminante al clasificador de Bayes. La probabilidad previa es más sencilla de estimar. La probabilidad de que una observación cualquiera pertenezca a la clase k se puede calcular como la proporción de observaciones en la clase k con respecto al número total de observaciones.

$$\widehat{\pi}_k = \frac{n_k}{N}.$$

Sin embargo, para calcular $f_k(x)$, no es tan sencillo, en primer lugar, se tiene que considerar que $f_k(x)$ se distribuye de forma normal en todas las clases. Una vez teniendo esto en cuenta, se puede calcular a partir de la ecuación (Tharwat, 2016):

$$f_k(x) = P(Y = k\,|X = x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2}$$

ya que, dado un valor determinado de las variables predictoras, estima la probabilidad de que una observación pertenezca a alguno de los grupos de la variable cualitativa, $P(Y = k\,|X = x)$. Se asigna la observación a la clase k para la probabilidad predicha es mayor, sin embargo, en este caso se hace para intentar modelar la distribución de X según la clase de predictores, en nuestro caso el valor Y ($P(X = x\,|Y = k)$) aplicando el teorema de Bayes.

$$P(Y = k\,|X = x) = \frac{P(X = x\,|Y = k)\,P(Y = k)}{P(X = x)} = \frac{P(X = x\,|Y = k)\,P(Y = k)}{\sum_{j=1}^{K} P(X = x\,|Y = j)\,P(Y = j)}$$

En este modelo se asume que $P(X = x\,|Y = k)$ se puede ajustar a la distribución normal multivariada dada por la siguiente ecuación:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\,|\mathbf{\Sigma}|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)}$$

Donde μ_k es la media de los ejemplos de la categoría k y $\mathbf{\Sigma}$ es la covarianza.

Escribimos la ecuación anterior sabiendo que: $P(Y = k) = \pi_k$

$$f_k(x) = \frac{\pi_k \dfrac{1}{(2\pi)^{p/2}\,|\mathbf{\Sigma}|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)}}{\sum_{j=1}^{K} \dfrac{1}{(2\pi)^{p/2}\,|\mathbf{\Sigma}|^{1/2}} \cdot e^{-\frac{1}{2}(x-\mu_j)^T\mathbf{\Sigma}^{-1}(x-\mu_j)}}$$

Se obtiene el límite de decisión tomando los logaritmos por los dos lados de la ecuación además de maximizarla:

$$\delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + x^T\mathbf{\Sigma}^{-1}\mu_k$$

En el caso de dos clases, el límite de decisión se trata de una función lineal de x, donde estas dos dan el mismo valor. Esta función se da de la siguiente manera:

$$\{x : \delta_k(x) = \delta_l(x)\}, 1 \leq j, l \leq K$$

En el caso de que se tenga $K > 2$, hay que hallar las medias y la varianza de pK, y las proporciones de K:

$$\binom{p}{2} K = \left(\frac{p\,(p-1)}{2} \right) K$$

1.1.2. Discriminante cuadrático

Se trata de un método supervisado y es muy parecido al análisis discriminante Lineal (LDA) (Greyrat, 2022). El objetivo de este discriminante es clasificar a los individuos en dos o más categorías en función de las variables predictoras. A diferencia del LDA, este método considera de manera equitativa todas las covarianzas de las variables predictoras, sin embargo, en el análisis discriminante cuadrático (QDA) presenta una estructura más flexible que permite variar las covarianzas entre los diferentes grupos. Por lo tanto, estas covarianzas se calculan por separado, y la ecuación sería la siguiente:

$$\Sigma_i = \frac{1}{N_i - 1} \Sigma_{x \in w_i} (x - \mu_i)\,(x - \mu_i)^T$$

Tras obtener la fórmula anterior, la ecuación general para el cálculo del análisis discriminante cuadrático es la siguiente:

$$P(w_i) = ln\,(P(x\,|w = w_i)) + ln\,(P(w_i))\,, i = 1, 2$$
$$= -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{m}{2} ln\,(2\pi) - \frac{ln\,|\Sigma_i|}{2} + ln\,(P(w_i))$$

Donde π_k es el número de casos que pertenecen a cada clase del total.

Ejemplo

Los datos de los que se parten son los siguientes:

num	x1	x2	x3
1	1	2	1
2	2	1	1
3	4	5	5
4	5	4	6

Se va a calcular las medias de cada uno de los grupos.

x_1	x_2	x_3	μ_1
1	2	1	4/3
2	1	1	4/3
4	5	5	4,67
5	4	6	5

A continuación, se va a calcular la probabilidad que tiene cada clase.

$$P(y_1) = 2/6 = 1/3$$

$$P(y_2) = 2/6 = 1/3$$

Como todas las clases tienen el mismo numero de entradas, todas tienen la misma probabilidad.

Ahora se van a calcular las matrices de covarianzas

$$\Sigma_1 = \frac{1}{2}\left(\left(\begin{pmatrix}1\\2\end{pmatrix} - \begin{pmatrix}4/3\\4/3\end{pmatrix}\right) * \left(\begin{pmatrix}1 & 2\end{pmatrix} - \begin{pmatrix}4/3 & 4/3\end{pmatrix}\right) + \right.$$

$$+ \left(\begin{pmatrix}2\\1\end{pmatrix} - \begin{pmatrix}4/3\\4/3\end{pmatrix}\right) * \left(\begin{pmatrix}2 & 1\end{pmatrix} - \begin{pmatrix}4/3 & 4/3\end{pmatrix}\right) +$$

$$\left. + \left(\begin{pmatrix}1\\1\end{pmatrix} - \begin{pmatrix}4/3\\4/3\end{pmatrix}\right) * \left(\begin{pmatrix}1 & 1\end{pmatrix} - \begin{pmatrix}4/3 & 4/3\end{pmatrix}\right)\right) =$$

$$= \begin{pmatrix}1/3 & -1/6\\-1/6 & 1/3\end{pmatrix}$$

$$\Sigma_2 = \frac{1}{2}\left(\left(\begin{pmatrix}4\\5\end{pmatrix} - \begin{pmatrix}4,67\\5\end{pmatrix}\right) * \left(\begin{pmatrix}4 & 5\end{pmatrix} - \begin{pmatrix}4,67 & 5\end{pmatrix}\right) + \right.$$

$$+ \left(\begin{pmatrix}5\\4\end{pmatrix} - \begin{pmatrix}4,67\\5\end{pmatrix}\right) * \left(\begin{pmatrix}5 & 4\end{pmatrix} - \begin{pmatrix}4,67 & 5\end{pmatrix}\right) +$$

$$\left. + \left(\begin{pmatrix}5\\6\end{pmatrix} - \begin{pmatrix}4,67\\5\end{pmatrix}\right) * \left(\begin{pmatrix}5 & 6\end{pmatrix} - \begin{pmatrix}4,67 & 5\end{pmatrix}\right)\right) =$$

$$= \begin{pmatrix}1/3 & 0\\0 & 1\end{pmatrix}$$

Lo siguiente sería calcular las funciones discriminante para cada clase, para ello se van a necesitar las matrices inversas de las matrices de covarianzas y los respectivos determinantes de estas matrices.

Para el grupo 1 sería:

$$|\Sigma_1| = \frac{1}{3} * \frac{1}{3} - \left(-\frac{1}{6}\right) * \left(-\frac{1}{6}\right) = \frac{1}{9} - \frac{1}{36} = \frac{3}{36} = \frac{1}{12}$$

$$\Sigma_1^{-1} = \frac{1}{|\Sigma_1|}\begin{pmatrix}1/3 & 1/6\\1/6 & 1/3\end{pmatrix} = 12\begin{pmatrix}1/3 & -1/6\\-1/6 & 1/3\end{pmatrix} = \begin{pmatrix}4 & 2\\2 & 4\end{pmatrix}$$

Para el grupo 2 sería:

$$|\Sigma_2| = \frac{1}{3} * 1 = \frac{1}{3}$$

$$\Sigma_2^{-1} = \begin{pmatrix}3 & 0\\0 & 1\end{pmatrix}$$

Una vez que se tengan los determinantes y las inversas se prodece al cálculo de las funciones discriminantes cuadráticas.

Para la clase 1 $\alpha_1(x)$

$$\alpha_1(x) = -\frac{1}{2} * \log\left(\frac{1}{12}\right) - \frac{1}{2}(x-\mu_1)^T * \Sigma_1^{-1}(x-\mu_1) + \log\pi_1 =$$

$$= -\frac{1}{2} * \log\left(\frac{1}{12}\right) - \frac{1}{2} * \begin{pmatrix} x - 4/3 \\ 4/3 \end{pmatrix}^T * \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix} * \begin{pmatrix} x - 4/3 \\ 4/3 \end{pmatrix} + \log \pi_1$$

Para la clase 2 $\alpha_2(x)$

$$\alpha_2(x) = -\frac{1}{2} * \log\left(\frac{1}{3}\right) - \frac{1}{2}(x - \mu_2)^T * \Sigma_2^{-1}(x - \mu_2) + \log \pi_2 =$$

$$= -\frac{1}{2} * \log\left(\frac{1}{3}\right) - \frac{1}{2} * \begin{pmatrix} x - 4,67 \\ 5 \end{pmatrix}^T * \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} x - 4,67 \\ 5 \end{pmatrix} + \log \pi_2$$

1.1.3. Comparación LDA y QDA

De manera sintetizada se va a presentar una tabla mostrando las diferencias entre el LDA y el QDA.

Tabla 1.1: Comparación LDA y QDA.

LDA	QDA
Límites lineales	Límites curvos
Menor flexibilidad	Mayor flexibilidad
Menor problema de varianza	Mayor problema de varianza
Mejores clasificaciones cuando hay menos individuos	Mejores clasificaciones cuando hay más individuos
Matriz de dispersión con $p(p+1)/2$ parámetros	Matriz de dispersión con $Kp(p+1)/2$ parámetros

1.2. Método biplot.

Esta técnica fue introducida por Gabriel en los años 70 y consiste en representar gráficamente una gran cantidad de datos (Nieto-Librero, 2013). Se basa en la descomposición de los valores singulares de una matriz de datos, similar a otras técnicas factoriales, con el objetivo de obtener una representación en un espacio de baja dimensión, minimizando la pérdida de información. Estos métodos permiten analizar la información de las variables, es decir, las representadas en las columnas, frente a los individuos, las variables representadas por las filas. En resumen, se busca aproximar cada elemento (x_{ij}) de la matriz original X_{ij} lo más cercano posible a su valor real. Para lograr esta estimación, se utilizan una serie de vectores llamados marcadores, que se aplican tanto a las filas como a las columnas de la matriz.

Considerando los marcadores $a_1, a_2, ..., a'$ como las filas de la matriz A, y $b_1, b_2, ..., b'$ como los marcadores columna de la matriz B, entonces se pude expresar la matriz X como:

$$X = AB^T$$

Sin embargo, esta matriz no es única, ya que puede variar al hacer rotaciones o escalamientos. La descomposición de los valores singulares de la matriz X se expresa de la siguiente manera:

$$X = U\Lambda V^T$$

Dada la expresión de la descomposición de los valores singulares podemos igualar las matrices A y B a dicha descomposición, y de esta manera poder obtener diferentes procedimientos de factorización. Los componentes que se van a emplear son:

- U: Matriz cuyos vectores columna son ortonormales y pertenecen a los vectores propios XX^T

- V: Matriz cuyos vectores columna son ortonormales y corresponden a los vectores propios $X^T X$

- Λ: Matriz diagonal formada por los valores singulares de la matriz X. Esta matriz está constituída por las raíces cuadradas positivas de los valores propios que corresponden al producto matricial $X^T X$, los cuales se encuentran organizados de forma descendente.

Para que se pueda llevar a cabo este desarrollo, se tiene que verificar que la matriz U como la matriz V sean ortonormales, de esta manera se confirma la unicidad de la factorización. Esto sería de la siguiente manera:

$$U^T U = V^T V = I$$

Se puede ejecutar de diversas formas esta factorización, además de poder realizar representaciones por cada una de estas. Dichas representaciones se hacen dependiendo de las propiedades de los datos, además se pueden llegar a obtener conclusiones importantes. Existen tres factorizaciones distintas, las cuáles se realizan de diferente manera dependiendo de que asignación se les dé a las matrices A y B, además de considerar la descomposición de los valores singulares.

El primer tipo de factorización es el GH-Biplot. Este nombre se debe a que el autor utilizó G para representar la matriz de marcadores de fila y H para la matriz de marcadores de columna. Este método se basa en representar simultáneamente a los individuos y las variables, garantizando la máxima calidad de representación para los individuos. En este caso, la matriz A se define como $U\Lambda$, y la matriz B se define como V.

El siguiente método de factorización se trata del JK-Biplot. Este fue desarrolldo por Jan de Leeuw y Erik-Jan Wagenmakers, y al igual que el anterior, este también permite representar de forma simultánea los individuos y las variables, dónde estas tienen la mayor calidad de representación. Fue llamado de esta manera debido a que sus creadores utilizaron la J para indicar los marcadores fila y la K para las columnas. Este tipo se interpreta de tal manera que las variables que tienen ángulos en torno a 90º o 270º tienen una correlación nula, mientras que si hay un angulo de 0º o 180º, dichas varibales tienen una correlación de 1 o -1.

El último método se denomina HJ-Biplot y fue desarrollado por J. C. Gower y D. J. Hand. Este también perimite la representación sumultánea entre los individuos y las variables. Este se trata de una combinación entre los dos métodos anteriores, debido a esto, tiene una calidad mucho mayor que cualquiera de dichos métodos.

Las características de los marcadores varían según el tipo de Bilpot seleccionado, además estas características son esenciales al decidir qué tipo de Biplot emplear y para la interpretación subsiguiente de los resultados.

La calidad de representación depende del tipo de Biplot, al igual que los marcadores. Este se refiere a la capacidad que tiene cada uno de los tipos de Biplot para mostrar con precisión la distribución además de las relaciones que existen entre las variables y las observaciones en el espacio; es decir, se trata de evaluar qué tan bien los Biplot reflejan la estructura subyacente de los datos. Esto facilita una interpretación mucho más precisa de los resultados.

En la siguiente tabla se presenta la calidad de representación además de los marcadores fila y columna para cada uno de los tipos de Biplot.

Tabla 1.2: Tipos de Biplot.

	Filas		Columnas	
	Coordenadas	Calidad	Coordenadas	Calidad
GH Biplot	U	$\frac{S}{\tilde{S}}$	$V\Lambda$	$\frac{\sum_{s=1}^{S}\lambda_s^4}{\sum_{s=1}^{\tilde{S}}\lambda_s^4}$
JK Biplot	$U\Lambda$	$\frac{\sum_{s=1}^{S}\lambda_s^4}{\sum_{s=1}^{\tilde{S}}\lambda_s^4}$	v	$\frac{S}{\tilde{S}}$
HJ Biplot	$U\Lambda$	$\frac{\sum_{s=1}^{S}\lambda_s^4}{\sum_{s=1}^{\tilde{S}}\lambda_s^4}$	$V\Lambda$	$\frac{\sum_{s=1}^{S}\lambda_s^4}{\sum_{s=1}^{\tilde{S}}\lambda_s^4}$

Capítulo 2

Datos Reales

2.1. Obtención de la base de datos

Para esta base de datos se obtuvieron muestras de pacientes con ictiosis. Para la extracción de estas muestras los dermatólogos frotaron la piel de cada paciente en un área de 2-4 cm^2 en diferentes partes del cuerpo, como la axila, la escapula y la espalda. Para ello emplearon hisopos secos estériles que se colocaron en tubos de 2 mL estériles y se almacenaron congelados. Posteriormente se extrajo el ADN de dichas muestras. Se mandaron secuenciar para así obtener los archivos fastqc. Estos se analizaron y finalmente resultó una hoja de excel para cada uno de los pacientes de los que se extrajeron las muestras. Cada una de las hojas contenía una lista de bacterias, las cuales son las que contiene el paciente y se muestra en porcentaje relativo. Además, por cada una de las zonas de las que se ha extraído la muestra, es decir, por cada zona del cuerpo, bien sea zona sana o enferma o parte del cuerpo, axila o escapula, que se ha examinado, se tiene la abundancia relativa que presenta cada bacteria en cada una de estas zonas, cuanto más presente tenga un tipo de bacteria, más porcentaje va a tener.

Esta limpieza constó de una recodificación mediante programación en python de todas las bacterias, ya que eran nombres muy largos y dificultaban a la hora de hacer los análisis. Estos incluían el filo, la clase, el orden, la familia y el género de cada una de las bacterias que presenta cada paciente. Se sustituyó este nombre por un número asignado para cada una de ellas, este número fue adjudicado en función del orden alfabético de las bacterias. Sin embargo, no se podían unificar todas las distintas hojas de excel en una sola porque los pacientes no tenían ni las mismas bacterias, ni el mismo número de estas. Por ello, se unificaron de forma que las bacterias que no estaban presentes en algunos de los pacientes se les añadía dicha bacteria, pero como se mide en porcentaje relativo, se le ponía un cero, de tal forma que una vez hecho esto con todas las bacterias que mostraban todos los pacientes, se obtuvo el mismo número de bacterias para cada individuo, sin haber modificado el porcentaje relativo, es decir, que aún habiendo añadido bacterias a cada individuo, si se suman los porcentajes de las bacterias de cada región, sigue dando 1.

Una vez hecha la recodificación, se juntaron las distintas hojas de excel en una sola, es decir, se pusieron todos los pacientes en una sola hoja, y se realizaron dos bases de datos distintas. La primera de ellas se organizó de la siguiente manera; se puso un pacientes tras otro en la primera columna, mientras que en la segunda columna se dispusieron las 246 bacterias, es decir que hay un total de 34 pacientes por 246 bacterias filas. En las siguiente cuatro columnas se colocaron las regiones y las condiciones de cada una de las regiones, es decir, sano y afectado axila y sano y afectado escápula. Las últimas columnas se corresponden a la edad, al tipo de ictiosis que presenta cada paciente, el gen responsable de la enfermedad y el género de cada individuo.

La otra base de datos tiene un total de 34 filas, una para paciente. Al contrario que la base de datos anterior, esta se dispone en posición horizontal, es decir, las bacterias en vez de estar todas en la misma columna, una tras otra, en este caso, hay una bacteria por columna, con cada porcentanje relativo para cada uno de los pacientes, por lo tanto ahora hay 246 columnas por cada una de las regiones, y consta de dos regiones y dos condiciones para cada una de las 246 bacterias.

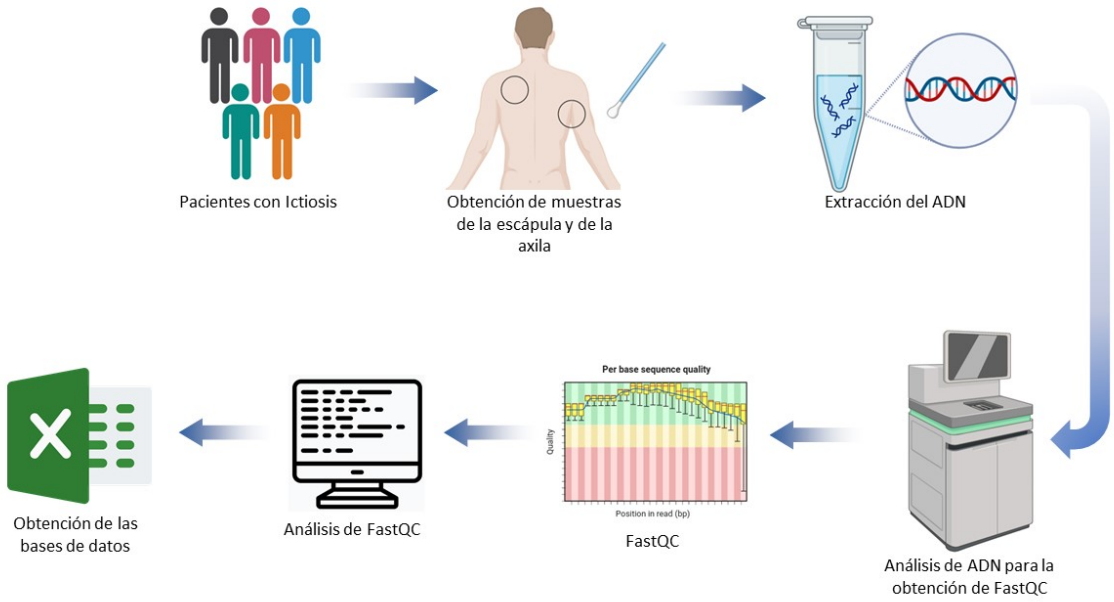


Figura 2.1: Obtención de la base de datos.

2.2. Descripción de la base de datos

En este trabajo se ha utilizado una base de datos compuesta por una gran cantidad de bacterias con sus respectivas frecuencias relativas.

La base de datos está compuesta por 34 pacientes, y en cada uno de ellos se analizaron un total de 246 bacterias en dos regiones diferentes, la axila y la escápula, y en cada una de ellas se analizó tanto la zona sana como la afectada. También se dispone del género, la edad y el tipo de ictiosis que presentan cada uno de ellos.

En dos de los pacientes hay dos muestras de cada región, por lo tanto, en estos individuos se analizaron dos veces las bacterias que tenían presentes.

Se crearon dos bases de datos, una para entrenar al modelo, y otra para poder predecir.

La primera se construyó de tal forma que en la primera columna se distribuyeron los pacientes, y se crearon tantas filas por cada paciente como bacterias hay, por lo tanto, hay un total de 246 filas por cada paciente. En todas estas filas, se especifica el género, el gen que predispone el tipo de ictiosis, la edad y el tipo de ictiosis que presenta, además de las frecuencias relativas por cada bacteria y por cada zona y región.

Paciente	Bacteria	Sano_axila	Afectado_axila	Sano_escápula	Afectado_escácupla	Género	Diagnóstico	Edad
P1 P1 ... P1	B1 B2 ... B246	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer Varón / Mujer ... Varón / Mujer	IE/ARCI/SAM ... IE/ARCI/SAM ... ... IE/ARCI/SAM ...	Variables cuantitativas [0,20]
P2 P2 ... P2	B1 B2 ... B246	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer Varón / Mujer .... Varón / Mujer	IE/ARCI/SAM ... IE/ARCI/SAM ... ... IE/ARCI/SAM ...	Variables cuantitativas [0,20]
P3 P3 ... P3	B1 B2 ... B246	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer Varón / Mujer .... Varón / Mujer	IE/ARCI/SAM ... IE/ARCI/SAM ... ... IE/ARCI/SAM ...	Variables cuantitativas [0,20]
....	....	....	....	....	....	....	....	....
P34 P34 ... P34	B1 B2 ... B246	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer Varón / Mujer .... Varón / Mujer	IE/ARCI/SAM ... IE/ARCI/SAM ... ... IE/ARCI/SAM ...	Variables cuantitativas [0,20]

Figura 2.2: Disposición de los datos de entrenamiento.

La base de datos de predicción se distrubuye de forma completamente diferente a la anterior. En la primera columna están distribuidos los pacientes del 1 al 40, donde cada uno corresponde a una fila. En las columnas estan distribuidas las zonas, tanto sanas como afectadas. Dentro de cada una de las regiones, se encuentran las 246 bacterias. A continuacón de las zonas, se encuentran el Género, el Diagnóstico y la Edad de los individuos.

	Sano Axila				Afectado Axila				Sano Escápula				Afectado Escápula	Género	Diagnóstico	Edad
Bacterias / Paciente	B1	B2	...	B246	B1	B2	...	B246	B1	B2	...	B246	...	...	...	...
Paciente 1	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer	IE/ARCI/SAM ...	Variables cuantitativas [0,20]
Paciente 2	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer	IE/ARCI/SAM ...	Variables cuantitativas [0,20]
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Paciente 40	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	...	Variables cuantitativas [0,1]	Variables cuantitativas [0,1]	Varón / Mujer	IE/ARCI/SAM ...	Variables cuantitativas [0,20]

Figura 2.3: Disposición de los datos de predicción.

Capítulo 3

Software

En este trabajo se ha utilizado principalmente R como lenguaje de programación, más especificamente la interfaz de RStudio. R programa se conoce por su utilización en el ámbito estadístico y por la visualización de datos. Se trata de un software abierto y cuenta con una variedad de de herramientas para realizar análisis avanzados.

Por otro lado, RStudio es un entorno gráfico diseñado para trabajar con R principalmente. Este entorno se trata de una interfaz que facilita la escritura, ejecución y depuración del código R. También ofrece herramientas para la visualización de datos además de la generación de infromes, por lo que lo convierte en una opción popular para los analsitas de datos.

En este proyecto se ha desarrollado una aplicación web, en el que se ha llevado a cabo con el paquete denominado Shiny. Este es una herramienta del entorno de programación en R el cual su funcion es simplficar el proceso de generar aplicaciones web interactivas sin la necesidad de conocimientos profundos en el diseño web.

3.1. Exploración del entorno y lenguaje de R

Inicialmente R es un programa de código abierto. Esto significa que se puede obtener y utilizar sin ningún tipo de restricción. Este lenguaje está diseñado específicamente para realizar análisis estadísticos además de generar representaciones gráficas.

Otra ventaja importante es la capacidad que tiene este programa a la hora de compartir el código, dado que es relativamente sencillo. Esto puede significar que analistas de datos, investigadores o estadísticos de cualquier parte del mundo pueden colaborar y contribuir a los estudios de una manera muy eficiente.

Este programa ha aumentado su popularidad y esto se puede deber a que el leguaje de este es bastante sencillo en compración a otros lenguajes, ademas de que se puede desarrollar y probar modelos rápidamente con resultados satisfactorios. Entre estos modelos incluyen análisis de Big data, pruebas estadísticas clásicas, modelos lineales y no lineales o anásis de series temporales . Además tambien permite que lenguajes como C, C++, Python o FORTRAN tengan integración en este programa para que de esta manera se pueda mejorar su eficiencia. También se puede limpiar, analizar, visualizar y presentar datos con este programa.

Lo primero que se necesita para el uso de este programa es su instalación, este es totalmente gratuíta dado que es un lenguaje abierto como ya se mencionó anteriormente. Una vez que ya se haya instalado en la computadora, se puede interactuar con el programa mediante la interfaz. Muchos usuarios prefieren tener Rstudio dado que es una interfaz mas amigable a la hora de empezar a

utilizar este programa. Una vez que ya se tenga todo lista para usarlo, se escribe el código. Este código va en forma de scripts. Estos scripts es lo que le dicen al programa qué es lo que tiene que hacer. Por ejemplo, se puede escribir un script para importar datos o realizar un determinado gráfico referente a los datos importados anteriormente. Hay que tener en cuenta que R trabaja con objetos, estos pueden ser variables, conjuntos de datos, gráficos, ... Estos objetos se pueden modificar utilizando diferentes funciones. Estas funciones son bloques las cuales realizan distintas labores. Estas funciones bien pueden ya existir en el entorno de R o bien se pueden crear dependiendo las necesidades del usuario. Una de las características mas importantes de este lenguaje son los paquetes. La función de estos paquetes es ampliar la capacidad de R, dado que estos son colecciones de funciones. También se incluyen datos adiciones en las areas que elija el usuario. Algunos ejemplos de estas areas pueden ser el análisis de series temporales o el aprendidaje automático. Por último, R tiene una gran variedad de análisis estadísticos, como pueden ser calculos simples o cálculos mas avanzados. Otra peculiaridad de este programa es la gran diversidad de herramientas para la visualización de datos, dado que cada usuario los puede modificar a conveniencia y a lo que se necesite en cada momento.

En resumen, R es una herramienta muy poderosa ademas de versátil para la visualización de dtaos y el análisis de datos, que permite al usuario tener la flexibilidad para ejecutar una gran variedad de tareas que pueden ser desde simples análisis hasta proyectos de investigación complejos.

3.2. Desarrollo del software

El conjunto de datos original se normalizó de tal manera que todos los datos y todas las variables fuesen comparables entre sí, dado que estas son esenciales para realizar el estudio y de esta forma, todos los datos pertenecientes a dicha base de datos se puedan comparar. Si no se realiza este proceso, además de no poderse comparar los datos, iban a estar sesgados, por lo que no se podrían sacar las conclusiones oportunas al realizar el estudio.

En este software se desarrolla con el programa de R con el cuál se analiza la base de datos del microbioma mencionada en el capitulo anterior. Esta aplicacón tiene como objetivo dividir dicha base de datos en tres partes diferenciadas, una será para la parte de entrenamiento, otra para el testeo y la ultima parte será para predicción del modelo. Se emplea el procedimiento de de dividir dicha base de datos aleatoriamente utilizando una una semilla para que no haya variaciones tanto en la partición de los conjuntos de datos como en los resultados a la hora de ejecutar el código.

Estos tres conjuntos se diviran aleatoriamente tanto las filas como el número total de estas.

- Conjunto de Entrenamiento: Este conjunto contiene una cantidad de filas aleatoria del total de la base de datos. Estas filas contienen todas las variables del conjunto original, es decir que no se pierde nada de información. El conjunto de entrenamiento es el que se encarga de entrenar y ajustar el modelo.
- Conjunto de Testeo: Este conjunto al igual que el anterior, contiene una cantidad de filas aleatoria del total de la base de datos, además, contiene todas las variables del conjunto original, es decir que no se pierde nada de información. El conjunto de Testeo tiene la función de evaluar el rendimiento del modelo que previamente se ha entrenado.
- Conjunto de Predicción: Este conjunto contiene las filas que no se han utilizado en los dos conjuntos anteriores, es decir, las filas restantes. En este conjunto se elimina la variable objetivo, esta variable es la que se encarga de clasificar cada una de las filas por tipo de enfermedad. Este último conjunto es el que se encarga de realizar las predicciones.

3.3. Shiny

Shiny es un paquete del programa de R el cual permite crear aplicaciones web interactivas y dinámicas (Chang, 2020). Con este programa se pueden crear aplicaciones web sin llegar a tener conocimientos de desarrollos web.

Con este paquete, se puede realizar interfaces de usuario interactivas mediante el entorno de R. Esto permite fusionar análisis estadísticos complejos con la representación interactiva, además de los controles de usuario intuitivos.

Shiny emplea un enfoque de progrmación reacctiva. Esto significa que componentes de dicha aplicación, como pueden ser los gráficos o los controles de entrada, están interrelacionados de manera dinámica. Otra función que tiene esta aplicación web es cuando un usuario está con el manejo de esta, y realiza cambios como pueden ser el importar datos, dichos cambios en el programa se realizan automaticamente en la aplicación, es decir que no se necesita regargar la aplicación para que se actualice con los nuevos datos, si no que directamente se actualiza en tiempo real.

3.3.1. Funcionamiento de Shiny

La aplicación de shiny funciona mediante el agoritmo del análisis de dos componentes principales: La primera parte denominada Interfaz de Usuario (ui.R) es la encargada del aspecto y de la disposición de las represenciones visuales con las que interaccionan los consumidores (Armas, 2017). Se tiene una gran variedad de opciones como pueden ser botones o menús plegables, a la hora de hacer la interfaz de usuarios. Además se tiene otra gran variedad en las representaciones gráficas como tablas o mapas.

La configuración de la interfaz se relaciona con:

- La estructura es la que se encarga de las barras laterales, paneles, las pestañas de la aplicación web, además de otra serie de características. Hay tres tipos de estructuras:
  - pageWithSideBar: se ha quedado bastante obsoleto
  - fluidPage
  - fixedPage: fuerza la anchura de la página
- El aspecto es el encargado de los colores, tipos de fuentes de la aplicación web, además de otras serie de elementos. Existe un paquete denominado shinythemes qie es el que contiene diferentes temas que cambian colores y los tipos de fuentes

Otro componente principal de esta aplicación es el servidor (server.R). Este es el encargado de procesar todo el trabajo mas complicado de dicha web. En este constituyente es donde se realizan todos los procedimientos para que se pueda llevar a cabo las órdenes del usuario (Mulero, 2024). Se pueden meter todas las entradas que se introducen el R adémas de otras tareas necesarias y la aplicación web devuelve al beneficiario por la interfaz de dicha web los resultados de las órdenes.

La principal peculiaridad de las aplicaciones creadas con shiny son su interactividad, esto significa que es posible mofificar los datos sin necesidad del código.Esto está relacionado con la programación reactiva, esto significa que cada vez que haya una modificación, se actualiza todo el proceso. Hay tres tipos de programaciones reactivas.

- Fuentes reactivas: Son los inputs se insertan en la interfaz de usuarios y que posteriormente se mandan al servidor.
- Conductores reactivos: Se trata de una transformación de los inputs que se usan en el servidor. Estos conductores cogen los inputs para procesarlos en medida de las ordenes dadas por el usuario.

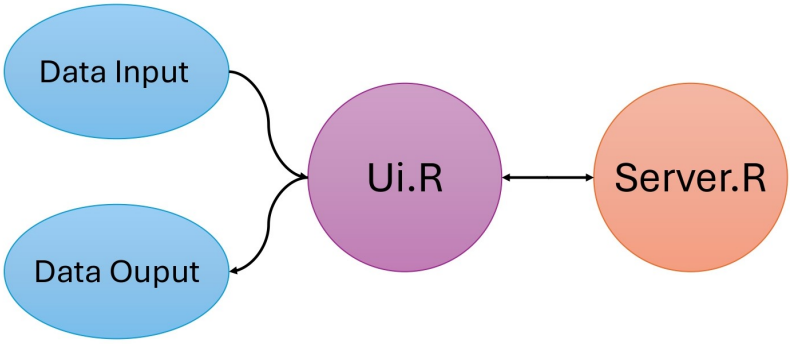


Figura 3.1: Funcionamiento de Shiny.

- Puntos finales reactivos: Son los outputs que envía el servidor y que se mandan de vuelta a la interfaz de usuarios.

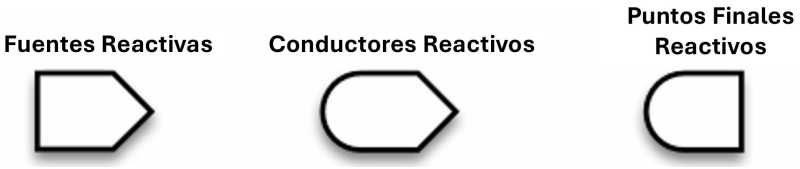


Figura 3.2: Programacion reactiva de Shiny.

El método de programación reactivo permite una interacción dinámica y ágil a través de las expresiones reactivas. Estas últimas se ocupan de separar partes del código garantizando su ejecución únicamente cuando se modifiquen las entradas en las que se basan. Debido a esto se previene realizar las operaciones superfluas las cuales podrían afectar al desempeño de la aplicación web.

El esquema mas simple es coger una fuente reactiva como input y posteriormente conseguir un final reactivo. En tos casos, pueden haber tantos inputs como sean necesarios, y por cada una de estas entradas se genera un output o incluso mas si el usuario lo desea. Además, las entradas pueden ser número y parámetros lógicos. Por otro lado, el paquete shiny dispone de sus propios intputs, los cuales en este caso se denominan widgets.

Tabla 3.1: Widgets.

Tipo de input	Uso
actionButton	Botón de Acción
checkboxGroupInput	Grupo de verificacción de casillas
checkboxInput	Casilla individual de verificación
dateInput	Calendario para seleccionar una fecha
dateRangeInput	Conjunto de Calendarios para seleccionar un rango de fechas
fileInput	Campo para enviar archivos
helpText	Texto de ayuda en el que se puede introducir un cuestinario
numericInput	Casilla para introducir numeros
radioButtons	Conjunto de botones de la radio
selectInput	Caja con varias opciones de las cuales se pueden elegir
sliderInput	Barra deslizante
submitButton	Botón para enviar
textInput	Casilla para intoducir texto

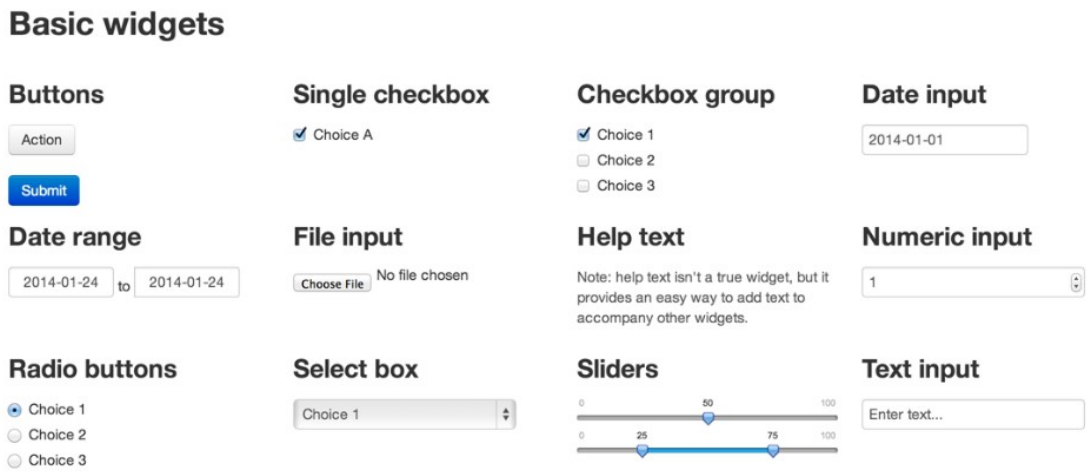


Figura 3.3: Widgets Inputs

Una vez introducido el widget necesario en la interfaz de usuario se usará como input$nombre dentro del servidor de la aplicación.

Las entradas mencionadas anteriormente se importan en la interfaz de usuarios y se mandan al servidor, y esto se usa para conseguir los outputs finales. Los cálculos efectuados en el servidor con las entradas introducidas por el usuario, tienen como resultado los outputs los cuales son de tipo reactivo (los nombres empiezan por “render” y terminan dependiendo el tipo de objeto que se devuelve).

Tabla 3.2: Comparación.

Tipo de objeto que se obtiene	Output	Uso
renderImage	imageOutput	Imágenes
renderPlot	poltOutput	Gráficos
renderPrint	verbatimTextOutput	Imprimir la salida
renderTable	tableOutput	Dataframes, Matrices o cualquier otro tipo de estructuras
renderText	textOutput	Strings
renderUI	uiOutput/ htmlOutput	HTML

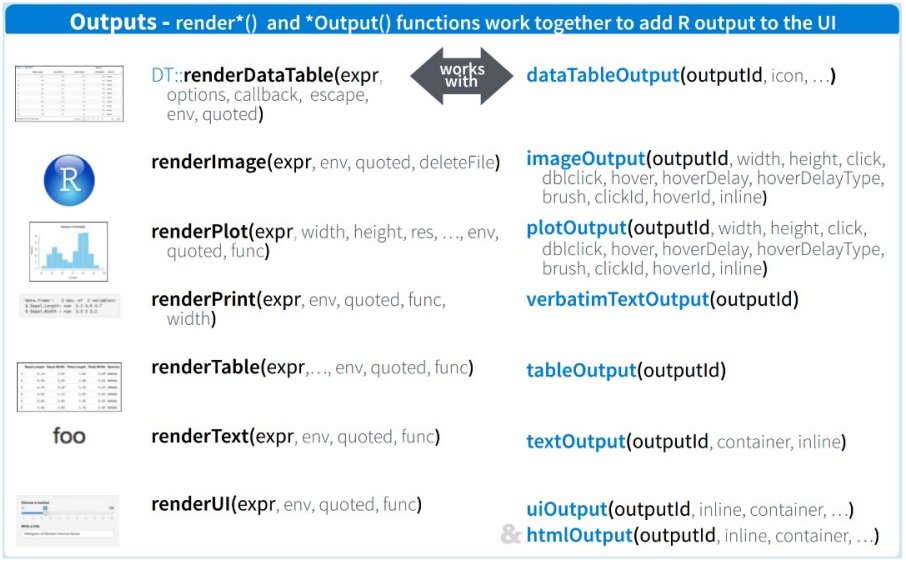


Figura 3.4: Widgets de los outputs.

En las funciones render, se muestran como inputs$nombre depende de lo que haya introducido el usuario en la interfaz de usuario, mientras que las funciones output de render se denominan como output$nombre.

3.3.2. Manual de la interfaz de la App

En este apartado se va a proporcionar un manual detallado para que de esta manera los usuarios que desean utilizarla y que no tengan muchos conocimientos de programacion, la puedan manejar sin ninguna complicación. En cuanto a los usuarios que tienen conocimientos previos, se les va a explicar el código que se ha empleado para crear el análisis disdriminante cuadrático. Algo importante a mencionar es que el lenguaje empleado para el desarrollo de esta aplicación ha sido programado en R exclusivamente.

Para empezar, la interfaz de la aplicación creada por el paquete Shiny, está diseñada para que sea fácil de usar para los usuarios. Esta interfaz se denomina interfaz de usuario (UI). Se tiene que cargar la base de datos con la que se va a proceder a realizar los cálculos y análisis específicos de esta aplicación, además también se tiene que ajustar los parámetros para que dichos cálculos y análisis se realicen correctamente.

A continuación, se va a proceder a ver el funcionamiento de la interfaz. Lo primero que se van a encontrar los usuarios cuando accedan, es el título el cuál se encuentra en mayúsculas.

Una vez que los usuarios se hayan enterado en lo que consiste esta aplicación, la interfaz muestra un campo que sirve para cargar los datos, hay que darle a “Browser”, como se puede observar en la figura de abajo, para importar los datos que se van a analizar con la aplicación. Este botón para cargar los datos, permite navegar a través de los archivos del ordenador para elegir el Excel deseado. Esta base de datos debe de estar en formato de Excel (.xlsx) para que la aplicación no de problemas y se puedan ejecutar los algoritmos correctamente.

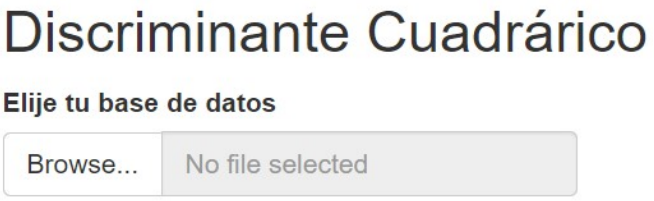


Figura 3.5: Salida para cargar la base de datos.

Posteriormente, los usuarios se van a encontrar varios cuadros interactivos en los que van aparecen los parámetros que se tienen que ajustar para poder realizar el análisis adecuadamente. Estos parámetros se ajustan a los requerimientos de los usuarios. Estas especificaciones son determinantes para el algoritmo ya que de estos dependen los resultados y las conclusiones que se van a obtener.

Algunos de los ajustes que se pueden realizar son “porcentaje de datos de entrenamiento” o “porcentaje de datos de testeo”. Esto lo que permite es ajustar al gusto de los usuarios, la cantidad de filas que desean emplear para realizar el entrenamiento y posteriormente el testeo de los datos. En ambos campos, es necesario introducir valores numéricos dado que es lo que permite la interfaz. Una vez proporcionados las cantidades de datos que se desean emplear, la aplicación usa esos valores para proceder a realizar el algoritmo.

Otro parámetro que aparece es el denominado “porcentaje de datos de prediccion”. Este no hay que ajustarlo, dado que, el programa suma el porcentaje de los dos apartados anteriores y se lo resta al total, dando como valor este último. De esta manera, los usuarios saben el porcentaje exacto que tienen para esta última sección.

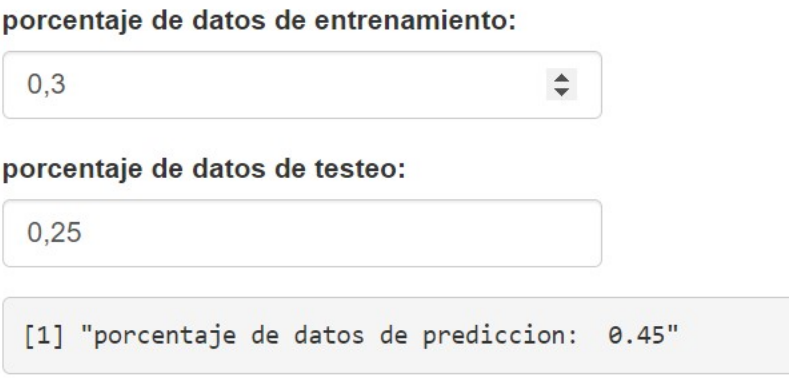


Figura 3.6: Porcentaje de las secciones del análisis.

Una vez que el usuario halla determinado todos los parámetros que se han mencionado anteriormente, la aplicación mostrará los resultados del análisis. Hay tres apartados diferentes para que los resultados de los datos sean mas accesibles a los usuarios.

En la primera de las secciones se exhibe los resultados alcanzados en una tabla en la que se observa los centroides del análisis. Estos se han realizado mediante unos cálculos y análisis específicos. Todos estos análisis se realizan una vez determinados los parámetros mencionados con anterioridad. El resultado final se presenta en un formato comprensible para los usuarios, este formato es en forma de texto. Lo que permite este formato es captar las conclusiones y por ello los descubrimientos que pueden acarrear dichas conclusiones, del análisis realizado.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula
-0.03	0.08	0.05	-0.00
-0.06	-0.09	-0.08	-0.24
0.06	-0.06	0.00	-0.02
-0.08	0.04	-0.09	-0.26
-0.07	-0.00	-0.08	-0.23
-0.08	-0.12	-0.08	-0.13

Figura 3.7: Centroides.

El siguiente apartado, se muestra la bondad de ajuste, esto nos permite saber que bien se adapta el modelo a los datos de entrenamiento. Este se puede observar justo debajo de la tabla mencionada anteriormente.

```
[1] 0.09214092
```

Figura 3.8: Bondad de ajuste del modelo.

La última de las secciones se exhibe una tabla, la cual se genera a partir de otra función diferente. Esta sección ofrece datos completarios y resultados más minuciosos del análisis. La tabla que saca el análisis está estructurada de tal manera para que los usuarios puedan investigar y estudiar los datos de una manera más organizada. Al tener esta estructura, los resultados proporcionados permiten a los usuarios obtener una visión mas detallada, además de poder indagar más en los análisis efectuados.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula	X.predict.modelo_qda..prediccion...class
0.00	-0.17	-0.17	-0.34	ICTI CONFETI
-0.24	-0.17	0.00	-0.33	IE
0.53	-0.32	-0.27	-0.75	IE
0.00	-0.17	-0.17	-0.25	ICTI CONFETI
0.00	0.00	0.00	-0.25	SAM
0.00	0.00	0.00	-0.17	ICTI CONFETI
0.00	-0.17	-0.17	5.36	ARCI
0.00	0.00	0.00	-0.17	ICTI CONFETI

Figura 3.9: Predicción de los datos.

Una vez explicados todas las partes de la interfaz de la app, se va a mostrar a continuación una visión general de dichos resultados de la interfaz una vez ya cargada una base de datos.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula
-0.03	0.08	0.05	-0.00
-0.06	-0.09	-0.08	-0.24
0.06	-0.06	0.00	-0.02
-0.08	0.04	-0.09	-0.26
-0.07	-0.00	-0.08	-0.23
-0.08	-0.12	-0.08	-0.13

```
[1] 0.09214092
```

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula	X.predict.modelo_qda..prediccion...class
0.00	-0.17	-0.17	-0.34	ICTI CONFETI
-0.24	-0.17	0.00	-0.33	IE
0.53	-0.32	-0.27	-0.75	IE
0.00	-0.17	-0.17	-0.25	ICTI CONFETI
0.00	0.00	0.00	-0.25	SAM
0.00	0.00	0.00	-0.17	ICTI CONFETI

Figura 3.10: Resultados de los datos.

Esta App se ha diseñado con el propósito de hacer más accesible a los usuarios realizar amálisis de una manera mas sencilla, dado que se trata de una interfaz que tiene una organización muy simple además de clara. Los componentes interactivos favorecen la realización de las funciones requeridas por el usuario de una manera más eficiente, lo que conlleva sacar conclusiones de una forma más óptima.

3.4. Código de R

El código que se va a explicar a continuación, se encuentra en los anexos. En el apartado mencionado anteriormente, se va a poder visualizar el algoritmo completo.

Antes de proceder al algoritmo, se realizó una normalización de los datos dado que debido a que los rangos y a las medidas que tenían, no se podían realizar los análsis. Por todo ello, se llevó a cabo la tipificación de los datos. Este proceso consiste en unificar los datos de tal manera que tengan una media de cero y desviación típica de uno.

A continuación, se va a proceder al desarrollo del algoritmo del análisis cuadrático discriminante en Rstudio.

Inicialmente se utilizó la librería “readxl” en la cual se eligió la función “read_excel” para importar la base de datos. Estos se almacenan en una variable para que de esta manera se pueden cargar los datos y manejarlos para realizar los análisis de una forma más sencilla. También se necesita cargar la librería “MASS”. Esta es la encargada de contener las funciones que se van a utilizar en el análisis que se va a describir a continuación.

Una vez cargados y preparados los datos para los análisis, se definirá una semilla utilizando la función “set.seed”. Esta función asegura la consistencia de los resultados, ya que al fijar la semilla, los números aleatorios generados serán los mismos en cada ejecución del código. A continuación, los datos se dividirán en conjuntos de entrenamiento, prueba y predicción. Esta división de los datos se hace cuando el usuario procede a introducir el porcentaje que quiere destinar para cada uno de los conjuntos. Esta parte lo que hace es calcular una proporción aleatoria de las filas totales para cada uno de los conjuntos de entrenamiento y testeo, esto se hace con la funcion “sample”.

Estos indices se van a acumualr en una variable denominada “indices_aleatorios” Para finalizar la parte de división de los datos, se le aplica la función “round”, dicha función redondea el resultado de la multiplicación realizada anteriormente, al número entero más cercano. Esto asegura que los conjuntos de entrenamiento y testeo tengan un número entero de filas. En cuanto al conjunto de predicción se suma las filas destinadas a las partes dichas anteriormente, es decir el número total de filas de entrenamiento y testeo y este, se le resta el número total de filas totales, de esta manera, las filas restantes son las asignadas para esta última parte.

Una vez ya se tengan divididos todos los conjuntos, se almacenan el las variables denominadas “tamaño_entrenamiento”, “tamaño_testeo”, “tamaño_predicción”. Estos objetos contienen las filas que corresponden a cada uno de los conjuntos. Se van a utilizar para realizar los análisis correspondientes.

Posteriormente, se utiliza la función “indices_aleatorios” para que sean aleatorias las filas que se generaron en el paso anterior. A continuación, se almacena en un objeto denominado “entrenamiento”, dado que estos datos corresponden a ese conjunto.

En cuanto a los datos de testeo, se realiza lo mismo que a los datos de entrenamiento, es decir, se usa la función “indices_aleatorios” para que las filas generadas antes de este paso sean aleatorias. Este proceso se acumula en la variable denominada “testeo”.

Finalmente, el apartado de predicción se logra restando el número total de filas del conjunto de datos inicial a la suma de las filas del conjunto de entrenamiento y de testeo. Una vez obtenidas las filas designadas a la parte de predicción, este conjunto se guarda en un objeto denominado “predicción”.

En resumen, el objetivo de las últimas líneas de código descritas, es dividir el conjunto de datos que se tiene inicialmente en tres subconjuntos distintos (entrenamiento, testeo y predicción), basándose en una selección de índices aleatorios que se ha generado previamente. Esto asegura que cada uno de los subconjuntos contiene filas seleccionadas al azar de la base de datos original.

3.4.1. Funcion QDA

Se va a realizar un análisis QDA (Quadratic Discriminant Análisis), para ello se utiliza la función “qda” y para ajustarlo a los datos, se pone en primer lugar la variable respuesta seguido de “ ” lo que indica que se utilizarán todas las demás variables en entrenamiento como predictores. Todo esto se acumula en un objeto denominado “modelo_qda”. En esta función se lleva acabo un análisis interno el cual se va a proceder a explicar a continuación.

Función de Entrenamiento

A continuación se va a proceder a explicar como funciona la función de entrenamiento mencionada anteriormente.

En primer lugar, se va a definir cómo se va a llamar la función, en este caso se va a denominar “f_entrenamiento”. Posteriormente, se calcula cuantas columnas tiene la base de datos que se va a utilizar. A continuación, se van a seleccionar las columnas de la base de datos que sean numéricas, ya que esas columnas son las que se van a utilizar para realizar el posterior análisis. Estas variable nuéricas se acumulan en otro objeto para poder utilizarse a posteriori.

El siguiente paso sería sacar los valores únicos de los tipos que hay en la variable respuesta, en este caso sería saber cuántos tipos de icthiosis hay en toda la base de datos. Esto se haría con funciones como “unique” y “unlist” dado que estas nos indican el número de categorías de la variable dependiente. Los valores unicos se guardan en otra variable que se va usar más tarde.

Una vez que se tengan todos los subtipos de la enfermedad, se va a realizar un conteo de las filas que pertenecen a cada uno de los subtipos. Para ello, se puede usar la función “sapply”. A continuación, se suman todas las variables correspondientes al objeto creado anteriormente. Esta sunma se acula en otra varible para poder utilizar estos datos posteriormente.

Posteriormente, se calculan las medias de las columnas. En estas se encuentran las bacterias por las diferentes zonas del cuerpo de los pacientes. Estas zonas son: “sano_axila”, “afectado_axila”, “sano_escapula” y “afectado_escapula”. La finalidad de esto es calcular las medias de las columnas numéricas para cada tipo de ictiosis y las organiza en una matriz donde cada fila corresponde a una clase. Para calcular las medias de estas columnas, se utilizan funciones como“sapply”, “apply”, “mean”.

Lo siguiente del código es, para cada uno de los subtipos de enfermedad, filtrar las variables numéricas que pertenecen a cada una de las clases. También calcula la media de cada columna numérica para las filas filtradas, se puede utiliza “sapply” para combinar estas medias en una matriz. Para finalizar, se traspone la matriz resultante para que las filas correspondan a las clases y las columnas correspondan a las medias de las variables numéricas. El resultado es una matriz donde cada fila representa a un subtipo de la enfermedad y cada columna representa la media de una variable numérica para esa clase, y se guarda en una variable.

Para continuar, se tienen que centrar los datos y para ello se resta la media de la clase correspondiente a cada muestra. Para ello se realiza un código en el que para cada fila, en el conjunto de datos, encuentra la clase correspondiente a la fila “i”, después, se resta la media de las variables numéricas de esa clase a los valores de la fila “i”. El objetivo de este paso es centrar los datos de la fila. Posteriormente, se le aplica alguna función como “lapply” para que cree una lista de vectores ya centrados. Además, se utiliza la función “unlist” para convertir la lista en un vector plano. Posteriormente se convierte el vector en una matriz “d“ con “A” filas. Cada fila contiene los valores centrados de las variables numéricas para una muestra del conjunto de datos. El resultado es una matriz donde cada fila representa una muestra de “x” con sus valores centrados, es decir, se le ha restado la media de la clase correspondiente a cada valor.

El siguiente paso sería realizar las matrices de covarianzas de las diferencias centradas para cada clase de ictiosis. Para lograr esto, se emplea alguna función como "lapply", que ejecuta una función en cada elemento del vector y se aplica a cada uno de los subtipos de ictiosis y devuelve los resultados en forma de lista. Dentro de la función se selecciona la columna la cuál contiene la clase de cada muestra y genera un vector lógico, donde la clase de las muestras es igual a “i”. Una vez que se tenga el vector con la columna deseada, se seleccionan las filas de “d”, esta es la matriz que contiene las diferencias centradas, correspondientes a la clase “i”. Una vez que ya se tenga todo seleccionado, se traspone la matriz para poder calcular el producto de las matrices. El resultado de este producto es la matriz de covarianzas no normalizadas de las diferencias centradas para la clase “i”.

Posteriormente, se calculan las probabilidades de cada clase, es decir, la proporción de muestras en cada clase respecto al total de dichas muestras.

El siguiente paso sería realizar los cálculos para poder obtener las funciones discriminates. Estas se realizan para cada una de las clases y siguen la siguiente ecuación:

$$P(w_i) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{m}{2} ln(2\pi) - \frac{ln|\Sigma_i|}{2} + ln(P(w_i))$$

Además de introducir la fórmula en Rstudio, se convierta en una lista de valores con las puntuaciones para cada clase, en un vector simple.

Por último, se aplica los datos de entrenamiento previamente obtenidos a la función de entrena-

miento para que devuelva finalmente los valores reales.

Función de Testeo

El siguiente paso sería realizar la función de testeo. Aquí se utilizan otros datos distintos al paso anterior. En primer lugar se crea la función y se toman los argumentos que se crean convenientes. A continuación, se escogen las variables numéricas de la base de datos, dado que solo se necesitan esas variables para crear el modelo, también se seleccionan las clases únicas de la variable respuesta.

Una vez que se tenga el análisis de los datos de prueba, se realiza la bondad de ajuste de los datos de testeo. La bondad de ajuste nos muestra el porcentaje sobre qué tan bien está prediciendo el modelo realizado sobre los datos reales

Para finalizar, se realiza una tabla de confusión. Esta muestra cuántas veces cada clase real fue predicha como cada clase.

3.4.2. Predicción

Para finalizar el código, se realizan nuevas predicciones con los datos seleccionados para la predicción.

Estos datos se almacena en un objeto denominado "predicciones_nuevas", y este contiene los resultados de la predicción, incluyendo las clases predichas (predicciones_nuevas$class) y las probabilidades asociadas a cada clase.

3.4.3. Interfaz

Se va proceder a la explicación de la interfaz de usuario de la app, la cual se ha creado mediante Shiny. Por ello, lo primero que se va a realizar es cargar la librería "Shiny". Una vez cargado el paquete, se va a proceder al diseño del entorno gráfico.

En primer lugar se va a crear una interfaz que se ajuste a la pantalla del ordenador de cada usuario, dado que cada uno tiene un dispositivo con distintas dimensiones, esto se hace con la función "fluidPage". Lo siguiente sería incluir el título de la aplicación nada más abrir el entorno gráfico y se realiza mediante la función "titlePanel". Este título se muestra arriba a la izquierda en letras grandes.

Otra función que hay que incluir es el denominado "fileInput". Este se encarga de que el usuario elija la base de datos con la que desea realizar los análisis. Permite navegar a través de las diferentes carpertas y archivos que tiene el usuario en su dispositivo para que pueda escoger los datos correspondientes. Además se puede especificar el tipo de archivo de estos datos, como pueden ser ".xlsx" o "xls".

La interfaz le pide al usuario el porcentaje de los datos que desea destinar a entrenamiento y testeo. Para poder realizar estos cuadros, es necesario la función de "numericInput". Este se trata de un cuadro de entrada numérica donde cada usuario puede indicar el número que desea emplear. Este valor debe ser entre 0 y 1 dado que las probabilidades no puedes ser menores a 0 y mayores a 1. En su defecto viene 0,30 y 0,25 respectivamente para entrenamiento y testo, pero el usuario va poder cambiar estos datos siempre que quiera.

Para poder saber el porcentaje destinado a la predicción es necesaria la función "verbatimTextOutput". Este nos muestra el porcentaje restante del conjunto de entrenamiento y testeo.

Una vez ya ingresados los datos necesarios para que la app calcule el análisis, también es necesaria la visualización de los resultados. Se van a observar tres salidas distintas. La primera de estas

muestra una tabla denominada “centroides”. Como bien su nombre indica, esta nos muestra los centroides del análisis, en este caso serían los centroides del análisis discriminante cuadrático. Para realizar esta salida se utiliza la función de “tableOutput”.

La siguiente salida nos muestra la bondad de ajuste, es decir, nos indica la precisión del modelo. Este valor estará comprendido entre 0 y 1. Cuanto más cerca de 1 esté mejor será la precisión del modelo. Esta salida se presenta con la función ya mencionada anteriormente “verbatimTextOutput’.

Por último, se van a mostrar las predicciones del modelo mencionado. Nos va a mostrar los resultados que indican cuál es la probabilidad mas alta de que ocurra uno de los valores de la variable independiente, es decir, de la que se está intentando predecir. Esto se basa en los valores observados de las variables dependientes del conjunto de datos. Para llevar a cabo esto, se usan los datos no utilizados hasta ahora, es decir los denominados “prediccion”.

En resumen, este código configura una interfaz que permite al usuario subir un archivo de datos en formato determinado, además de poder determinar los porcentajes de datos que se usarán para el entrenamiento y la prueba, y luego visualizar resultados como el porcentaje de precisión de las predicciones, los centroides obtenidos y las predicciones del modelo. Estos componentes constituyen la base de una aplicación interactiva en Shiny para el análisis discriminante cuadrático.

3.5. Algoritmo

Para poder realizar todos los cálculos y el código descrito previamente, es necesario una serie del ecuaciones para que los análisis se realicen de forma adecuada. Estas formulas se muestran a continuación.

Para realizar el análisis discriminante se necesita una matriz X de N muestras $[x_i]_{i=1}^{N}$, donde tienen una dimensión de m columnas. En este caso se trata de una matriz de 8856 filas y 5 columnas.

El primer paso sería hacer la media de cada clase $\mu_i(m$ x 1) dada la siguiente ecuación (Tharwat, 2016):

$$\mu_i = \frac{1}{n_i} \sum_{i=1}^{n_i} x_i, x_i \in w_i, \forall i = 1, 2, ..., c$$

A continuación, se calcula la probabilidad de que ocurra cada una de las clases ($P(w_i)$), y se realiza de la siguiente manera:

$$P(w_i) = \frac{n_i}{N}$$

El siguiente paso es realizar la matriz de covarianzas para cada una de las clases (Σ_i), siguiendo la siguiente fórmula:

$$\Sigma_i = \frac{1}{n_i} \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T \, \forall i = 1, 2, ..., c$$

Para finalizar el análisis discriminante se sacan las ecuaciones discriminantes para cada una de las clases:

$$P(w_i) = ln(P(x|w = w_i)) + ln(P(w_i)), i = 1, 2$$

$$= -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{m}{2} ln(2\pi) - \frac{ln|\Sigma_i|}{2} + ln(P(w_i))$$

3.6. Representación

3.6.1. Representación QDA

En cuanto a la representación del QDA, se va a describir el código para poder realizarla. Lo primero sería cargar la librería para cargar la base de datos que se va usar para realizar los gráficos, la librería se denomina "readxl".

De la base de datos que se ha cargado previamente, se obtienen las variables únicamente numéricas. Esto se hace con la función "sapply", indicando dentro de esta el nombre que se le haya querido dar a la base de datos y a continuación se pone "is.numeric", ya que esto es lo que hace que solo seleccione las numéricas.

La otra variable que se necesita para la representación del QDA. Esta es la que nos indica el tipo de enfermedad que tiene cada individuo, es la que se encuentra en la última posición de la base de datos. Esta variable se acumula en un objeto ya que el siguiente paso sería juntar las variables numéricas que se extrajeron previamente con esta última variable, que sería la variable predictora. Una vez se hayan juntando todas estas variables se acumulan en nuevo objeto, ya que estos datos son los que se van a usar para realizar los gráficos.

A continuación se carga la librería denominada "MASS", esta es la que contiene funciones como "qda", la cuál se va a utilizar para realizar el modelo de los datos que se han juntado anteriormente. De esta manera se comprueba que no hay ningún error y se va a poder realizar el modelo sin ningún problema.

Por último se carga la librería "klaR". Esta es la que nos va a permitit realizar las representaciones del modelo QDA. También se va a tener que definir el número de intervalos. Estos dependen del número de variables numéricas que quieras representar, sería el numero de variables numéricas más uno. Finalmente se realiza la visualización del modelo. Esto se lleva a cabo mediante la función "partimat". Dentro de esta, hay que introducir los datos con los que se quiere que haga el análisis, además de otra serie de varibales.

3.6.2. Representación Biplot

Por último, en cuanto al QDA, se va a describir el código para la representación de este método. En primer lugar, se carga la librería para importar los datos que se va usar para realizar los gráficos. Lla librería se denomina "readxl".

De la base de datos que se ha cargado previamente, se obtienen las variables únicamente numéricas. Esto se hace con la función "sapply", indicando dentro de esta el nombre que se le haya querido dar a la base de datos y a continuación se pone "is.numeric", ya que esto es lo que hace que solo seleccione las numéricas.

La otra variable que se necesita para la representación del QDA. Esta es la que nos indica el tipo de enfermedad que tiene cada individuo, es la que se encuentra en la última posición de la base de datos. Esta variable se acumula en un objeto ya que el siguiente paso sería juntar las variables numéricas que se extrajeron previamente con esta última variable, que sería la variable predictora. Una vez se hayan juntando todas estas variables se acumulan en nuevo objeto, ya que estos datos son los que se van a usar para realizar los gráficos.

Una vez que se tengan los datos para realizar los cálculos y posteriormente los gráficos, se calcula la descomposición singular de los datos, con la función "svd".

Una vez que se tenga la descomposición de los valores singulares se seleccionan las columnas de los primeros vectores singulares izquierdos. Además, se crea una matriz diagonal con estos dos primero valores singulares. Estos dos conjuntos se multiplican estre sí para dar los datos originales en un espacio de menor dimensión. Esto se acumula en on objeto para utilizarlo a posteriori.

Se realiza exactamente lo mismo, pero con con los vectores singulares derechos. Este da como recultado las relaciones entre las columnas originales en el nuevo espacio bidimensional

A continuación, se calculan los valores máximos absolutos de las dos primeras columnas de las matrices que se han creado en el paso previo. Se calcula el valor máximo de estos valores y se guarda en un objeto ya que este, se va a dividir entre el espacio de menor dimensión, es decir, el que se obtuvo de los valores singulares izquierdos, que se halló previamente.

Lo siguiente sería realizar el gráfico con los datos que se hallaron al redicir la dimensión de los valores singulares izquierdos. Se van a delimitar los ejes, tanto los máximos cómo los mínimos. Se dbujan la flechas desde el origen con la función “arrows” y se añade el texto mediante la función “text”.

Finalemte se realiza el gráfico mediante la función “plot”.

Todo este código se realiza para reducir la dimensión de las flechas que representan el biplot y poder ver las variables que representan cada una de estas flechas.

En resumen, este código realiza un análisis de datos, el cuál incluye la normalización de una matriz, la creación de un gráfico con flechas y texto, la generación de una cuadrícula de datos, la aplicación de una descomposición en valores singulares y la visualización de los resultados del modelo QDA.

Capítulo 4

Resultados

En este último apartado se van a presentar los resultados obtenidos usando el algoritmo de clasificación mostrado en la aplicación creada mediante la herramienta de Shiny en RStudio, descrita en el capítulo anterior.

En primer lugar, se realizó una tipificación de los datos debido a los diferentes rangos y medidas que presentaban, lo que impedía realizar los análisis. Por esta razón, se llevó a cabo la estandarización de los datos. Este proceso consiste en unificar los datos de tal manera que tengan una media de cero y una desviación estándar de uno, de este modo se elimina el sesgo que pudiese tener previamente.

En cuanto a la base de datos utilizada para el análisis, se compone de 34 pacientes, todos menores de 18 años. Por cada uno de ellos se les extrajo una muestra de diferentes regiones, estas son la axila y la escápula, además, por cada una de estas zonas se analizaron muestras sanas y muestras afectadas en la piel por la enfermedad. Por otro lado, se recogieron un total de 246 bacterias entre los dos tipos de regiones dichas previamente.

Después de haber realizado la limpieza de la base de datos, se procedió a realizar el algoritmo que dió lugar a la aplicación web, ejecutada con la herramienta de Shiny, donde se mostrarán los resultados de esta a continuación. Gracias a Shiny se van a poder explicar las conclusiones de una manera mucho más efectiva, debido a que se han podido obtener unos resultados más precisos.

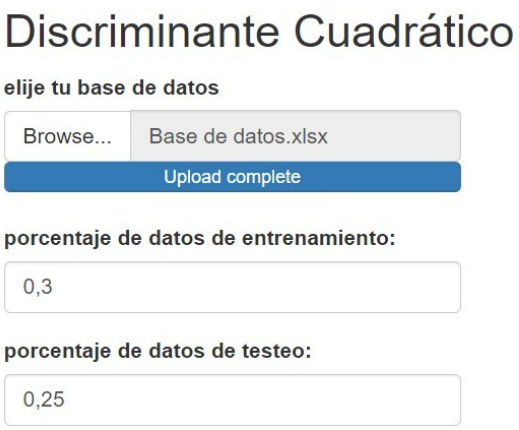


Figura 4.1: Resultado obtenido de los datos.

```
[1] "porcentaje de datos de prediccion:  0.45"
```

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula
-0.03	0.08	0.05	-0.00
-0.06	-0.09	-0.08	-0.24
0.06	-0.06	0.00	-0.02
-0.08	0.04	-0.09	-0.26
-0.07	-0.00	-0.08	-0.23
-0.08	-0.12	-0.08	-0.13

Figura 4.2: Resultado obtenido de los datos.

En la Figura 4.1 se puede observar lo primero que visualiza nada más meter la base de datos en la aplicación. En este caso, la base de datos se denomina “Base de datos”.

Para estos datos se tenía un total de 34 pacientes, de los que de dos de estos se tienen muestras sacadas dos veces, por lo que a efectos prácticos hacen un total de 36. Por otro lado, se obtuvieron un total de 246 bacterias en cada uno de estos pacientes, además, como se mencionó anteriormente, se les extrajo una muestra de diferentes regiones, las cuales son la axila y la escápula, y por cada una de estas zonas se analizaron muestras sanas y muestras afectadas. Esto significa que tenemos 5 columnas, 4 de variables independientes y la última de ellas corresponde al diagnóstico perteneciente de cada uno de los pacientes. Esta correspondería a nuestra variable dependiente.

En cuanto al número total de filas, se tienen a efectos prácticos 36 individuos, y cada uno de ellos tiene 246 bacterias, lo que hace un total de 8856 filas.

También se dividió la base de datos en tres conjuntos distintos, estos son entrenamiento, testeo y predicción. El usuario decidirá el porcentanje de filas que quiera destinar a cada uno de dichos conjuntos anteriores. En caso de que no asigne ningún valor a dichos conjuntos se designaran los que vienen en su defecto, que serían 30 % de los datos para entrenamiento, es decir, algo más de 2500 filas, 25 % de los datos para testeo, 2214 filas, y el resto para la parte de predicción, que sería el 45 % de los datos, es decir, 3985 filas. Con este número de filas es como se van a presentar los siguientes resultados.

En primer lugar, se puede observar una tabla con las 4 zonas muestreadas de los pacientes en las columnas, y en las filas los diferentes tipos de enfermedad, estas son, ARCI, ICTI CONFETI, IE, KID, NETHERTON y SAM, en ese orden respectivamente.

La Figura 4.3 nos muestra las medias por cada zona y tipo de enfermedad. Se puede observar que los valores más bajos se encuentran todos ellos en la zona de “Afectado_Escápula” con los resultados de -0.26, -0.24 y -0.23, en las enfermedades de IE, ICTI CONFETI y KID.

En cuanto al valor mas alto estaría en la zona de “Afectado_axila” con un resultado de 0.08, en la enfermedad de ARCI, seguido de la zona de “Sano_axila” con un valor de 0.06, en la enfermedad de IE y con una décima de diferencia la siguiente zona que le continúa es la de “Sano_Escapula” en la enfermedad de ARCI.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula
-0.03	0.08	0.05	-0.00
-0.06	-0.09	-0.08	-0.24
0.06	-0.06	0.00	-0.02
-0.08	0.04	-0.09	-0.26
-0.07	-0.00	-0.08	-0.23
-0.08	-0.12	-0.08	-0.13

```
[1] 0.09214092
```

Figura 4.3: Bondad de ajuste obtenido del modelo.

Debajo de la tabla de las medias descrita anteriormente, se puede observar la bondad de ajuste. Lamentablemente, no es nada buena, ya que nos da un valor de 0.0921, es decir, un 9.21 % de los datos predichos los coloca correctamente con la enfermedad adecuada. Esto se puede deber a que no hay tantos pacientes debido a que se trata de una enfermedad rara y hay pocos casos diagnosticados.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula	X.predict.modelo_qda..prediccion...class
0.00	-0.17	-0.17	-0.34	ICTI CONFETI
-0.24	-0.17	0.00	-0.33	IE
0.53	-0.32	-0.27	-0.75	IE
0.00	-0.17	-0.17	-0.25	ICTI CONFETI
0.00	0.00	0.00	-0.25	SAM
0.00	0.00	0.00	-0.17	ICTI CONFETI
0.00	-0.17	-0.17	5.36	ARCI
0.00	0.00	0.00	-0.17	ICTI CONFETI
-0.24	-0.24	-0.18	-0.50	ICTI CONFETI
4.78	-0.17	0.00	-0.33	IE
0.00	0.00	0.00	-0.17	ICTI CONFETI
0.00	-0.17	-0.17	-0.33	ICTI CONFETI
-0.17	-0.17	-0.18	-0.39	ICTI CONFETI
0.00	0.00	-0.17	-0.24	ICTI CONFETI
0.00	0.00	0.00	-0.22	ICTI CONFETI
0.00	-0.17	-0.17	3.74	ARCI

Figura 4.4: Resultados de los datos de prediccion obtenido del modelo.

Los resultados que se observan en la Figura 4.4 se corresponden a la predicción de los datos. Para esta última parte se han utilizado el 45 % de los datos.

Cada una de las filas se corresponde a una entrada diferente, en este caso sería a una bacteria en un individuo en concreto. En cuanto a las columnas, estan representadas las variables de la base de datos, a excepción de la última. Esta denominada “X.predict.modelo_qda..prediccion...class”, se corresponde a la enfermedad que predice en función del resto de variables de cada fila, es decir, de cada entrada.

Los valores numéricos mostrados en la tabla se corresponden a la base de datos una vez ya tipificada. En este caso, se predicen 295 entradas como ARCI, 2557 como ICTI CONFETI, 130 como IE, 25 como KID, 151 como NETHERTON y 827 como SAM. Estos resultados son bastante anómalos dado que la gran mayoría de los datos corresponden a los tipos IE Y ARCI, ya que son los tipos más comunes. Mientras que los resultados mostrados por las predicciones muestran que los

tipos más frecuentes son SAM e ICTI CONFETI, destacando por una gran diferencia este último. Estos dos tipos son bastante raros, por lo que sorprende la gran cantidad de predicciones que han obtenido. Se puede deber a que no hay muchos individuos con estos tipos de enfermedad en la base de datos y por ello no hay suficientes datos para que se puedan clasificar de manera correcta, además, puede ser que los pocos datos que haya se parezcan entre ellos y por ello los clasifica de manera incorrecta.

Otra manera de la que se podría implementar la bondad de ajuste y por ello predecería de una manera más correcta sería realizando el análisis con los tipos de ictiosis más comunes, además de cambiar el número de entradas que se quieran destinar a entrenamiento, testeo y predicción.

Dado que la bondad de ajuste es muy baja, se va a realizar de nuevo el modelo, pero esta vez eligiendo de la base de datos los tipos de enfermedad más comunes, siendo estos ARCI e IE. Dado que del resto de enfermedades hay uno o como máximo dos pacientes y aunque sea una enfermedad rara y no haya muchos casos diagnosticados, se cree que no es muy representativo tener un número tan bajo de pacientes.

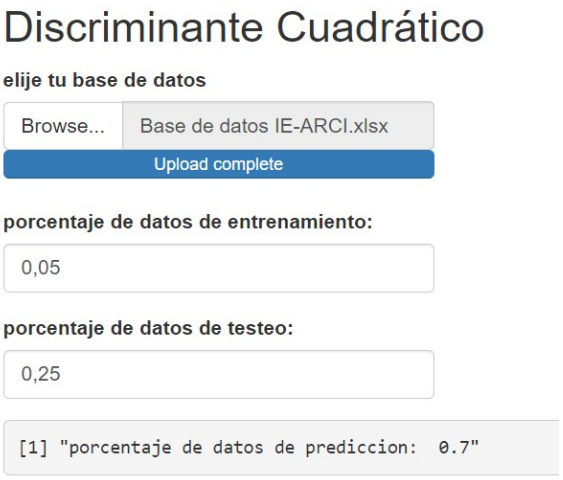


Figura 4.5: Resultados de los datos de prediccion obtenido del modelo.

Ahora se ha reducido la base de datos, de seis tipos que se tenía en la variable dependiente se ha pasado a dos. Por ello, se ha reducido el número de entradas de la base de datos, es decir, ha disminuido el número de filas. Para el siguiente análisis se tienen un total de 7626 filas, de las cuales 4182 pertenecen a la clase de ARCI y 3444 son del tipo IE. Estos datos son mucho más equitativos en comparación a los resultados mostrados anteriormente.

Como ya se mencionó previamente, la base de datos está compuesta por 5 columnas, 4 variables independientes y la última corresponde a la variable dependiente, es decir, la que nos muestra qué enfermedad tiene cada entrada. Las variables independientes serían “Sano_axila”, “Sano_Escapula”, “Afectado_axila” y “Afectado_Escápula”.

Se dividió la base de datos en tres conjuntos distintos, estos son entrenamiento, testeo y predicción. El usuario decidirá el porcentaje de filas que quiera destinar a cada uno de los conjuntos anteriores. En este caso se le van a asignar unos valores de 0.05, 0.25, y 0.7 respectivamente, como se puede observar en la Figura 4.5.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula
-0.08	0.07	0.01	-0.07
0.11	-0.09	0.08	0.01

```
[1] 0.5314795
```

Figura 4.6: Medias y bondad de ajuste del modelo.

En primer lugar se puede observar una tabla con las 4 zonas muestreadas de los pacientes en las columnas, estas son, “Sano_axila”, “Afectado_axila”, “Sano_Escapula” y “Afectado_Escápula” y en las filas los diferentes tipos de enfermedad, estas son, ARCI e IE en ese orden respectivamente.

La tabla muestra las medias de los valores de la base de datos, por cada zona y tipo de enfermedad. Se puede observar que los valores más bajos se encuentra en la zona de “Afectado_axila” con un valor de -0.09, en la enfermedad de IE, seguido muy de cerca por las zonas “Sano_axila” y “Afectado_Escápula”, ambos en la enfermedad de ARCI con valores de -0.08 y -0.07 respectivamente. Estos valores nos pueden indicar diferencias entre las zonas sanas y las zonas afectadas. Se puede incluir la zona afectada de la axila dado que esta es una zona donde suelen haber muchas más bacterias que en otras partes del cuerpo como puede ser la escápula.

En cuanto al valor más alto se encuentra en la zona de “Sano_axila” con un resultado de 0.11, en la enfermedad de ARCI, seguido de la zona de “Sano_Escapula” con un valor de 0.08 en la misma enfermedad dicha anteriormente. Estos valores nos pueden indicar que las zonas sanas en esta enfermedad tienen mayores valores que frente a las zonas afectadas y frente a la otra enfermedad (IE).

Debajo de la tabla de puede observar la bondad de ajuste. Este aunque sigue siendo bastante baja, ha mejorado con los resultados mostrados previamente, incrementando en más de 40 puntos porcentuales. Con estos datos y los valores introducidos del porcentaje que se quiere destinar a los conjuntos del análisis, la bondad de ajuste ha obtenido un valor de 0.53147, es decir un 53.147 % de los datos destinados a la predicción se han predicho de manera correcta.

Sano_Axila	Afectado_axila	Sano_Escapula	Afectado_Escapula	X.predict.modelo_qda..prediccion...class
0.00	0.00	0.00	5.92	IE
5.92	-0.33	3.15	-0.43	IE
0.00	0.00	-0.17	-0.17	ARCI
-0.24	-0.24	-0.17	-0.26	ARCI
0.00	0.00	0.00	-0.17	ARCI
-0.21	-0.39	-0.26	-0.66	ARCI
0.00	0.00	0.00	-0.17	ARCI
-0.18	-0.24	-0.17	3.35	IE
-0.30	0.00	-0.20	-0.45	IE
0.00	0.00	0.00	5.92	IE
0.00	0.00	0.00	-0.17	ARCI
0.00	-0.17	0.00	-0.17	ARCI
-0.27	-0.17	-0.21	-0.30	IE

Figura 4.7: Resultados de los datos de predicción obtenido del modelo.

Los resultados se pueden observar en la Figura 4.7, que se corresponden a la predicción de los datos. Para esta parte se ha utilizado un 70 % de los datos totales.

Cada una de las filas se corresponde a una entrada diferente, en este caso sería a una bacteria en un individuo en concreto. En cuanto a las columnas, están representadas las variables de la base de datos, a excepción de la última. Esta, denominada “X.predict.modelo_qda..prediccion...class”, se corresponde a la enfermedad que predice en función de cada entrada.

Los valores numéricos mostrados en la tabla se corresponden a la base de datos una vez ya tipificada, con los dos tipos de enfermedad más comunes seleccionados. En este caso, se predicen 4843 entradas como ARCI y 496 entradas como IE. Se sigue notando sesgo a la hora de predecir. Esto puede ser porque no hay variables que sean claramente identificatorias para cada clase de enfermedad, porque aunque haya diferencias entre los valores como ya se mencionó anteriormente, no hay diferencias significativas como para que se pueda clasificar según estas variables.

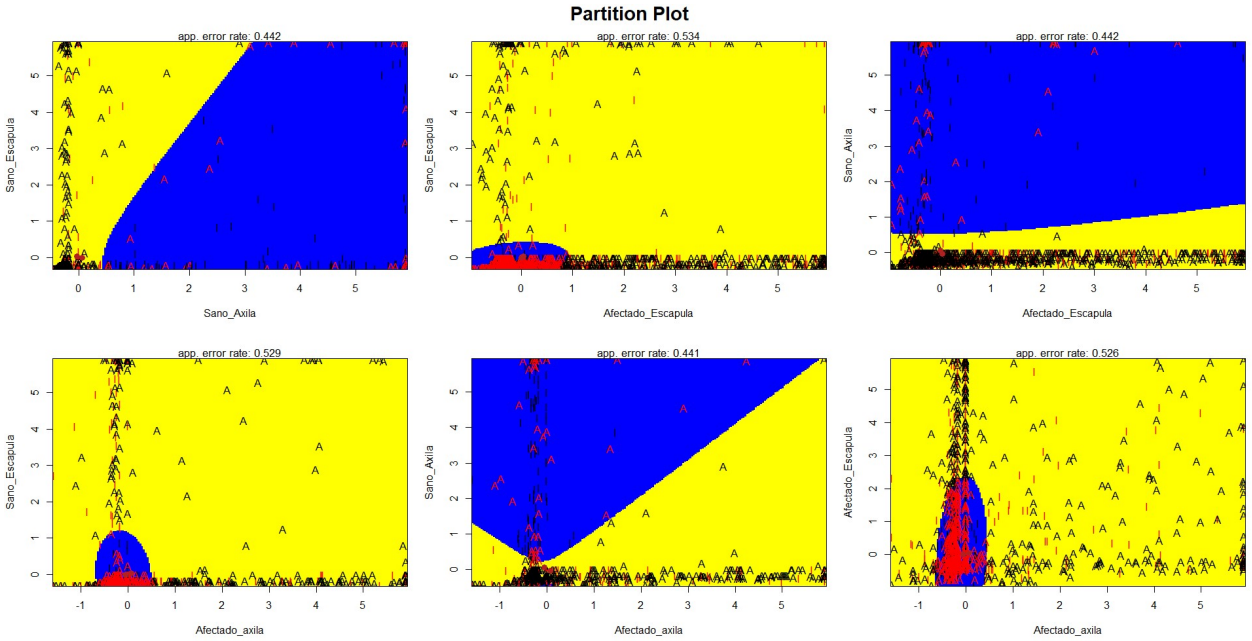


Figura 4.8: Resultados IE vs ARCI.

Los anteriores gráficos muestran áreas de decisión coloreadas en azul y amarillo que representan las regiones donde el modelo QDA clasifica los puntos como pertenecientes a las clases IE y ARCI, respectivamente. Se ha decidido realizar este gráfico dado que son los tipos más comunes y tienen un número de pacientes más similar que el resto de variables.

En el gráfico se puede observar A e I, estas son las iniciales de las enfermedades mencionadas anteriormente. Cada una de las letras son cada una de las bacterias por cada tipo de enfermedad. Las que están en negro significan que están bien representadas, mientras que las que se encuentran en rojo significan que no las clasifica de manera correcta.

Los gráficos que representan Sano_Escapula vs Sano_Axila, Afectado_axila vs Sano_Axila y Sano_Axila vs Afectado_Escapula son los que menos tasa de error tienen. Todos ellos tienen en común Sano_Axila, por lo que podemos deducir que esta variable es una muy buena clasificadora en contraposición al resto.

Algunos gráficos muestran áreas donde una clase es dominante, pero estas áreas son limitadas y no abarcan una porción significativa del espacio de características.

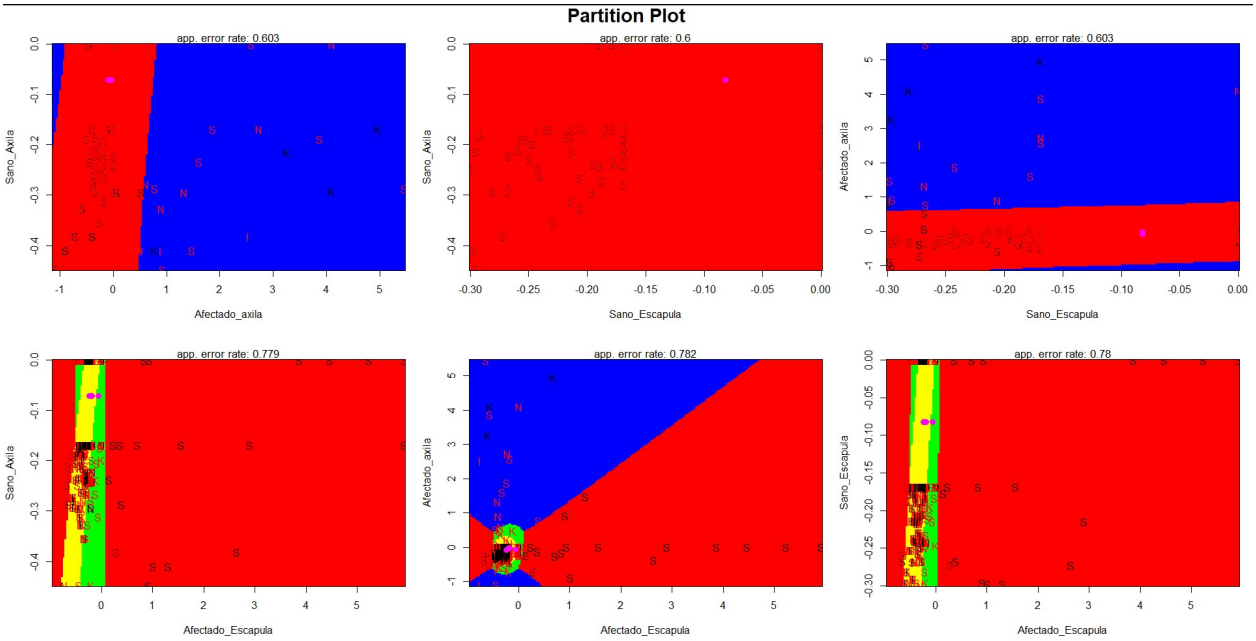


Figura 4.9: Resultados Neth, Sam, Confeti y Kid.

Los anteriores gráficos muestran áreas de decisión coloreadas en azul, amarillo, verde y rojo que representan las regiones donde el modelo QDA clasifica los puntos como pertenecientes a las clases Netherton, SAM, Confeti y KID, respectivamente. Se ha decidido realizar este gráfico dado que son los tipos menos comunes, pero tienen un número de pacientes más similar que el resto de enfermedades.

En el gráfico se puede observar “N”, “S”, “C” y “K”. Estas son las iniciales de las enfermedades mencionadas anteriormente. Cada una de las letras son cada una de las bacterias por cada tipo de enfermedad. Las que están en negro significan que están bien representadas, mientras que las que se encuentran en rojo, significa que no las clasifica de manera correcta.

En estos gráficos se observan mayores tasas de error que en los anteriormente descritos. Esto viene dado, ya que en estos datos no hay un gran número de pacientes, por lo que cuanto menor sea el número de pacientes para poder predecir menor será la precisión.

Se puede observar que los datos correspondientes a la enfermedad SAM, representada mediante “S”, están bastantes dispersos, y no se aprecia un límite claro en el modelo. En el gráfico de Sano_Escapula vs Sano_Axila, solo se observa esta enfermedad, por lo que podemos deducir que si un paciente tiene una gran parte de las bacterias en las zonas sanas, lo mas probable es que tenga este tipo.

En cuanto a los tipos Confeti y Netherton, se caracterizan por estan presentes en la zona de Afectado_Escapula y en una zona concreta de esta. Es decir, si hay algún paciente que tiene bacterias entre las frecuencias -1 y 0 y abundan en la zona afectada de la escápula, lo mas probable es que tenga alguna de estas dos enfermedades.

El tipo de enfermedad KID, se caracteriza por estar presente en la parte de Afectado_axila, y con valores por encima de 3. Con esto se quiere decir que si en un paciente abundan bacterias en la zona afectada de la axila y tiene valores altos de las bacterias, lo más probable es que tenga la enfermedad de KID.

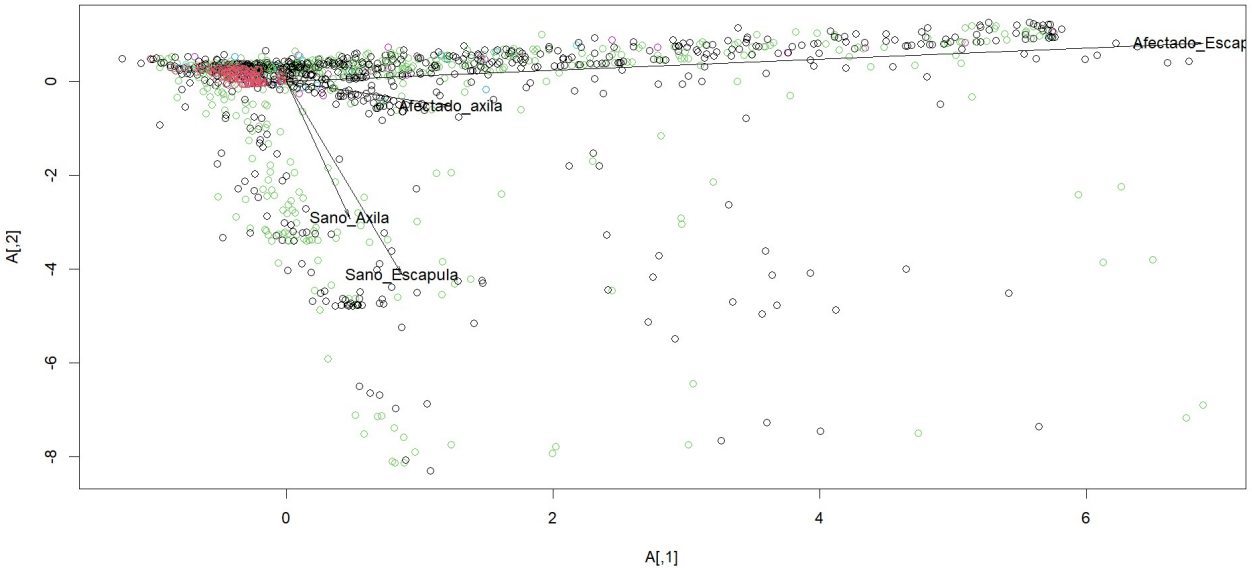


Figura 4.10: Representación biplot.

En la Figura 4.10 se puede observar un biplot. Este se trata de un gráfico que combina las características de un gráfico de dispersión con un análisis de componentes principales, permitiendo visualizar la relación entre las observaciones y las variables en un espacio reducido de dos dimensiones.

Cada punto en el gráfico representa una observación, en este caso las observaciones son bacterias de dierentes pacientes con diferentes tipos de enfermedad. Cada uno de estos tipos viene representado por un color distinto.

Las flechas representan las variables proyectadas en el espacio de las componentes principales. Los vectores indican la dirección y magnitud de cada variable original en el espacio de las componentes principales. En las flechas viene representada las zonas de las que se extrajo la muestra para ver si los pacientes tenían bacterias, estas zonas son,“Sano_Axila”, “Sano_Escapula”, “Afectado_axila” y “Afectado_Escapula”. Estas indican la posición de las variables originales proyectadas en el espacio de las componentes principales.

En este gráfico, las observaciones (las bacterias) cercanas a la punta de un vector tienen valores altos en esa variable. Se pueden visualizar perfectamente las bacterias que distinguen a la zona de .Afectado_Escapula". Se trata de la zona más dintinguible y que presenta una mayor dependencia de las bacterias.

Las zonas de “Sano_Axila” y “Sano_Escapula” tienen un ángulo pequeño entre los vectores que las representan, esto indica que hay correlaciones positivas, es decir, que hay relación entre esas zonas. Y que las bacterias que están relacionadas con una de esas zonas también estan relacionadas con la otra.

En cuanto a la zona “Afectado_axila”, se encuentra en medio del resto de vectores, esto nos indica que no es buen clasificador. Se puede deducir que es porque en esta zona abundan bacterias ya en condiciones normales, así que si se tiene una enfermedad no es una zona de la que sea prioritario coger muestras para analizar. Aun así, esta zona tiene un grupo de bacterias que se observan que son representativas para distinguirla.

Capítulo 5

Conclusiones

En este trabajo se ha desarrollado una aplicación web mediante la herramienta de RStudio con el propósito de predecir y diagnosticar la ictiosis a partir de una base de datos. A continuación se van a presentar las conclusiones obtenidas en relación con los objetivos planteados.

- Se ha proporcionado una descripción detallada del análisis discriminante lineal y cuadrático (QDA), ya que este último se creó a partir del primero. Se han explicado todos los fundamentos teóricos matemáticos y estadísticos. La comprensión es crucial para su correcta implementación e interpretación de los resultados obtenidos. Además, se han ilustrado los principios básicos de cómo el QDA diferencia entre múltiples clases en un conjunto de datos al asumir que cada clase tiene su propia matriz de covarianzas.
- Se ha aplicado el método QDA a la clasificación de la base de datos de ictiosis, logrando categorizar los diferentes Diagnósticos con las variables de los datos.

  Los resultados obtenidos han demostrado la eficacia del QDA en la clasificación de datos complejos del microbioma, habiendo proporcionado una herramienta útil para estudios futuros en el campo de la microbiología.

  Aunque la clasificación no ha sido muy precisa, sienta un precedente para su implementación en este campo, sin embargo, se puede utilizar con otros conjuntos de datos biológicos y seguir mejorando la capacidad de realizar diagnósticos precisos y personalizados.
- Se ha desarrollado una aplicación web en RStudio que utiliza algoritmos de análisis de datos, específicamente QDA, para predecir y diagnosticar la enfermedad, en este caso la ictiosis, a partir de la base de datos correspondiente. La aplicación permite a los clínicos cargar datos nuevos y obtener predicciones sobre la presencia de ictiosis, facilitando así el proceso de diagnóstico.

  La herramienta creada tiene el potencial de ser un recurso valioso para los clínicos, mejorando la precisión del diagnóstico de ictiosis y, por ello, la calidad de vida de los pacientes afectados por esta enfermedad, ya que se les puede dar un tratamiento más específico para cada tipo de ictiosis.

En conclusión, este trabajo ha logrado desarrollar una aplicación web eficaz que integra el análisis discriminante cuadrático para la clasificación y diagnóstico de la ictiosis. La descripción matemática del QDA y su aplicación a bases de datos biológicos han sido exitosamente implementadas, proporcionando una herramienta práctica para el ámbito clínico. Este proyecto no solo contribuye al campo de la bioinformática y la medicina, sino que también abre nuevas posibilidades para la aplicación de técnicas estadísticas avanzadas en el diagnóstico de enfermedades.

Bibliografía

Amat-Rodrigo, J. (2016). *Análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA)* [Accessed: 26-06-2024]. https://rpubs.com/Joaquin_AR/233932

Armas, M. (2017). *Clase 3 - Shiny* [Accessed: 26-06-2024]. https://rstudio-pubs-static.s3.amazonaws.com/339312_4e857d2d6a054cfd88cf749afc9f8a95.html

Chang, W. (2020). *shiny: Web Application Framework for R* [Accessed: 15-06-2023]. https://cran.r-project.org/web/packages/shiny/index.html

Cutro, J. (2010). *Evolución Historia de la Minería de Datos | Dataprix* [Accessed: 26-06-2024]. https://www.dataprix.com/es/mineria-datos-aplicada-encuesta-permanente-hogares/171-evolucion-historia-mineria-datos

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, *84*, 165-175. https://doi.org/10.1080/01621459.1989.10478752

Greyrat, R. (2022). *Análisis Discriminante Cuadrático* [Accessed: 26-06-2024]. https://barcelonageeks.com/analisis-discriminante-cuadratico/

Gutiérrez-Cerrajero, C., Sprecher, E., Paller, A., Akiyama, M., Mazereeuw-Hautier, J., Hernández-Martín, A., & González-Sarmiento, R. (2023). Ichthyosis. *Nature Reviews Disease Primers*, *9*, 2. https://doi.org/10.1038/s41572-022-00412-3

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to Statistical Learning: with Applications in R*. New York: Springer.

Marin, J. (2016). *Capítulo 23 Analisis Discriminante: El procedimiento discriminante*. https://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/23discr.pdf

Mulero, J. (2024). *Qué es shiny El diseño de la aplicación El contenido de la aplicación La interactividad de shiny Aplicaciones interactivas diseñadas con shiny*. https://rua.ua.es/dspace/bitstream/10045/54325/1/shiny.pdf

Nieto-Librero, A. B. (2013). Estudio de la precisión de los resultados del HJ-Biplot a través de los métodos bootstrap. https://gredos.usal.es/bitstream/handle/10366/128225/TFM_MAADM_Nieto_Librero_AnaBelen.pdf

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*, 145-180. https://doi.org/10.1504/IJAPR.2016.079050