Universidad Nacional Autónoma de México

Facultad de Ciencias

Técnicas de análisis multivariado en *"Credit Scoring"*

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I A

P R E S E N T A :

MARTHA ANGELICA MONTES FONSECA

DIRECTORA DE TESIS:
DRA. RUTH SELENE FUENTES GARCÍA

2012

1. **Datos del alumno**
   Montes
   Fonseca
   Martha Angelica
   Universidad Nacional Autonoma de Mexico
   Facultad de Ciencias
   Actuaría
   303221469
2. **Datos del tutor**
   Dra
   Ruth Selene
   Fuentes
   García
3. **Datos del sinodal 1**
   Mat
   Margarita Elvira
   Chávez
   Cano
4. **Datos del sinodal 2**
   Dr
   Ricardo
   Ramírez
   Aldana
5. **Datos del sinodal 3**
   M en C
   Jésica
   Hernández
   Rojano
6. **Datos del sinodal 4**
   Act
   Francisco
   Sánchez
   Villarreal
7. **Datos del trabajo escrito**
   Técnicas de análisis multivariado en "Credit Scoring"
   135 p
   2012

Índice general

Introducción

El Análisis Multivariado es el conjunto de métodos estadísticos cuya finalidad es analizar simultáneamente conjuntos de datos multivariados en el sentido de que hay varias variables medidas para cada individuo u objeto estudiado. Su razón de ser radica en la necesidad de un mejor entendimiento del fenómeno u objeto de estudio, obteniendo información que los métodos estadísticos univariantes y bivariantes son incapaces de conseguir. Los objetivos pueden ser de diversos tipos: Establecer perfiles, separar de grupos, segmentar, determinar influencias eficientes entre varias variables, entre otras.

En general se clasifica en dos grandes grupos:

- Métodos explicativos como regresión lineal, análisis discriminante, regresión logística, modelos de respuesta probit, logit, modelos loglineales, entre otros.
- Métodos descriptivos como análisis de conglomerados, análisis de factores, análisis de componentes principales, análisis de correspondencias simples y múltiples, entre otros.

Este tipo de análisis ha avanzado mucho y hoy ocupa el corazón del análisis estadístico avanzado. Desafortunadamente es exigente en comprensión matemática para usarios en diversas áreas, por ello los resultados gráficos son muy importantes.

En este trabajo emplearemos técnicas de análisis multivariado para resolver un problema al cual nos enfrentamos día tras día: el *problema de la clasificación*, que es uno de los primeros objetivos que aparecen en la actividad científica. En muchas ocasiones la resolución de problemas y la toma de decisiones requieren de una primera tarea que consiste precisamente en clasificar el problema o la situación. Otras metodologías podrán aplicarse después y dependerán en buena medida de esa clasificación. Cuando hablamos de clasificar a un sujeto en un grupo determinado, a partir de los valores de una serie de variables medidas u observadas, y esa clasificación tiene un cierto grado de incertidumbre, resulta razonable pensar en la utilización de una metodología probabilística, que nos permita cuantificar esa incertidumbre.

Desde el punto de vista estadístico podemos distinguir dos enfoques diferentes del problema de la clasificación.

En el primero de ellos los grupos están bien definidos (a priori) y se trata de determinar un criterio para etiquetar cada individuo como perteneciente a alguno de los grupos, a partir de los valores de una serie limitada de parámetros (áreas de entrenamiento). A este metodo se le conoce como **Clasificación supervisada**. En este caso las técnicas más utilizadas se conocen con el nombre de análisis discriminante, aunque como veremos existen otras posibles alternativas, tales como la utilización de la regresión logística, así como otros métodos no paramétricos como *"K- nearest neighbour"*, *"K- means clustering"*, Aprendizaje de cuantización vectorial (LVQ), entre otros.

El segundo enfoque corresponde a aquel caso en el que a priori no se conocen los grupos y lo que precisamente se desea es establecerlos a partir de los datos que observamos. A estos metodos se les conoce como **Clasificación no supervisada**. Ahora las técnicas estadísticas más utilizadas en esa área se conocen con el término análisis de *"cluster"*, que podemos traducir como análisis de agrupaciones y también como análisis de conglomerados por algunos autores. En la figura 1 mostramos un ejemplo de clasificación.

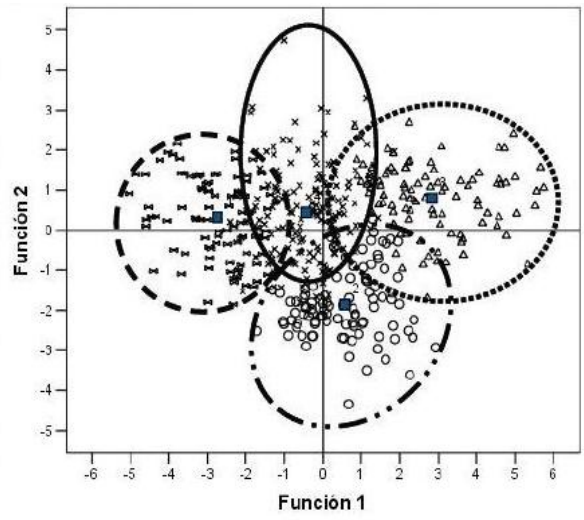

Figura 1: Ejemplo de clasificación

Como hemos comentado existen dos metodologias para tratar el problema de clasificación sin embargo en este trabajo comprende solo métodos de *clasificación supervisada* aplicados a un problema al que se enfrentan las entidades financieras, como es el *"Credit Scoring"*, que consiste en tomar una decisión entre, si a un nuevo solicitante de crédito se le otorga o se le rechaza dicho crédito, sobre una base de rendimiento esperado por la entidad financiera. Como se puede observar es necesario hacer uso de la teoría de clasificación ya que lo que se busca es hacer dos grupos uno con los individuos que sí se les otorga un crédito y el otro grupo con los individuos a los que se les rechaza el crédito

solicitado.

Para llevar a cabo la elaboración de este trabajo se estructuró en 5 capítulos organizados de la siguiente manera:

Capítulo 1: En este capítulo se define el concepto de *"Credit Scoring"*, se hace una breve reseña histórica de cómo fue surgiendo a lo largo del tiempo y cuáles fueron los motivos que llevarón a utilizar *"Credit Scoring"*. Por otro lado se hace una breve descripción sobre algunos métodos que se han visto relacionados en el *"Credit Scoring"*, aunque hay que señalar que sólo se estudiaran cuatro técnicas como son: Análisis Discriminante, Regresión Logística, *"K-nearest neighbour"* (KNN) y *"Learning Vector Quantization"* (LVQ).

Capítulo 2: En este capítulo se hace el desarrollo de las técnicas estadísticas a explorar, en particular para dos de ellas, análisis discriminante y regresión logística. Para ambas técnicas se muestra la teoría y un ejemplo sencillo en el que se puede apreciar la aplicación de cada técnica.

Capítulo 3: En este capítulo se hace la aplicación de los métodos explorados en el capítulo 2 al problema de interés, ya que se cuenta con una base de datos Australiana que contiene los datos de clientes a los a quienes se les otorgó una tarjeta de crédito conjuntamente con clientes que no se les concedió dicha tarjeta. Asímismo antes de comenzar con la aplicación se realiza un análisis exploratotio de la base de datos, para observar el comportamiento de las variables. Para ambos métodos se intenta llegar a una conclusión acerca de la clasificación de los grupos.

Capítulo 4: En este capítulo debido a los resultados obtenidos en el capítulo 3 se desarrolla la teoría de dos metodos estadísticos no paramétricos de automatización como son *K- nearest neighbor"* y *"Learning vector quantization"*. Se presenta el algoritmo que utiliza cada método para llevar acabo la clasificación, también se presenta un ejemplo de cada método con la finalidad de ilustrarlos.

Capítulo 5: En este capítulo se hace la aplicación de los métodos estudiados en el capítulo 4 a la base de datos de *"Credit Scoring"*, primero se revisa si la base de datos cumple con cada uno de los supuestos establecidos por cada uno de los métodos para después observar la forma de clasificación de cada método y se menciona si alguno de estos métodos es un buen clasificador.

Finalmente se concluye comentando cuál de los cuatro métodos es un mejor clasificador para la base de datos.

Capítulo 1

"Credit scoring"

1.1. ¿Qué es el *credit scoring*?

Primero comenzaremos con traducir estas palabras, algunos autores las traducen como calificación de crédito, otros como puntaje de crédito, pero ¿Qué es realmente el *Credit Scoring*? Es un método ó modelo que, mediante un conjunto de técnicas, de manera automática, evalúa el riesgo de crédito de un solicitante de financiamiento o de alguien que ya es cliente de la entidad. Estos métodos ayudan al prestamista a definir a qué prestatario se le otorgará un crédito, la cantidad de créditos y las estrategias operacionales que mejorarán la rentabilidad de los prestatarios[1]. En función de toda la información disponible es capaz de predecir la probabilidad de incumplimiento asociada a una operación crediticia. Estas técnicas determinan el riesgo de prestarle a un determinado cliente. Estas técnicas se aplican de manera individual ya que tienen por objeto medir el riesgo ya sea por empresa o por individuo, sin importarles lo que ocurra con el resto de los prestatarios.

Hablamos de créditos personales o para pequeñas o medianas empresas ya que la evaluación de grandes empresas implica la revisión de aspectos cualitativos de difícil estandarización, por lo cual el resultado se expresa como una calificación o *rating* y no como un *score*.

El resultado de la evaluación se refleja en la asignación de alguna medida que permita comparar y ordenar a los evaluados en función de su riesgo, a la vez que cuantificarlo. Por lo regular estos modelos le asignan a los evaluados un puntaje o *score* o una calificación o *rating*. Algunos métodos los asignan a grupos en donde cada grupo tiene un perfil de riesgo distinto.

[1] *Credit Scoring and its Applications*; Thomas Lyn C.,Edelman David B. siam, Monographs of Mathematical Modeling and Applications

En el caso de los modelos de aplicación de *scoring*, las entidades financieras generalmente determinan un *cut off* o punto de corte para determinar qué solicitudes se aceptan (por tener un puntaje mayor o igual al *cut off*) y cuáles no. En la figura 1.1 se muestra un ejemplo con dos grupos, donde a cada grupo se le ajusta una distribución normal mostrando donde podría ser el punto óptimo de corte para la toma de una decisión

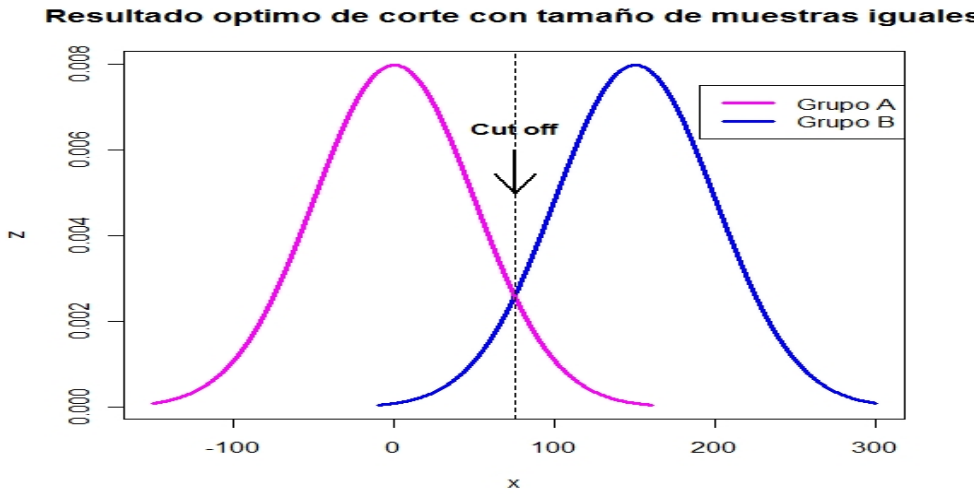

Figura 1.1: Resultado Óptimo(*Cutt Off*)

Estas evaluaciones estaban basadas en la experiencia de gerentes de crédito y no solo se tomaba en cuenta la información histórica sino que tratan se realizar proyecciones de la probable situación futura de los prestatarios y de su capacidad de pago del crédito. La recolección de información estadística sobre los individuos va reemplazando a la experiencia humana en la valuación crediticia. Si bien las técnicas de *scoring* no proyectan hacia delante la situación de los clientes, sino se basan en información histórica, presentan una gran ventaja en términos de costo para la evaluación de créditos minoristas principalmente para consumo y de tarjetas de crédito. Esto no quiere decir que la evaluación por parte de los funcionarios de las instituciones financieras hallan desaparecido si no que las técnicas de *"Credit Scoring"* se usan para los productos de alto volumen .

Hay veces que se dice que el *Credit Scoring* evalúa la solvencia del prestatario, pero no es así, ya que la solvencia no es atributo de los individuos como los ingresos. La solvencia es una evaluación realizada por una entidad crediticia de un prestatario que refleja las circunstancias de ambos. Además de que mediante ésta el prestamista ve al prestatario desde distintos posibles escenarios económicos, así los prestamistas evalúan a algunos prestatarios como solventes y a otros no.

Un prestamista debe tomar dos tipos de decisiones, primero sobre la concesión del

crédito a un nuevo solicitante y segundo sobre quiénes son los candidatos para otorgar el crédito; incluyendo la posibilidad de incrementar sus limites de crédito, analizar la rentabilidad de los clientes, monitorear el riesgo, detectar posibles problemas de cobranza, entre otras. En este trabajo en lo que nos enfocaremos es en revisar cómo tomar la primera decisión. Es decir, queremos saber a quién otorgarle un crédito; mediante técnicas estadísticas analizaremos variables para poder rescatar dos clases y así poder tomar una decisión acerca de a qué individuos se les debe otorgar el crédito y a quién no. Bajo estos métodos la finalidad es la clasificación "correcta" de los individuos, no estamos buscando el *score*, pero si podemos recuperar las clases después se puede crear el *score* y así, cuando llegue un nuevo individuo, podamos determinar si se le debe otorgar el crédito o no se le debe conceder.

El planteamiento de un análisis de *Credit Scoring* puede resumirse de la siguiente manera: el objetivo de la entidad financiera o del prestamista es maximizar el beneficio derivado de la concesión de un volumen de crédito, lo cual no necesariamente tiene que estar relacionado con el riesgo. Es decir, que un solicitante de crédito presente cierto riesgo no necesariamente implica que no conviene otorgarle financiamiento. Por ejemplo, un cliente de una entidad financiera que se financia con tarjeta de crédito y que es poco riesgoso, le dará más rentabilidad a la entidad o al prestamista que alguien que no es para nada riesgoso y más aún si nunca se financia con la tarjeta. Por lo tanto el problema se reduce a decidir si se concede o no se concede cada crédito individual sobre la base del rendimiento esperado de los solicitantes de distinto tipo de riesgo. Para optimizar esta decisión el prestamista debe conocer la probabilidad que tiene cada solicitante de presentar problemas de incumplimiento.

Un problema importante procede de la naturaleza de las muestras disponibles, debido a que los prestamistas no conceden todos los créditos solicitados, sino que sólo le conceden un crédito a los individuos que cumplen con un criterio de selección determinado. Por lo tanto una muestra de créditos concedidos es una muestra con truncamiento. Para obtener estimaciones óptimas de las probabilidades de incumplimiento se necesita disponer de una muestra censurada que contenga información tanto de los individuos que han recibido el crédito como de los que no han recibido.

1.2. Aspectos Históricos

¿Y cómo surgió el *Credit Scoring*? Mientras que la historia del crédito comienza desde hace 500 años, la historia del *Credit Scoring* es sólo desde hace un poco más de 50 años; de primera instancia el *Credit Scoring* se crea para identificar grupos en una cierta población, cuando no se podían ver las características que definen a un grupo, esto fue lo que llevó a tener una primera aproximación a la solución de este problema. Fue introducido por Fisher(1936); en 1941, Durand fue el primero en reconocer que no

podía usar la misma técnica de discriminante entre buenos y malos préstamos. Durante la década de 1930 algunas empresas introdujeron sistemas numéricos de puntuación para tratar de superar las inconsistencias de decisiones de crédito a través de análisis de créditos.

Los primeros sistemas de *scoring* se desarrollaron en la década de 1950 mediante la implementación de *scores* internos de comportamiento por parte de los bancos pioneros en Estados Unidos, los cuales eran utilizados por la gestión de las cuentas de dichos bancos en base a la propia información que manejaba. Posteriormente surgieron los *scorings* de aceptación conforme a los cuales se ranqueaba a los solicitantes de crédito según la propia calificación de buenos o malos que efectuaba una determinada institución bancaria.

No pasó mucho tiempo después de que terminó la guerra para contactar a la automatización de las decisiones de crédito y a las técnicas de clasificación desarrolladas en estadística y ver la ventaja de usar modelos estadísticos en decisiones de préstamos.

La llegada de las tarjetas de crédito a fines de la decada de 1960 ayudo a los bancos y a otros emisores de tarjetas de crédito a ver las utilidades del *Credit Scoring*. La única oposición vino de los qué como Capón (1982) argumentaban que el empirismo de la fuerza bruta del *Credit Scoring* atentaba contra las tradiciones de nuestra sociedad. El creía que debería haber una mayor dependencia de la historia del crédito y debía ser posible explicar por qué algunas características eran necesarias en un sistema de puntaje y en otros no. En la década de 1980 el éxito del *Credit Scoring* en las tarjetas de crédito significó para los bancos empezar a utilizar *scoring* en otros productos, como préstamos personales mientras que en los últimos años el *scoring* ha sido utilizado para préstamos hipotecarios y préstamos para pequeñas empresas. [Thomas, 2002]

La utilización del *Credit Scoring* para la evaluación de riesgo de crédito y ordenar a los deudores y solicitantes en función de su riesgo e incumplimiento comenzó en la decada de 1970 pero se generalizó a partir de 1990. Ésto se ha debido tanto al desarrollo de mejores recursos estadísticos y computacionales, como a los avances de la informática y de otras técnicas que permitieron que se trataran de construir *scorecards*[2]. En la década de 1980 la regresión logística y la programación lineal, dos principales baluartes de las tarjetas de hoy en día, se habían introducido. Más recientemente las técnicas de inteligencia artificial como los sistemas expertos y las redes neuronales se han puesto a prueba.

En la actualidad hay diversas metodologías de cómo evaluar el riesgo crediticio o la conveniencia de otorgar un crédito como son: análisis discriminante, regresión lineal,

[2]Plantillas o programas para asignar un *score* o *rating*

regresión logística, modelos probit, modelos logit, modelos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, algoritmo de particionamiento recursivo (árboles de decisión), sistemas expertos, algoritmos genéticos, redes neuronales y finalmente el juicio humano. La literatura sugiere que todos los métodos arrojan resultados similares, por lo que la conveniencia de usar uno u otro depende de las características particulares de cada caso.

Entre todas las metodologías disponibles, los modelos probit junto con las regresiones lineales y logísticas, el análisis discriminate y los árboles de decisión, se encuentran entre los métodos más usados por la industria para confeccionar estos modelos. Aunque estos métodos son los más utilizados frecuentemente se emplean de manera combinada. Estos métodos no se utilizan de manera autoritaria para aceptar o rechazar una solicitud de crédito, si no que sus resultados se combinan con revisiones posteriores. En otros casos, previo al cálculo del *score*, se aplican filtros que acotan el universo de solicitantes a ser evaluados con estos modelos. En ocasiones se combinan diversas metodologías como por ejemplo en los árboles de regresión: a través de un árbol se segmenta la muestra de deudores y luego a los deudores de cada segmento se les estima una regresión logística o modelo probit con distintas características.

1.3. Métodos estadísticos

En las décadas de 1950 y 1960, los únicos métodos de *scoring* utilizados eran los estadísticos: discriminación estadística y métodos de clasificación. Aún hoy son los más comunes. Su principal ventaja radica en que permiten utilizar las propiedades de los estimadores y las herramientas de los intervalos de confianza y las pruebas de hipótesis.

Estos métodos permiten discriminar la importancia relativa de las diferentes variables del *scorecard*. Así, es posible identificar y descartar características irrelevantes y asegurar que las más importantes sean tenidas en cuenta a la hora de tomar una decisión. Los procedimientos estadísticos más ampliamente utilizados para la estimación de las probabilidades de incumplimiento es el análisis discriminante.

1.3.1. Paramétricos

Análisis discriminante lineal: La primera técnica utilizada fue el análisis discriminante lineal, basado en el trabajo de Fisher (1936). Este procedimiento tiene como objetivo clasificar a los individuos en dos grupos, aquellos de los que se espera que devuelvan el crédito y aquellos de los que no. En particular si Y es una variable aleatoria discreta y X es un vector de variable explicativas, el análisis discriminante se basa en la distribución condicional de X en Y bajo el supuesto de que la distribución $X|Y$ es normal. Puede considerarse como una forma de regresión lineal, lo que posteriormente

llevó a la investigación de otras formas de regresión con supuestos menos restrictivos.

Entre estos, el método más exitoso y común es la regresión logística.

Regresión logística: Los modelos de regresión logística son modelos estadísticos multivariantes tanto explicativos como predictivos en los que se desea conocer la relación entre una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial) y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

1.3.2. No paramétricos

Árboles de clasificación: Otra técnica muy utilizada en los últimos 20 años ha sido la "partición recursiva" o "árboles de clasificación". Es un método no paramétrico que no requiere supuestos distribucionales, permite detectar interacciones, modela relaciones no lineales y no es sensible a la presencia de datos faltantes((Breiman, Friedman, Olshen Stone 1984), (Kass 1980)). Su principio básico es generar particiones recursivas por reglas de clasificación hasta llegar a una clasificación final, tal que es posible identificar perfiles (nodos terminales) en los que la proporción de clientes malos es muy alta (o baja) y de esta forma asignar su probabilidad.

En este método, se segmenta el conjunto de postulantes en diferentes subgrupos en función de sus atributos. Luego, se clasifica a cada subgrupo en "satisfactorio" o "no satisfactorio". Además de que ofrece una ventaja fundamental al ser un método de fácil entendimiento para personas que no cuentan con conocimientos avanzados de estadística. Un mismo modelo permite hacer diferentes usos, como mantenimiento de clientes considerdos como buenos (probabilidades bajas de incumplimiento), cobranza proactiva y discriminada por nivel de riesgo para los clientes considerados como malos o con probabilidades altas de llegar a incumplimiento.

Aquí hay distintas formas de segmentar. Las más comunes son el estadístico de Kolmogorov-Smirnov, el índice básico de impureza, el índice de Gini, el índice de entropía y la maximización de la media suma de cuadrados.

K - vecinos más cercanos: El método *KNN* (K nearest neighbors Fix y Hodges, 1951) es un método de clasificación supervisada que sirve para estimar la función de densidad $F(x/C_j)$ de las predictoras x por cada clase C_j.

Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la

distribución de las variables predictoras.

KNN es un algoritmo algo diferente de los demás en el sentido de que los propios datos proporcionan el "modelo". Para predecir un nuevo registro se encuentran los vecinos mas próximos al calcular la distancia euclidiana y luego realizando una media ponderada o mayoría de votos para obtener la predicción final.

Learning Vector Quantization: Esta red es un híbrido que emplea tanto aprendizaje no supervisado, como aprendizaje supervisado para clasificación de patrones. El aprendizaje por cuantificación vectorial (LVQ) es una técnica usada para reducir la dimensionalidad de los datos, supone una extensión del aprendizaje competitivo donde los prototipos están etiquetados y las clases están predefinidas. El objetivo es determinar un conjunto de prototipos que representan lo mejor de cada clase. Ahora, además de considerar la cercanía de un prototipo se puede evaluar la clase de éste e imponer, por lo tanto, correcciones de acercamiento o alejamiento.

Es una versión bajo la supervisión de la cuantización de vectores que se pueden utilizar cuando se han etiquetado los datos de entrada.

1.4. Métodos no estadísticos

Los sistemas expertos: En la década de 1970, se realizaron diversas investigaciones en el área de inteligencia artificial, intentando programar computadoras para que replicaran habilidades humanas. En este marco, uno de los intentos más exitosos fue el de los "sistemas expertos". En ellos, se daba a la computadora una base de datos y un mecanismo para generar reglas. Luego, el programa utilizaba esta combinación para analizar nuevas situaciones y encontrar formas de tratarlas para imitar la manera en que los expertos tomaban decisiones. Se construyeron sistemas pilotos para diagnósticos médicos. Y, como esto es esencialmente un problema de clasificación, se aplicaron estas ideas en el *Credit Scoring*.

Las redes neuronales: En los años 1980, otra variante de inteligencia artificial recibió atención: las redes neuronales. Las redes neuronales son formas de modelar el proceso de decisión como un sistema de unidades de procesamiento conectadas entre sí. Aquí, cada unidad recibe un input y produce un output. Así, si los inputs son las características del cliente y el output es su desempeño crediticio, vemos claramente la aplicación en el mundo del *scoring*.

Los algoritmos genéticos: Una forma alternativa de pensar en el *Credit Scoring* es la siguiente. Por un lado, tenemos una cierta cantidad de parámetros (por ejemplo, los posibles puntajes dados a diversos atributos). Por el otro, tenemos una manera de

medir qué tan bueno es el conjunto de parámetros (por ejemplo, el error de clasificación cuando un *scorecard* es aplicado a una muestra de antiguos clientes).

Así, con un procedimiento sistemático de búsqueda a través de la población de posibles soluciones, podremos encontrar aquella que resulte más próxima a resolver el problema de optimización. En términos muy sencillos, esta es la idea de los algoritmos genéticos propuesta por Holland (1975).

Programación lineal: Freed & Glover (1981) descubrieron que encontrar la función lineal de las características que mejor discrimina entre grupos puede plantearse como un problema de programación lineal. Es una técnica para gran cantidad de variables, sin embargo es de difícil comprensión.

El enfoque de la programación lineal mide la bondad y ajuste tomando la suma de los errores absolutos. Si uno quiere tomar el número de casos en los cuales la discriminación es incorrecta como medida de bondad y ajuste, entonces se debe introducir variables enteras en el programa lineal, y esto lleva a modelos de programación entera.

1.5. Métodos a explorar

Nosotros exploraremos principalmente cuatro técnicas: La primera es el Análisis Discriminante clásico debido a Fisher (1936). El segundo enfoque más flexible está basado en los modelos con variables dependientes cualitativas, de los cuales el más utilizado es el modelo de Regresión Logística. Estudiaremos estas técnicas no por que sean las mejores ni las más eficaces, si no por que son los más comunes y las más utilizadas en el entorno del *Credit Scoring*. Así mismo buscamos comparar los métodos más comunes con métodos de caracter "automatizado", sencillos para evaluar su mejoría con respecto a la clasificación, si existe.

Por esta razón empleamos dos metodologías más, así el tercer enfoque es un método no paramétrico automatizado "*K- nearest neighbor*" (KNN) que utiliza variables cuantitativas y debido a que utiliza la información proporcionada por cada solicitante para realizar el modelo resulta interesante observar si KNN logra rescatar con mayor facilidad el comportamiento de las clases para después poder construir el *score* mediante la información proporcionada. El cuarto enfoque es de la misma manera un método no paramétrico "*Learning Vector Quantization*" (LVQ), este tiene una visión un poco más completa en cuanto a la automatización ya que mediante un conjunto de inicialización de vectores localiza los individuos que deberán pertenecer a cada clase. Es un método que busca reducir la dimensionalidad de las bases de datos.

Además de que ambos métodos funcionan bien para bases de gran dimensión así como

para problemas de difícil decisión como puede ser el caso de *Credit Scoring* para la toma de una decisión correcta.

Capítulo 2

Algunos métodos para determinar un "credit scoring"

En este capítulo trataremos de explicar brevemente algunas técnicas estadísticas que se utilizan para la construcción de *Credit Scoring*, estas técnicas estadísticas permiten identificar y eliminar las características de poca importancia y garantizar que todas las características importantes permanezcan en el individuo. En la actualidad hay diversas técnicas que se utilizan para evaluar el riesgo crediticio para otorgar un crédito como son: análisis discriminate, regresión lineal, regresión logística, modelos probit, modelos logit, modelos no paramétricos de suavizado, métodos de programación matemática, algoritmo de particionamiento recursivo (árboles de decisión), entre otras.

Describiremos brevemente algunas de estas técnicas, comenzaremos con análisis discriminante para después continuar con regresión logística.

2.1. Análisis Discriminante

Es una técnica estadística multivariante cuya finalidad es analizar si existen diferencias significativas entre grupos de objetos respecto a un conjunto de variables medidas sobre los mismos para, en el caso de que existan, explicar en qué sentido se dan y proporcionar procedimientos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos analizados.

Esta técnica permite estudiar las diferencias entre dos o más grupos de individuos definidos a *priori* con respecto a una o varias variables. Dicha técnica equivale a un análisis de regresión donde la variable dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos, y las variables independientes son continuas y determinan a que grupos pertenecen los individuos.

Un segundo objetivo es construir una regla de decisión que asigne un objeto nuevo, que no sabemos clasificar previamente, a uno de los grupos prefijados con un cierto grado de riesgo, de acuerdo a las variables medidas.

Es necesario considerar una serie de supuestos y restricciones que utiliza esta técnica:

- Se tiene una variable categórica y el resto de variables son independientes respecto de ella.
- Es necesario que existan al menos dos grupos y para cada grupo se necesitan dos o más casos.
- El número de variables discriminantes debe ser menor que el número de objetos menos 2: $x_1, x_2, \ldots, x_p$, donde $p < (n-2)$ y n es el número de objetos.
- Ninguna variable discriminante puede ser combinación lineal de otras variables discriminantes.
- El número máximo de funciones discriminantes es igual al mínimo entre el número de variables y el número de grupos menos 1 (con q grupos, (q - 1)funciones discriminantes).
- Las matrices de covarianzas dentro de cada grupo deben de ser aproximadamente iguales.

Como nosotros estamos estudiando el proceso de otorgamiento de un crédito, éste se lleva a cabo mediante la elección entre dos acciones: dar el crédito a un nuevo solicitante o rechazar el crédito a dicho solicitante, debido a esto aplicaremos el análisis discriminante entre dos grupos ó mejor conocido como análisis discriminante simple.

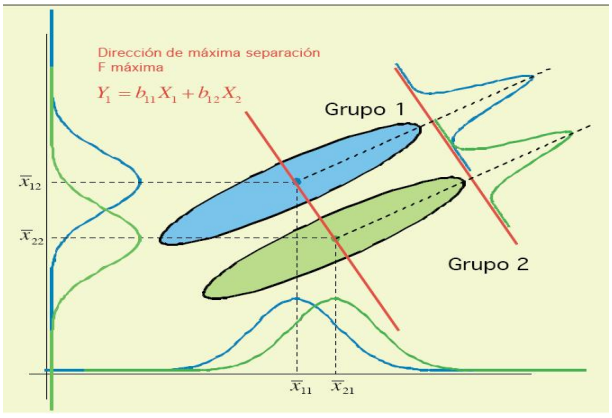

Figura 2.1: Diferencia Multivariada entre grupos

En la figura 2.1 se aprecian dos grupos para los cuales la distribución normal se ajusta de tal manera que se puede obtener una clasificación apropiada debido a la diferencia multivariante que existe entre los grupos.

En esta sección exploraremos esta técnica de análisis multivariado desde diferentes enfoques, se comienza describiendo el enfoque que utilizó Fisher para el caso de k grupos, y a su vez se desglosa en otros dos enfoques, el primer enfoque basado en el trabajo de Fisher (1936) trata de encontrar una función que mejor discrimine los dos grupos entre los individuos que tienen un buen historial crediticio y otro con clientes que tienen un historial malo de crédito; el segundo enfoque se refiere a la teoría de decisión donde se busca encontrar la regla que minimice el costo esperado para decidir si aceptar un nuevo cliente o rechazarlo.

2.1.1. Clasificación de grupos

A partir de q grupos donde se asigna a una serie de objetos y de p variables medidas sobre cada grupo $(x_1, x_2, \ldots, x_p)$ donde x_i es un vector de dimensión n; se trata de obtener para cada objeto una serie de puntuaciones que nos indique al grupo al que pertenecen tal que $(y_1, y_2, \ldots, y_m)$ sean funciones lineales de las variables observadas, es decir,

$$\begin{aligned} y_1 &= \omega_{10} + \omega_{11}x_1 + \ldots + \omega_{1p}x_p \\ y_2 &= \omega_{20} + \omega_{21}x_1 + \ldots + \omega_{2p}x_p \\ \vdots & \qquad\qquad \vdots \\ y_m &= \omega_{m0} + \omega_{m1}x_1 + \ldots + \omega_{mp}x_p \end{aligned} \tag{2.1}$$

Donde $m = \text{mín}(q-1, p)$ mismos que discriminen lo mejor posible a los q grupos.

Descomposición de la varianza

La variabilidad total de la muestra se puede descomponer en la variabilidad *dentro* de los grupos y *entre* los grupos, esto es necesario para poder realizar la discriminación de los q grupos, entonces:

Sabemos por definición que:

$$Cov(x_j, x_j^T) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T \tag{2.2}$$

Se puede considerar la media de la variable x_j para $j = 1, \ldots, p$ en cada uno de los grupos $I_1, I_2, \ldots, I_q$ es decir:

$$\bar{x}_{kj} = \frac{1}{n_k} \sum_{i \epsilon I_k} x_{ij} \qquad \Rightarrow \qquad \sum_{i \epsilon I_k} x_{ij} = n_k \bar{x}_{kj}$$

para $k = 1, \ldots, q$,

entonces la media total de la variable x_j se puede expresar como función de las medias de cada grupo.

$$\begin{aligned} \bar{x}_j &= \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \epsilon I_k} x_{ij} \\ &= \frac{1}{n} \sum_{k=1}^{q} n_k \bar{x}_{kj} = \sum_{k=1}^{q} \frac{n_k}{n} \bar{x}_{kj} \end{aligned} \tag{2.3}$$

Entonces (2.2) nos queda de la siguiente manera

$$Cov(x_j, x_j^T) = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \epsilon I_k} (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T \tag{2.4}$$

Sumando algunos "unos" adecuados en cada uno de los términos, se obtiene:

$$\begin{aligned} Cov(x_j, x_j^T) &= \frac{1}{n} \sum_{k=1}^{q} \sum_{i \epsilon I_k} (x_{ij} - \bar{x}_j + \bar{x}_{kj} - \bar{x}_{kj})(x_{ij}^T - \bar{x}_{kj}^T + \bar{x}_{kj}^T - \bar{x}_j^T) \\ &= \frac{1}{n} \sum_{k=1}^{q} \sum_{i \epsilon I_k} \Big((x_{ij} - \bar{x}_{kj}) + (\bar{x}_{kj} - \bar{x}_j) \Big) \Big((x_{ij} - \bar{x}_{kj})^T + (\bar{x}_{kj} - \bar{x}_j)^T \Big) \\ &= \frac{1}{n} \sum_{k=1}^{q} \sum_{i \epsilon I_k} \Big((x_{ij} - \bar{x}_{kj})(x_{ij} - \bar{x}_{kj})^T + (x_{ij} - \bar{x}_{kj})(\bar{x}_{kj} - \bar{x}_j)^T \\ &+ (\bar{x}_{kj} - \bar{x}_j)(x_{ij} - \bar{x}_{kj})^T + (\bar{x}_{kj} - \bar{x}_j)(\bar{x}_{kj} - \bar{x}_j)^T \Big) \end{aligned}$$

Obsérvese que el segundo y tercer producto son iguales a cero. Analicemos solamente el segundo término ya que el tercer producto es el transpuesto del segundo; por lo tanto

llegaremos a lo mismo, veamos:

$$\begin{aligned}
& \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}(x_{ij}-\bar{x}_{kj})(\bar{x}_{kj}-\bar{x}_j)^T \\
= & \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}\left(x_{ij}\bar{x}_{kj}^T - x_{ij}\bar{x}_j^T - \bar{x}_{kj}\bar{x}_{kj}^T + \bar{x}_{kj}\bar{x}_j^T\right) \\
= & \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}x_{ij}\bar{x}_{kj}^T - \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}x_{ij}\bar{x}_j^T - \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}\bar{x}_{kj}\bar{x}_{kj}^T + \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}\bar{x}_{kj}\bar{x}_j^T \\
= & \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_{kj}^T - \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_j^T - \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_{kj}^T + \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_j \\
= & \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_{kj}^T - \frac{1}{n}\sum_{k=1}^{q}n_k\bar{x}_{kj}\bar{x}_{kj}^T - \bar{x}_j\bar{x}_j^T + \bar{x}_j\bar{x}_j^T \quad = \quad 0
\end{aligned}$$

Como vimos el segundo y el tercer producto son cero entonces sólo nos quedan el primer producto y el último, es decir;

$$\begin{aligned}
= & \frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}(x_{ij}-\bar{x}_{kj})(x_{ij}-\bar{x}_{kj})^T + \sum_{k=1}^{q}\sum_{i\epsilon I_k}\frac{1}{n}(\bar{x}_{kj}-\bar{x}_j)(\bar{x}_{kj}-\bar{x}_j)^T \\
= & \underbrace{\frac{1}{n}\sum_{k=1}^{q}\sum_{i\epsilon I_k}(x_{ij}-\bar{x}_{kj})(x_{ij}-\bar{x}_{kj})^T}_{d(x_j,x_j^T)} + \underbrace{\sum_{k=1}^{q}\frac{n_k}{n}(\bar{x}_{kj}-\bar{x}_j)(\bar{x}_{kj}-\bar{x}_j)^T}_{e(x_j,x_j^T)} \\
= & d(x_j,x_j^T) + e(x_j,x_j^T)
\end{aligned}$$

Por lo tanto la covarianza *total* es igual a la covarianza *dentro* ($d(x_j, x_j^T)$) de grupos más la covarianza *entre* ($e(x_j, x_j^T)$) grupos.

Módelo de Clasificación

Comencemos buscando un vector ω, talque cuando proyectemos los puntos sobre él, se obtenga la máxima variabilidad *entre* los grupos en relación a la variabilidad *dentro* de los grupos.

En notación de matrices tenemos:

$$T = E + D$$

donde

T= matriz de covarianzas *total*
E= matriz de covarianzas *entre* grupos
D= matriz de covarianzas *dentro* de grupos

Entonces el cociente que tenemos que maximizar es el siguiente:

$$\lambda = \frac{\omega^T E \omega}{\omega^T D \omega} \tag{2.5}$$

Esto puede ser escrito de la siguiente forma:

$$\omega^T E \omega = \lambda \omega^T D \omega$$

$$\omega^T (E\omega - \lambda D\omega) = 0 \tag{2.6}$$

Examinemos los λ y ω que son soluciones de (2.6), determinemos los valores de ω que maximize a λ. La solución $\omega^T = 0$ no es solución ya que daría $\lambda = \frac{0}{0}$. Otras posibles soluciones se encuentran por medio de:

$$E\omega - \lambda D\omega = 0 \qquad \Rightarrow \qquad (D^{-1}E - \lambda I)\omega = 0 \tag{2.7}$$

Observemos que la matriz $D^{-1}E$ no es simétrica. Por otro lado sabemos que los eigen valores de la matriz simétrica $(U^{-1})^T E U^{-1}$ son los mismos para la $D^{-1}E$ donde $D = U^T U$ que es la factorización de Cholesky.

Entonces si b es un eigen-vector de $(U^{-1})^T E U^{-1}$, entonces $\omega = U^{-1}b$ es un eigen-vector de $D^{-1}E$.

Por lo tanto, el eigen-vector asociado a la primera función disciminante es de la matriz $D^{-1}E$. Luego si tomo el vector asociado al eigen-valor más grande, se obtendrá la función que recoge el mejor poder de discriminación.

El eigen-valor asociado a la función discriminante indica la proporción de varianza total explicada por las m funciones discriminantes que recoge la variable y_i. Para obtener más funciones discriminantes, se siguen sacando los eigen-vectores de la matriz $(D^{-1}E)$ asociados a los eigen-valores elegidos en orden *decreciente* $(\lambda_1 > \lambda_2 > \cdots > \lambda_s)$.

La suma de todos los eigen-valores $\sum_{i=1}^{m} \lambda_i$, es la proporción de varianza total que queda explicada, o se conserva, al considerar las funciones discriminantes. Como consecuencia, el porcentaje explicado por y_i del total de la varianza explicada por $y_1, y_2, \ldots, y_m$ es:

$$\frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} * 100\,\%$$

Como la idea básica del análisis discriminante consiste en extraer a partir de $x_1, x_2, \ldots, x_p$ variables observadas en q grupos, m funciones $y_1, y_2, \ldots, y_m$ de forma que:

$$y_i = \omega_{i0} + \omega_{i1}x_1 + \omega_{i2}x_2 + \ldots + \omega_{ip}x_p$$

donde y_i es la función discriminante correspondiente al $i-ésimo$ eigen vector, tales que $corr(y_i, y_j) = 0$ para todo $i \neq j$. Las funciones $y_1, y_2, \ldots, y_m$ se extraen de modo que:

(i) y_1 sea la combinación lineal de $x_1, x_2, \ldots, x_p$ que proporciona la *mayor* discriminación posible entre los grupos.

(ii) y_2 sea la combinación lineal de $x_1, x_2, \ldots, x_p$ que proporciona la *mayor* discriminación posible entre los grupos, despues de y_1, tal que la $corr(y_i, y_2) = 0$.

En general, y_i es la combinación lineal de $x_1, x_2, \ldots, x_p$ que proporciona la discriminación posible *entre* los grupos despues de y_{i-1} tal que $corr(y_i, y_j) = 0$ para $j = 1, 2, \ldots, (i-1)$.

2.1.2. Enfoque de Fisher: Separación de dos grupos

El objetivo es encontrar una variable Y, combinación lineal de las variables observadas $Y = \omega^T X$ donde $X^T = (X_1, X_2, \ldots, X_p)$ el vector de variables observadas, tales que discriminen o separen lo máximo posible a los 2 grupos y que muestre las mayores diferencias entre las medias de los dos grupos de forma que nos permita decidir entre si le otorgamos el crédito a un solicitante o se lo rechazamos. Estas combinaciones lineales de las p variables deben maximizar la varianza *entre* los grupos y minimizar la varianza *dentro* de los dos grupos.

Bajo este enfoque se asume que los dos grupos tienen varianza común, es decir, Σ_1=Σ_2=Σ

Fisher sugirió encontrar una combinación lineal del vector X en cada población, de manera que sea máxima, la razón de la diferencia de medias de las combinaciones lineales respecto a la raíz cuadrada de su varianza, es decir, queremos encontrar un vector ω de pesos de manera que maximize el criterio.

$$\lambda = \frac{\text{La distancia entre la media muestral de dos grupos}}{(\text{varianza de la muestra de cada grupo})^{1/2}}$$

Queda expresado matemáticamente como sigue:

$$\lambda = \frac{\omega^T(\mu_2 - \mu_1)}{(\omega^T \Sigma \omega)^{1/2}}$$

Cabe señalar que la combinación lineal no es única, es decir, cualquier conjunto de coeficientes puede multiplicarse por cualquier constante, cY, sin que ocurra cambio alguno en λ.

En las aplicaciones los parámetros, usualmente no se conocen. Por lo tanto las muestras de las n_i observaciones de cada grupo, se utilizan para definir la regla basada en una muestra mediante la sustitución de μ_i por$\bar{X}_i$, el vector de medias estimada de cada grupo y Σ por S la varianza de la muestra.

Las medias de los valores de la nueva variable para cada grupo y la varianza son:

$$\bar{Y}_1 = \omega^T \bar{X}_1 \tag{2.8}$$

$$\bar{Y}_2 = \omega^T \bar{X}_2 \tag{2.9}$$

$$Var(Y) = \omega^T S \omega \tag{2.10}$$

Donde:

$$\bar{X}_i^T = (\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{ip}) \qquad \text{para} \qquad i = 1, 2 \qquad \text{y} \tag{2.11}$$

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \tag{2.12}$$

Entonces la correspondencia para la separación queda:

$$\lambda = \frac{\omega^T(\bar{X}_2 - \bar{X}_1)}{(\omega^T S \omega)^{1/2}} \tag{2.13}$$

Podemos demostrar que el vector de pesos que maximiza (2.13) esta dado por:

$$\omega^T \alpha(\bar{X}_2 - \bar{X}_1)^T S^{-1} \tag{2.14}$$

Este es el mejor vector de pesos que separa los grupos, es decir, el que separa a los solicitantes que se les otorgará el crédito de los solicitantes que no se les otorgará dicho crédito. Para este criterio no importa la distribución. Es válido para todas las distribuciones, ya que (2.13) involucra solo la media y la varianza.

Reglas de Clasificación

Sabemos que (2.13) nos indica cómo separar o discriminar un grupo de otro, así como el mejor vector que hace mayor dicha discriminación. Pero si tenemos un nuevo solicitante de crédito ¿Como decidimos si otorgarle el crédito a dicho solicitante ó negárselo?

De la función discriminante lineal de Fisher se obtiene la siguiente regla de asociación que asigna a los solicitantes de crédito a alguno de los dos grupos como sigue: Si G_1 es

el grupo que contiene a los solicitantes que se les concederá el crédito y G_2 el grupo de los solicitantes a los cuales se les negará el crédito. Entonces asignaremos a G_1 si los *scores* del nuevo solicitante distan menos de la media de G_1 que de la media de G_2, es decir,
Si

$$\begin{aligned} | Y - \bar{Y}_1 | &\leq | Y - \bar{Y}_2 | \quad \Rightarrow \\ | \omega^T X - \omega^T \bar{X}_1 | &\leq | \omega^T X - \omega^T \bar{X}_2 | \quad \Rightarrow \\ | \omega^T (X - \bar{X}_1) | &\leq | \omega^T (X - \bar{X}_2) | \end{aligned} \tag{2.15}$$

Por lo que si

$$| \omega^T (X - \bar{X}_1) | > | \omega^T (X - \bar{X}_2) | \tag{2.16}$$

Entonces asigno a G_2 [Dillon, 1984]

2.1.3. Discriminante basado en Distribuciones de Probabilidad

Sean P_1 y P_2 dos poblaciones donde se define una variable aleatoria X . Supondremos tambien que X es absolutamente continua y que las funciones de densidad de las poblaciones, f_1 y f_2 son conocidas. El problema puede enfocarse desde el punto de vista de la Teoría de decisión incluyendo además probabilidades a *priori*:
Consideremos las siguientes hipótesis:

(i) Las probabilidades a *priori* de que un individuo tomado al azar provenga de cada población son conocidas como: $\pi_1 + \pi_2 = 1$.

(ii) Las consecuencias asociadas a los errores de clasificación son: $c(2|1)$ y $c(1|2)$, donde $c(i|j)$ es el costo de clasificar en P_i de un objeto que pertenece realmente a P_j. Estos costos se suponen conocidos.

Las posibles decisiones en el problema son unicamente dos: asignar en P_1 ó en P_2.

El problema es separar el espacio muestral Ω en dos regiones R_1 y $R_2 = \Omega - R_1$. Una vez observado el dato podemos calcular:

La probabilidad de clasificar en π_i si viene de π_i

$$P(i|i) = \int_{R_i} f_i(x)dx$$

La probabilidad de clasificar en π_i si viene de π_j

$$P(i|j) = \int_{R_i} f_j(x)dx$$

Asignaremos el elemento al grupo 2 si su costo esperado es menor, es decir:

$$\frac{f_2(x)\pi_2}{c(2|1)} > \frac{f_1(x)\pi_1}{c(1|2)} \tag{2.17}$$

que indica que, siendo iguales el resto de términos, clasificaremos en la población P_2 si:

(i) La probabilidad a priori es más alta

(ii) La verosimilitud de que provenga de P_2 es más alta

(iii) El costo de equivocarnos al clasificarlo en P_2 en más bajo

Analicemos este enfoque suponiendo que la función de densidad es una normal multivariada con matrices de covarianza común y después con matrices de covarianza diferentes.

Normal Multivariada con matrices de Covarianza común

Vamos a aplicar el análisis suponiendo que las funciones de densidad son normales con distintos vectores de medias pero identica matriz de varianza, es decir,

$$X \sim N(\mu_1, \Sigma) \tag{2.18}$$
$$X \sim N(\mu_2, \Sigma) \tag{2.19}$$

El objetivo es hallar las dos regiones R_1 y R_2.

Entonces se trata de asignar la observación a aquella población que tenga la verosimilitud más alta. Por lo que la correspondiente función de densidad en este caso:

$$f_i(x) = (2\pi)^{-p/2}|\Sigma|^{-1/2} exp\Big(\frac{-(x-\mu_i)^{T-1}(x-\mu_i)}{2}\Big)$$

Usando (2.17), como ambos términos son siempre positivos, obtenemos lo siguiente:

$$\frac{f_2(x)\pi_2}{c(2|1)} > \frac{f_1(x)\pi_1}{c(1|2)} \qquad \Rightarrow \qquad \frac{f_2(x)}{f_1(x)} > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \tag{2.20}$$

sustituyendo las funciones en (2.20)

$$\frac{(2\pi)^{-p/2}|\Sigma|^{-1/2} exp\Big(\frac{-(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)}{2}\Big)}{(2\pi)^{-p/2}|\Sigma|^{-1/2} exp\Big(\frac{-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}{2}\Big)} > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \qquad \Rightarrow \tag{2.21}$$

por propiedades de exponencial

$$exp\Big(-\frac{1}{2}\Big[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\Big]\Big) > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \qquad \Rightarrow \ (2.22)$$

aplicando logaritmo, tenemos:

$$-\frac{1}{2}\Big[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\Big] > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow \quad (2.23)$$

desarrollando

$$-\frac{1}{2}\Big[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_2 - \mu_2^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2 - x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}x - \\ - \mu_1^T\Sigma^{-1}\mu_1\Big] > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow (2.24)$$

$$\frac{1}{2}\Big[x^T\Sigma^{-1}\mu_2 + \mu_2^T\Sigma^{-1}x - \mu_2^T\Sigma^{-1}\mu_2 - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1\Big] \\ > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \quad \Rightarrow \qquad (2.25)$$

$$\frac{1}{2}\Big[x^T\Sigma^{-1}(\mu_2-\mu_1) + (\mu_2-\mu_1)^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\Big] \\ > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow \qquad (2.26)$$

$$(\mu_2-\mu_1)^T\Sigma^{-1}x + \frac{\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2}{2} > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad (2.27)$$

$$(\mu_2-\mu_1)^T\Sigma^{-1}x - \frac{(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2+\mu_1)}{2} > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad (2.28)$$

entonces el coeficiente lineal es:

$$\omega^T = (\mu_2-\mu_1)^T\Sigma^{-1} \qquad (2.29)$$

Y suponiendo iguales los costos y las probabilidades a priori, es decir, $c(1|2) = c(2|1)$; $\pi_1 = \pi_2$, la regla resultante:

$$(\mu_2-\mu_1)^T\Sigma^{-1}x > \frac{(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2+\mu_1)}{2} \qquad (2.30)$$

Sin embargo, hay que tener encuenta que en *Credit Scoring* el costo de conceder un crédito a un candidato con mal riesgo de crédito es signaficativamente mayor que el

costo de denegar un crédito a un candidato con un buen riesgo de crédito. Si no se tiene información al respecto, los consideramos iguales.

Por lo tanto clasifico en P_2 si ocurre (2.30)

Observemos que el lado izquierdo de (2.30) es una suma ponderada de los valores de las variables, mientras que el lado derecho es una contaste. Lo anterior asume que las medias y las varianzas son conocidas, pero en la práctica esto rara vez ocurre, por lo que es necesario reemplazarlo por sus estimaciones, es decir μ_1 por $\bar{X}_1$, μ_2 por $\bar{X}_2$ y la Σ por S. Donde la media de la distribución se estima a partir de los vectores de medias muestrales y la matriz de covarianzas común a los dos grupos se les estima mediante la matriz de covarianza dentro de los grupos. Entonces (2.30) se convierte en:

$$(\bar{X}_2 - \bar{X}_1)^T S^{-1} x > \frac{(\bar{X}_2 - \bar{X}_1)^T S^{-1} (\bar{X}_2 + \bar{X}_1)}{2} \tag{2.31}$$

Veamos que el coeficiente lineal es el mismo que el obtenido en (2.14), solo que aqui ocupamos supuestos de Normalidad mientras que en (2.14) no nos importó la distribución sino solamente la media y la varianza.

Como supusimos que las dos poblaciones tienen distribuciones normales multivariantes, que la matriz de covarianza es la misma para los dos grupos y que las probabilidades a *priori* se estiman a partir de los datos muestrales como la proporción muestral en cada grupo, a este resultado se le conoce como **Discriminate Lineal** que coincide con el resultado de Fisher tratado en la subsección anterior.

Normal Multivariada con matrices de Covarianza diferentes

Otra de las limitaciones que es evidente en el caso anterior es que las matrices de covarianza son iguales para las dos poblaciones de los solicitantes, algo que en la práctica es difícil que ocurra, es decir, la matriz de covarianza de los solicitantes a los que sí se les otorgará el crédito generalmente es diferente que la matriz de covarianza de los que no se les concederá dicho crédito. Entonces para este caso supongamos que no se cumple la homogeneidad en las matrices de covarianza, es decir, $\Sigma_1 \neq \Sigma_2$, por lo que:

$$X_1 \sim N(\mu_1, \Sigma_1) \tag{2.32}$$
$$X_2 \sim N(\mu_2, \Sigma_2) \tag{2.33}$$

$$\frac{f_2(x)}{f_1(x)} > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \tag{2.34}$$

Recordemos la función Normal Multivariada que utilizaremos

$$f_i(x) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}exp\Big(\frac{-(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}{2}\Big)$$

Sustituyendo la función anterior en (2.34) obtenemos lo siguiente:

$$\frac{(2\pi)^{-p/2}|\Sigma_2|^{-1/2}exp\Big(\frac{-(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)}{2}\Big)}{(2\pi)^{-p/2}|\Sigma_1|^{-1/2}exp\Big(\frac{-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)}{2}\Big)} > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \qquad \Rightarrow \tag{2.35}$$

agrupando términos y por propiedades de exponencial tenemos los siguiente:

$$\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big)^{-1/2}exp\Big(-\frac{1}{2}\Big[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\Big]\Big) > \frac{\pi_1 c(2|1)}{\pi_2 c(1|2)} \qquad \Rightarrow \tag{2.36}$$

tomando logaritmos en ambos lados de la igualdad

$$ln\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big)^{-1/2}-\frac{1}{2}\Big[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\Big] > ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow \tag{2.37}$$

desarrollando y haciendo simplificaciones tenemos:

$$\begin{aligned} x^T\Sigma_1^{-1}x - x^T\Sigma_1^{-1}\mu_1 - \mu_1^T\Sigma_1^{-1}x + \mu_1^T\Sigma_1^{-1}\mu_1 - x^T\Sigma_2^{-1}x + x^T\Sigma_2^{-1}\mu_2 + \mu_2^T\Sigma_2^{-1}x - \\ - \mu_2^T\Sigma_2^{-1}\mu_2 > ln\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big) + 2ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow \end{aligned} \tag{2.38}$$

$$\begin{aligned} x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x - 2x^T(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_2^T\Sigma_2^{-1}\mu_2 \\ > ln\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big) + 2ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad \Rightarrow \end{aligned} \tag{2.39}$$

$$\begin{aligned} x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x - 2x^T(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) > \mu_2^T\Sigma_2^{-1}\mu_2 - \mu_1^T\Sigma_1^{-1}\mu_1 + \\ + ln\Big(\frac{|\Sigma_2|}{|\Sigma_1|}\Big) + 2ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \end{aligned} \tag{2.40}$$

Entonces del lado izquierdo se tiene una parte cuadrática de los valores mientras que del lado derecho es una constante. Esto pareciera ser una regla de decisión más general, uno podría esperar que sea mejor que la lineal, sin embargo existen también sus limitaciones. Por ejemplo, para este caso en la práctica uno tendría que estimar parámetros para la Σ_1 y la Σ_2, además es más eficiente para muestras grandes que para muestras pequeñas.

Como lo habíamos mencionado en la práctica es difícil conocer los parámetros por

lo que hay que estimarlos; entonces (2.40) en su versión muestral queda de la siguiente manera:

$$x^T(S_1^{-1} - S_2^{-1})x - 2x^T(S_1^{-1}\bar{X}_1 - S_2^{-1}\bar{X}_2) > \bar{X}_2^T S_2^{-1}\bar{X}_2 - \bar{X}_1^T S_1^{-1}\bar{X}_1 + \\ + ln\Big(\frac{|S_2|}{|S_1|}\Big) + 2ln\Big(\frac{\pi_1 c(2|1)}{\pi_2 c(1|2)}\Big) \qquad (2.41)$$

Por otro lado si suponemos que son iguales los costos y las probabilidades a priori, es decir, $c(1|2) = c(2|1)$; $\pi_1 = \pi_2$, la regla resultante es:

$$x^T(S_1^{-1} - S_2^{-1})x - 2x^T(S_1^{-1}\bar{X}_1 - S_2^{-1}\bar{X}_2) > \bar{X}_2^T S_2^{-1}\bar{X}_2 - \bar{X}_1^T S_1^{-1}\bar{X}_1 + ln\Big(\frac{|S_2|}{|S_1|}\Big) \qquad (2.42)$$

Cuando se supone que $\Sigma_1 \neq \Sigma_2$, como en este caso, recibe el nombre de **Discriminante Cuadrático**.

2.1.4. Pruebas de Hipótesis

Como hemos visto uno de los objetivos más fuertes del análisis discriminante es la determinación de la media de los grupos definidos a *priori* mediante la media de sus *scores* (características) así como, determinar si las matrices de covarianza para ambos grupos son iguales.

El valor medio de la función discriminante comúnmente se conoce como centroide. El centroide de un grupo denotado por $\bar{Y}_i$ donde i se utiliza para identificar el grupo que se esta estudiando, se obtiene aplicando el vector resultante del análisis discriminante al vector de *score* para cada grupo, es decir, $\bar{Y}_i = \hat{\omega}^T \bar{x}_i$. Por otro lado sabemos lo siguiente:

La distancia de Mahalanobis al cuadrado de un individuo al grupo i es la distancia al centroide del grupo, es decir:

$$D^2 = D^2(x, \bar{x}_i) = (x - \bar{x}_i)^T S^{-1}(x - \bar{x}_i)$$

Mientras que la distancia generalizada de Mahalonobis para dos grupos entre sus centroides es:

$$D^2 = D^2(\bar{x}_1, \bar{x}_2) = (\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 - \bar{x}_2)$$

Esta estadística se puede utilizar para determinar si las diferencias entre los grupos son estadísticamente significativas. Los valores grandes de la estadística nos llevaría a creer que los grupos están suficientemente separados en términos de su media. En estos casos es probable que nuevas observaciones puedan ser clasificadas con éxito, sobre la base de las características medidas. Sin embargo, una prueba estadística formal es la siguiente:

$$\mathbf{Z} = \frac{n_1 n_2}{n_1 + n_2} \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} D^2 \qquad (2.43)$$

Bajo la hipótesis $H_0 : \mu_1 = \mu_2$ y con matriz de varianza-covarianza común, el estadístico de prueba $\mathbf{Z}$ se distribuye F con p y $n_1 + n_2 - p - 1$ grados de libertad es decir:

$$\mathbf{Z} \sim F_{\xi_\alpha;(p,n_1+n_2-p-1)} \tag{2.44}$$

La prueba estadística de significancia de la diferencia entre los dos grupos tiene una relación con el enfoque adoptado en el problema de dos muestras univariadas para muestras independientes. Sin embargo, en este caso se considera el estadístico t al cuadrado expresado de la siguiente manera:

$$t^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_p^2[(n_1 + n_2)/n_1 n_2]}$$

Donde s_p^2 es la varianza ponderada de las muestras y

$$\bar{x}_1 = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}, \qquad \bar{x}_2 = \frac{\sum_{j=n_1+1}^{n_1+n_2} x_{2j}}{n_2}$$

Remplazando estos estimadores por los estimadores multivariados, resulta un estadístico conocido como la T^2 de Hotelling's para dos muestras.

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 - \bar{x}_2)$$

Observemos que lo anterior se puede expresar mediante la relación de la expresión (2.43) en terminos de D^2

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 \qquad \Rightarrow \qquad \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F_{\xi_\alpha;(p,n_1+n_2-p-1)}$$

Por lo tanto rechazamos H_0 con un cierto nivel de significancia deseado α si

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 > F_{\xi_\alpha;(p,n_1+n_2-p-1)} \tag{2.45}$$

Donde ξ_α,denota el cuantil de la función F [Dillon, 1984]

2.1.5. Ejemplos

A continuación mostraremos dos ejemplos de análisis discriminante, en los que observaremos la manera de clasificación, mediante el uso de rutinas de **R**.

Ejemplo 1

Consideremos la base de datos de 200 cangrejos, que se puede encontrar en el paquete MASS de **R** llamada "crabs". Esta base considera 6 variables como son:

sp = Especie ("B" ó "O"), es decir, Blue=Azul ó Orange=Naranja.

FL = Tamaño del lóbulo Frontal.

RW = Anchura Trasera.

CL = Longitud del Caparazón.

CW = Ancho del Caparazón.

BD = Profundidad del Cuerpo.

Observemos como se ven los datos en la Figura 2.2.

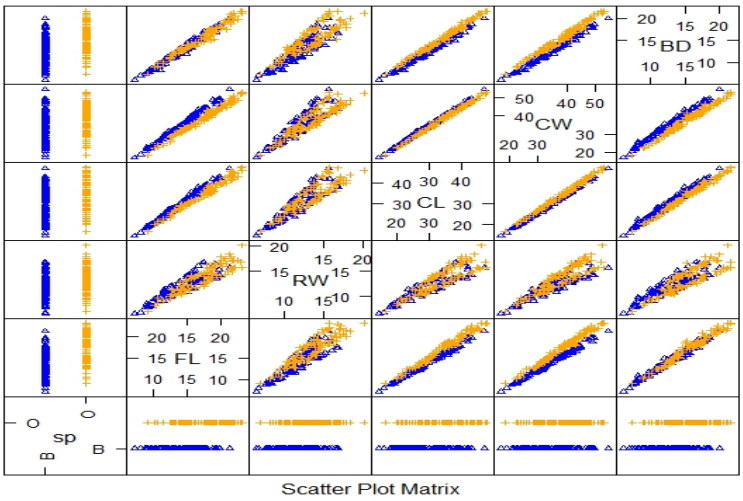

Figura 2.2: Diagrama de dispersión por grupos ("Azules" y "Naranjas")

Tomemos una submuestra aleatoria de 120 cangrejos, para despues clasificar los 80 cangrejos restantes y poder evaluar los errores de clasificación. Considerando que un cangrejo tiene la misma probabilidad de pertenecer a ambos grupos. Tratemos de ver si las variables son capaces de indicarnos a que especie pertenecen estos 80 cangrejos.

Para un primer modelo observemos qué ocurre si excluimos la variable FL, quitaremos esta variable sólo para observar la discriminación de los grupos aunque esta variable quiza nos aporte mayor información que RW, si observamos el gráfico 2.2 que muestra los grupos de clasificación podemos ver que RW es una de las variables en la que no se distingue una separación entre ambos cangrejos, pero quiza con la combinación de las restantes puede hacer un buena clasificación. Mediante la función de **R** obtenemos la siguiente función discriminante.

$$sp = 0.9065122CL + 0.4486312RW - 1.6489436CW + 1.8320112BD \tag{2.46}$$

Observemos que el coeficiente de RW es chico respecto a los coeficientes de las otras variables. Esto nos indica que esta variable es la que aporta menos información al momento de tomar la decisión de pertenecer a un grupo u otro.

Recordemos que estamos buscando la función que discrimine mejor a los grupos, es decir, una función que separe a los cangrejos azules de los cangrejos naranjas.

Las medias por cada variable y para cada grupo son la mostradas a continuación:

$\bar{X}_{ij}$				
$i \setminus j$	**CL**	**RW**	**CW**	**BD**
B	30.058	11.928	34.717	12.583
O	34.153	13.549	38.112	15.478

Donde:
$i = B, O$ son los cangrejos azules y los naranjas respectivamente.
$j = CL, RW, CW, BD$ variables utilizadas.

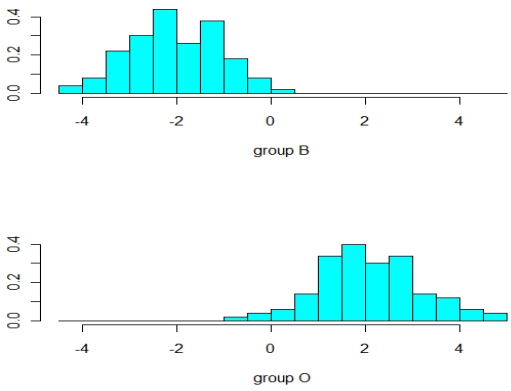

Figura 2.3: Histograma por grupo("Azules" y "Naranjas")

Observemos que en la Figura 2.3 para ambos grupos podemos ajustar una función Normal, lo cual nos indica que se ajustó una Función Discriminante Lineal, por medio de este gráfico podemos ver que tienen casi la misma varianza. Ahora veamos que tal clasificó a los 80 cangrejos que excluimos para hacer los grupos:

B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B B O O O O O O B O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O B B O O O O O O O O O

Veamos que lo primeros 40 cangrejos son azules, mientras que los siguientes 40 tenían que ser naranjas, y tuvo tres errores, clasifica tres naranjas como azules.

La función discriminante nos indica que la variable *RW* es la variable que menos información nos aporta al momento de hacer la separación, entre saber si los cangrejos pertenecen a la clase de los azules o de los naranjas.

A continuación presentamos el cuadro 2.1 en donde se ven los errores que se cometen utilizando distintos grupos de variables, así como los coeficientes que arroja la función discriminante de **R**.

Variables Utilizadas	**Coeficientes**	**Errores "B"**	**Errores "O"**
FL	1.5687028		
CL	0.3946165		
RW	0.1707172	0	0
CW	-1.5460297		
BD	1.3083372		
FL	1.7029509		
CL	0.2031467		
CW	-1.3899604	0	0
BD	1.3277983		
FL	1.725887		
CW	-1.262331	0	0
BD	1.434982		
FL	0.5709018		
BD	-0.2639611	12	15

Cuadro 2.1: Errores de clasificación con distintas variables

Obsérvese en el cuadro 2.1 que fuimos quitando la variable que menos información nos proporcionaba al momento de clasificar a nuestros cangrejos a alguno de nuestros grupos. Los discriminantes propuestos funcionaban bien, no tenían error de clasificación, con facilidad se podía decidir a que grupo pertenecia cada cangrejo. Sin embargo, al

excluir CW, nos indicó que ya no le era nada fácil clasificar. Debido a esta situación mostremos los gráficos del comportamiento de los grupos cuando tenemos a las 3 variables anteriormente señaladas (FL, CW, BD).

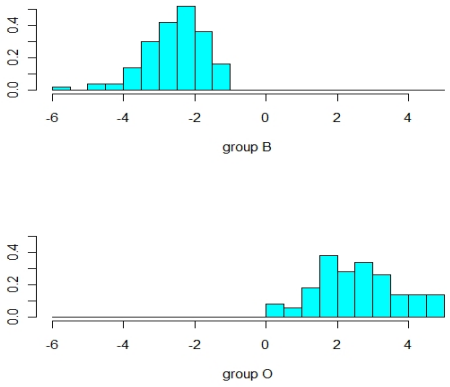

Figura 2.4: Histograma de las funciones Discriminantes con 3 variables("Azules" y "Naranjas")

En la Figura 2.4 se ve que los grupos no se empalman, hasta se ve claramente que uno está separado del otro. Ahora observemos cómo se ve cuando le quitamos CW.

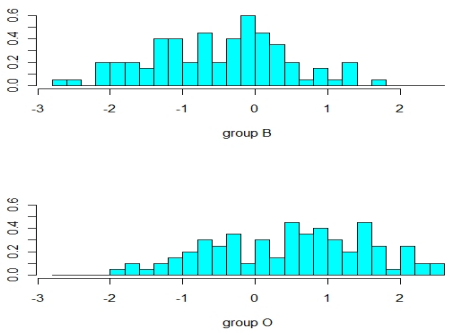

Figura 2.5: Al momento de quitar CW("Azules" y "Naranjas")

Es un cambio demasiado drástico en la Figura 2.5, con mucha dificultad podríamos decir a qué grupo pertenece uno de nuestros cangrejos. Por esto mismo es que presenta demasiados errores de clasificación.

Debido a esto, la función discriminante que mejor separa un grupo de otro es la mostrada a continuación:

$$sp = 1.725887FL - 1.262331CW + 1.3277983BD \tag{2.47}$$

En conclusión por todo lo anterior podemos decir que al menos necesita de FL, CW y BD para poder clasificar bien a cualquier cangrejo de alguna de estas especies, es decir, decidir si el cangrejo es azul o es naranja.

Ejemplo 2

Consideremos lo siguiente: Se midieron 6 variables en 100 billetes suizos auténticos y 100 falsos. Las variables son:

X_1 : Longitud del billete.

X_2 : Altura medida de lado izquierdo.

X_3 : Altura medida de lado derecho.

X_4 : Distancia del recuadro interno medida en la parte inferior.

X_5 : Distancia del recuadro interno medida en la parte superior.

X_6 : Longitud diagonal.

X_7 : Tipo ("A" o "F"), es decir, Auténtico=A, Falso=F.

Pretendemos ver si las variables son capaces de indicarnos si el billete es auténtico ó falso.

Figura 2.6: Billete Suizo

Presentamos el gráfico que nos muestra como se ven las dos poblaciones. Observese la Figura 2.7

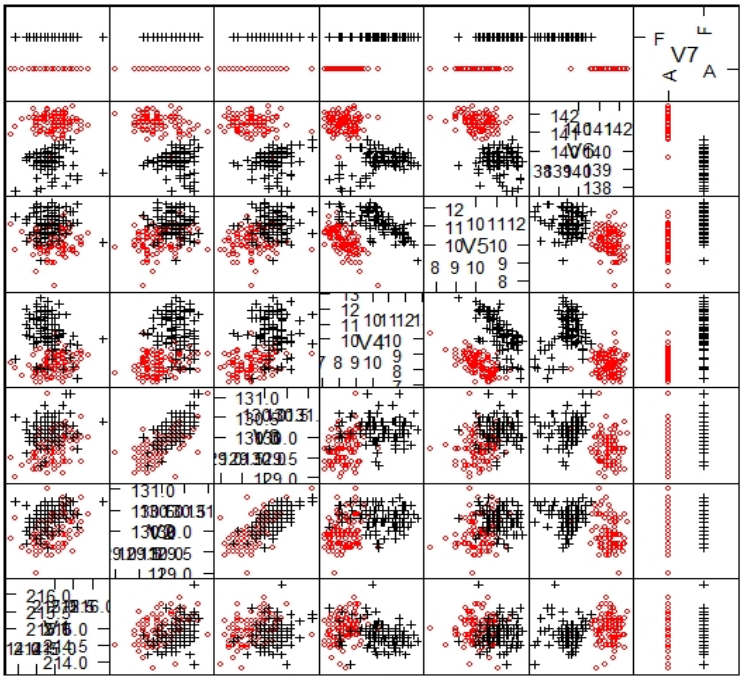

Figura 2.7: Diagrama de dispersión de las variables observadas.

Observemos que en la variable 6, la longitud de la diagonal, se nota más la separación de ambos grupos, es decir, esta variable es la que mejor nos podría indicar cómo discriminar un grupo de otro. Pero hagamos el análisis en **R** para verificar nuestra suposición, pués quiza necesita de alguna otra u otras variables para poder tomar la decisión de pertenencia de grupos.

Entonces tomemos una submuestra de 120 billetes 60 auténticos y 60 billetes falsos, para después clasificar los 80 restantes y verificar si clasifica bien, sin olvidar que la probabilidad de pertenencia a un grupo o al otro es la misma, así mismo tratar de encontrar la función discriminante que mejor separa los grupos.

Consideremos todas las variables para aplicar la función de **R**. Misma que nos da los siguientes resultados:

Función Discriminante:

$$X_7 = -0.005011X_1 - 0.832432X_2 + 0.848993X_3 + 1.117335X_4 + 1.178884X_5 - 1.556520X_6 \tag{2.48}$$

Observemos qué la variable X_1 es la que tiene el coeficiente más pequeño respecto a los demás, debido a esto es la que menos información nos aporta al momento de hacer la clasificación. Pero apesar de esto hace una buena clasificación.

Observemos la clasificación producida por una función de **R** de los 80 billetes restantes.

A A A A A A A A A A A A A A A A A A A A A A A A A A A
A A A A A A A A A A A A A F F F F F F F F F F F F F F F F F F F F F F F
F F F F F F F F F F F F F F F F F

Veamos que clasifico muy bién no presento ningun error, ahora veamos las medias por grupos de las variables utilizadas, en el siguiente cuadro 2.2:

$\bar{X}_{ij}$						
$i \setminus j$	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$
A	214.969	129.943	129.720	8.305	10.168	141.517
F	214.823	130.300	130.193	10.530	11.133	139.450

Cuadro 2.2: Medias por grupo y variables

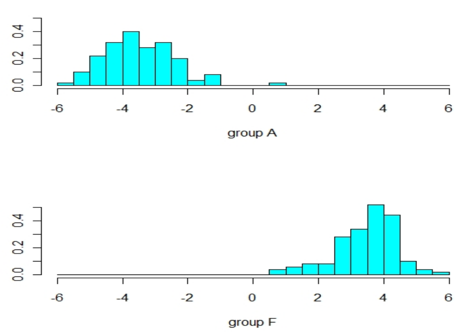

Figura 2.8: Comportamiento de los grupos (Billetes Auténticos y Falsos)

En la Figura 2.8 vemos el comportamiento de ambos grupos. Observemos que se puede ajustar una función Normal, aunque el ajuste es mejor en el grupo de los billetes Auténticos que en el grupo de los billetes Falsos, pués la distribución de los billetes Falsos tiene colas muy pesadas y un sesgo hacia la izquierda, concentrándose más alrededor de la media.

En el cuadro 2.3 presentamos los errores que se obtuvieron al correr la función de **R** y los coeficientes de las funciones quitando de una en una las variables que nos proporcionan menos información.

Variables Utilizadas	**Coeficientes**	**Errores "A"**	**Errores "F"**
X_2	-0.8343963		
X_3	0.8482564		
X_4	1.1176464	0	0
X_5	1.1790249		
X_6	-1.5569778		
X_3	0.3479619		
X_4	1.1030396		
X_5	1.1615401	0	0
X_6	-1.4873244		
X_4	1.133220		
X_5	1.203479	0	0
X_6	-1.493915		
X_5	0.3380229		
X_6	-1.8677551	0	0
X_6	1.978315	0	0

Cuadro 2.3: Errores de clasificación en distintos modelos

De la función original se fueron excluyendo las variables que menos información aportaba al problema de clasificación, hasta llegar a observar que la variable de la longitud de la diagonal es la que clasifica mejor, como lo habíamos observado al principio.

Mostremos como se ve el gráfico del comportamiento de los grupos bajo la variable longitud de la diagonal (X_6), en la Figura 2.9.

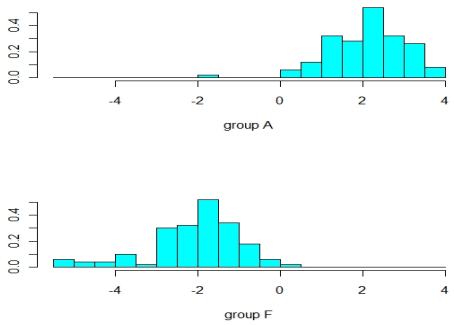

Figura 2.9: Solo la variable: Longitud de la Diagonal

Veamos que la separación es buena ya que clasificó bien a los 80 billetes de la muestra de prueba, entonces es la variable que más información aporta para hacer la discriminación de ambos grupos.

En conclusión podemos decir que la clasificación se puede hacer sólo contemplando la longitud de la diagonal de los billetes Auténticos y Falsos, y mediante ésta decidir si es un billete auténtico o falso. Por lo anterior, utilizando la menor cantidad de variables, la función discriminante es la siguiente:

$$X_7 = 1.978315X_6 \tag{2.49}$$

Como observamos en los ejemplos anteriores, los resultados dependerán de la información que nos aporten las variables observadas para poder realizar la discriminación de los grupos. En el ejemplo 1 fue necesario útilizar mínimo 3 variables mientras que en el ejemplo 2 basto con solo una variable y esto debido a la información contenida en cada variable.

2.2. Regresión Logística

Cuando no se verifican las condiciones de aplicación del análisis discriminante puede utilizarse el discriminante logístico basado en la regresión logística. También se puede utilizar esta técnica si se desea incluir variables discretas.

Los modelos de regresión logística son modelos estadísticos multivariantes tanto explicativos como predictivos en los que se desea conocer la relación entre:

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).
- Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas.

La variable dependiente o variable respuesta no es continua sino discreta generalmente toma valores 0 y 1. La ecuación del modelo no es una función lineal de principio, sino exponencial, que mediante una transformación logarítmica la podemos ver como una función lineal. El modelo puede asemejarse más a la realidad ya que muchos fenómenos se comportan más como una curva que como una recta.

El objetivo primordial de esta técnica es el de modelar cómo influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular, además de predecir la probabilidad de un evento de intéres, como la probabilidad de pertenecer a un grupo u otro, que justo es nuestro problema de estudio (*Credit Scoring*); así como identificar las variables predictoras útiles para tal predicción. El modelo será de utilidad puesto que, muchas veces el perfil de variables puede estar formado por caracteres cuantitativos y cualitativos; y se pretende hacer participar a todos ellos en una sola ecuación conjunta.

Se estima la probabilidad de que un evento ocurra; es decir, el valor esperado de y dado las variables regresoras debe tomar valores entre 0 y 1. Las estimaciones de probabilidad estaran siempre entre 0 y 1, así, el valor de la variable se puede definir como una probabilidad de que ocurra o no un evento sujeto a control.

Es decir

$$E(y|x_i) = 1 * P(y = 1|x_i) + 0 * P(y = 0|x_i) = P(y = 1|x_i)$$

Sea $P(y = 1|x_i) = \pi(x)$ y

además

$$E(y^2|x_i) = 1^2 * \pi(x) + 0^2 * (1 - \pi(x)) = \pi(x)$$

Por otro lado la varianza es:

$$Var(y|x_i) = E(y^2|x_i) - E^2(y|x_i) = \pi(x) - (\pi(x))^2 = \pi(x)(1 - \pi(x))$$

Entonces para una respuesta binaria el modelo de regresión sería:

$$E(y) = \pi(x) = \beta^T X$$

Sin embargo, el incoveniente principal es que $\pi(x)$ debe estar entre 0 y 1 y no hay ninguna garantía de que la predicción $\beta^T X$ verifique dicha restricción ya que el modelo puede prever probabilidades mayores a la unidad.

Debido a lo anterior, como deseamos que el modelo nos proporcione la probabilidad de pertenecer a alguno de los grupos, debemos transformar la variable respuesta de tal manera que esté entre cero y uno para que se pueda tomar una decisión.

2.2.1. Modelo de Regresión Logística

Si deseamos que el modelo nos proporcione directamente la probabilidad de pertenecer a cada uno de los grupos, hay que hacer lo siguiente.

Sean:

$$\beta^T = (\beta_0, \beta_1, \ldots, \beta_p) \qquad \text{y} \qquad X^T = (1, x_1, x_2, \ldots, x_p)$$

Si consideramos que:

$$\pi(x) = F(\beta^T X) \tag{2.50}$$

Entonces ahora buscamos una F que tenga esta propiedad. Se toma F donde dicha función cumpla con ser acotada entre cero y uno. Entonces $\pi(x)$ nos queda de la siguiente manera:

$$\pi(x) = \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}} \tag{2.51}$$

por otro lado

$$1 - \pi(x) = 1 - \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}} = \frac{1 + e^{(\beta^T X)} - e^{(\beta^T X)}}{1 + e^{(\beta^T X)}} = \frac{1}{1 + e^{(\beta^T X)}} \tag{2.52}$$

Considerando el logaritmo del cociente de las probabilidades anteriores:

$$\begin{aligned} log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) &= log\left(\frac{\frac{e^{\beta^T X}}{1+e^{\beta^T X}}}{\frac{1}{1+e^{\beta^T X}}}\right) = log\left(\frac{(1+e^{\beta^T X})(e^{\beta^T X})}{1+e^{(\beta^T X)}}\right) \\ &= log\big((1+e^{\beta^T X})(e^{\beta^T X})\big) - log\big(1+e^{\beta^T X}\big) \\ &= log\big(e^{\beta^T X}\big) + log\big(1+e^{\beta^T X}\big) - log\big(1+e^{\beta^T X}\big) \\ &= log\big(e^{\beta^T X}\big) = \boldsymbol{\beta^T X} \end{aligned} \tag{2.53}$$

$$\therefore \qquad log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2.54}$$

de tal manera que al hacer la transformación se tiene un modelo lineal que se denomina *logit*.

El resultado (2.54) representa en una escala logarítmica la diferencia entre las probabilidades de pertenecer a ambas poblaciones, por lo mismo, al ser una función lineal facilita la estimación y la interpretación del modelo. [Agresti, 2007]

Regla de Clasificación

En el problema para el cual se esta trabajando (*Credit Scoring*) es necesario tomar una decisión, la cual se refiere a la manera de clasificar a un nuevo solicitante de crédito. Es decir, se tienen dos grupos, uno de ellos formado por los individuos que se les concedió un crédito y el otro por los individuos que se les nego el crédito. Entonces si tenemos un nuevo individuo que solicita un crédito ¿cómo tomamos la decisión de otorgarle o de negarle un crédito?

Una vez que se han calculado las probabilidades de pertenencia a cada una de las poblaciones mediante el modelo de regresión logística, el individuo será asignado a aquella población para la que la probabilidad sea mayor, o sea,

- Asignaremos a la población 2 si

$$\pi(x) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}} \geq 0.5$$

- En caso contrario asignar a la población 1

$$\pi(x) < 0.5$$

El punto de corte es arbitrario sin embargo, en la práctica es más usual utilizar 0.5 como punto de corte, es decir, si el valor previsto está más próximo a 0 que a 1, clasificaremos al elemento en la primera población. En otro caso, lo haremos en la segunda.

Por ejemplo: supongamos que la población 1 esta formada por los individuos que no se les concedió el crédito y la población 2 por los individuos que si se les concedió, y tenemos un nuevo solicitante al cual el modelo de regresión logística le calculó una probabilidad de pertenencia de 0.3721, entonces este individuo lo asignaremos a la población 1, es decir, a este solicitante no le otorgaremos el crédito. Pero hay que tener en cuenta que estos métodos sólo son una medida y no hay que tomar la decisión de manera autoritaria. Uno podría medir ó cuantificar el riesgo de incumplimiento o bien dependerá de la entidad financiera.

2.2.2. Interpretación de los Coeficientes

Los coeficientes en una regresión logística son β^T, pero lo que realmente se interpreta son sus exponenciales, es decir, $exp(\beta^T)$, conocidos como "*momios*" y definidos como la probabilidad de que ocurra el evento.

La probabilidad de que ocurra (*momios*) se define como el cociente entre las probabilidades de que el evento ocurra $P(y = 1/x) = p(x)$ y la probabilidad de que no ocurra

$P(y=0/x)$

$$\begin{aligned} momios = \frac{\pi(x)}{1-\pi(x)} &= exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \\ &= e^{\beta_0} exp(\underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}_{\beta^{*T} X^*}) \\ &= e^{\beta_0} e^{\beta^{*T} X^*} \\ &= e^{\beta_0} \prod_{j=1}^{k} (e^{\beta_j})^{x_j} \\ \Rightarrow \qquad log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \end{aligned} \tag{2.55}$$

Esta relación exponencial proporciona una interpretación de β, entonces, los coeficientes estimados se interpretan como el cambio en el logaritmo del *momios* por un cambio unitario en la variable independiente asociada. Es decir, los *momios* a nivel $x+1$ es igual a la probabilidad en x multiplicado por $e^{\beta^{*T}}$, permaneciendo las demás variables constantes. Cuando $\beta^{*T} = 0 \Rightarrow e^{\beta^{*T}} = 1$ y el *momio* no cambia, como cambia x

Supongamos que consideramos dos elementos que tienen valores iguales en todas las variables menos en una. Sean $(x_{i1}, \ldots, x_{ih}, \ldots, x_{ip})$ los valores de las variables para el primer elemento y $(x_{j1}, \cdots, x_{jh}, \cdots, x_{jp})$ para el segundo, y todas las variables son las mismas en ambos elementos menos en la variable h donde $x_{ih} = x_{jh} + 1$, entonces:

El cociente de *momios* para la variable independiente x_{ih} dicotómica es igual a:

$$\begin{aligned} cociente\ de\ momios &= \Psi = \frac{\frac{\pi(x_{jh}=1)}{1-\pi(x_{jh}=1)}}{\frac{\pi(x_{ih}=0)}{1-\pi(x_{ih}=0)}} \\ &= \frac{e^{\beta_0} \prod_{j=1}^{p} (e^{\beta_j})^{x_j}}{e^{\beta_0} \prod_{i=1}^{k} (e^{\beta_i})^{x_i}} \end{aligned} \tag{2.56}$$

Como sabemos que son iguales todos excepto el término h queda como sigue:

$$= \frac{\prod_{j=1}^{h-1} (e^{\beta_j})^{x_j} \quad e^{\beta_h x_{jh}} \quad \prod_{j=h+1}^{p} (e^{\beta_j})^{x_j}}{\prod_{j=1}^{h-1} (e^{\beta_j})^{x_j} \quad e^{\beta_h x_{ih}} \quad \prod_{j=h+1}^{p} (e^{\beta_j})^{x_j}} = \quad e^{\beta_h}$$

El valor de Ψ indica cuánto se modifica, es decir, aumenta (o disminuye) el cociente de la probabilidad de que se presente el evento entre los individuos que pertenecen a la categoría $x = 1$, en relación con la de los que corresponden a la categoría de referencia ($x = 0$). La interpretación varia según el tipo de variables independientes consideradas.

La fórmula de Regresión Logística indica que los aumentos *logit* para cada β aumentan una unidad de x. La mayoría de nosotros no pensamos naturalmente en un *logit* (logaritmo de momios) de escala, por lo que tenemos que tener en cuenta las interpretaciones alternativas.

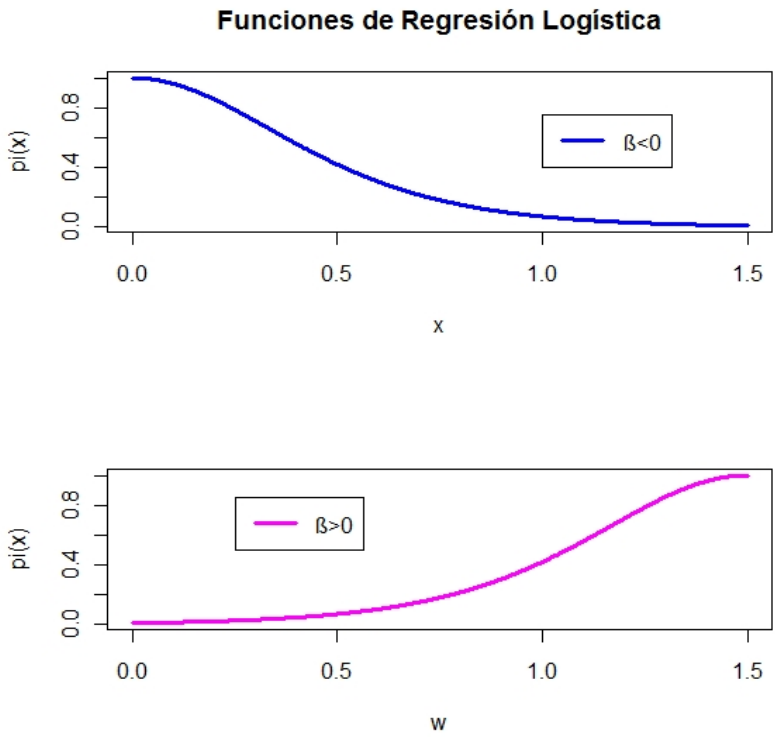

Figura 2.10: Funciones de Regresión Logística

Como se puede ver en la figura 2.10 el parámetro β de la ecuación determina la tasa de incremento o decremento de la curva para $\pi(x)$. El signo de β indica si la curva asciende ($\beta > 0$) ó desciende ($\beta < 0$) y la tasa de cambio aumenta como $|\beta|$ aumente. Cuando $\beta = 0$ la ecuación se reduce a una constante, entonces la $\pi(x)$ es idéntica en todas las x lo que hace que la curva se convierta en una línea horizontal. Por lo que la variable respuesta binaria Y es independiente de X.

2.2.3. Estimación de los Parámetros

Como el modelo no es lineal se requiere de un algoritmo iterativo para estimar los parámetros.

Para la estimación de los coeficientes del modelo y de su error estándar se recurre al cálculo de estimaciones de máxima verosimilitud, es decir, estimaciones que maximicen la probabilidad de obtener los valores de la variable dependiente Y proporcionados por los datos de nuestra muestra. Estas estimaciones no son de cálculo directo, como ocurre en el caso de las estimaciones de los coeficientes de la regresión lineal múltiple por el método de mínimos cuadrados. Para el cálculo de estimaciones máximo-verosímiles se recurre a métodos iterativos, como el método de Newton-Raphson. Dado que el cálculo es complejo, normalmente hay que recurrir al uso de rutinas de programación o a paquetes estadísticos. De estos métodos surgen no sólo las estimaciones de los coeficientes de regresión, sino también de su error estándar y de las covarianzas entre las covariables del modelo.

Metodo de Newton-Rapson

Se trata de un método iterativo, empleado en diversos problemas matemáticos, como en la determinación de raíces de ecuaciones, y en nuestro caso, en las estimación de los coeficientes de la regresión logística β mediante máxima verosimilitud.

Sea

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} & \dots & x_{1,p} \\ 1 & x_{2,2} & x_{2,3} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n-1,2} & x_{n-1,3} & \dots & x_{n-1,p} \\ 1 & x_{n,2} & x_{n,3} & \dots & x_{n,p} \end{pmatrix}$$

Con un vector de coeficientes

$$\beta = (\beta_0, \beta_1, \dots, \beta_p)$$

el proceso inicia construyendo la función de verosimilitud de la ecuación de regresión logística. Consideremos la verosimilitud:

$$L = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{n_i - y_i}$$

$$
\begin{aligned}
&= \Big\{\prod_{i=1}^{n}[1-\pi(x_i)]^{n_i}\Big\}\Big\{\prod_{i=1}^{n}\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big)^{y_i}\Big\} \\
&= \Big\{\prod_{i=1}^{n}[1-\pi(x_i)]^{n_i}\Big\}\Big\{\prod_{i=1}^{n}exp\Big[log\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big)^{y_i}\Big]\Big\} \\
&= \Big\{\prod_{i=1}^{n}[1-\pi(x_i)]^{n_i}\Big\}exp\Big[\sum_{i=1}^{n}y_i log\Big(\frac{\pi(x_i)}{1-\pi(x_i)}\Big)\Big] \qquad (2.57)
\end{aligned}
$$

En el modelo de (2.54) el $i-$*ésimo* término dado por $\beta^T X = \sum_j \beta_j x_{ij}$, por lo que el término exponencial en la última expresión es igual a: $exp[\sum_i y_i(\sum_j \beta_j x_{ij})] = exp[\sum_j(\sum_i y_i x_{ij})\beta_j]$. También $[1-\pi(x_i)] = \big(1+exp(\sum_j \beta_j x_{ij})\big)^{-1}$, $\Rightarrow$ el logaritmo de máxima verosimilitud es igual a:

$$
\begin{aligned}
\mathbf{log(L)} &= log\Bigg(\prod_{i=1}^{n}\Big[1+exp\Big(\sum_j \beta_j x_{ij}\Big)\Big]^{-n_i}exp\Big[\sum_j\Big(\sum_i y_i x_{ij}\Big)\beta_j\Big]\Bigg) \\
&= log\Bigg(\prod_{i=1}^{n}\Big[1+exp\Big(\sum_j \beta_j x_{ij}\Big)\Big)\Bigg)^{-n_i} + log\Bigg(exp\Big(\sum_j\Big(\sum_i y_i x_{ij}\Big)\beta_j\Big)\Bigg) \\
&= \sum_j\Big(\sum_i y_i x_{ij}\Big)\beta_j - \sum_i n_i log\Big[1+exp\Big(\sum_j \beta_j x_{ij}\Big)\Big] \qquad (2.58)
\end{aligned}
$$

Para poder obtener el estimador de β que es justo lo que buscamos necesitamos derivar con respecto de β y después igualar a cero para obtener dicho estimador. Entonces lo que resulta es:

$$
\begin{aligned}
\frac{\partial L}{\partial \beta} &= \sum_i y_i x_{ij} - \sum_i n_i x_{ij}\left[\frac{exp(\beta^T X)}{1+exp(\beta^T X)}\right] \\
&= \sum_i y_i x_{ij} - \sum_i n_i x_{ij}\hat{\pi}(x_i) = 0 \qquad (2.59)
\end{aligned}
$$

Sea $\hat{m}_i = n_i\hat{\pi}(x_i)$ entonces la última expresión la podemos escribir:

$$
X^T Y = X^T \hat{m} \quad \Rightarrow \quad \frac{\partial L}{\partial \beta} = X^T(Y-\hat{m}) \qquad (2.60)
$$

Obtengamos la segunda derivada, llamada matriz Hessiana o matriz de información la cual nos servira para garantizar que efectivamente sea un máximo.

$$
\frac{\partial^2 L}{\partial \beta^2} = \sum_i \frac{x_{ij}n_i x_{ij} exp\Big(\sum_j \beta_j x_{ij}\Big)}{\Big[1+exp\Big(\sum_j \beta_j x_{ij}\Big)\Big]^2} = -\sum_i x_{ij}n_i x_{ij}\pi_i(1-\pi_i) \qquad (2.61)
$$

Donde $\Lambda = n_i\pi_i(1-\pi_i)$ denota una matriz diagonal, tal que los elementos de la diagonal están dados por los productos de $\pi_i(1-\pi_i)$ lo cual nos lleva a ver que:

$$\frac{\partial^2 L}{\partial \beta^2} = -X^T \Lambda X \tag{2.62}$$

Entonce una vez teniendo todo lo anterior procederemos a describir el método requerido:

1. Se le asigna un valor inicial empírico a los coeficientes de regresión, en general 0 a todos ellos.

2. En cada iteración t la matriz de nuevos coeficientes de regresión resulta de sumar matricialmente un gradiente a la matriz de coeficientes propuestos del paso anterior. Este gradiente es el resultado del cociente entre la primera derivada y la segunda derivada resultantes de la función de verosimilitud de la ecuación de regresión.
$$\beta^{(t)} = \beta^{(t-1)} + (X^T \Lambda^{(t)} X)^{-1} X^T (Y - m^{(t)}) \tag{2.63}$$
Donde: $m^{(t)} = n_i \pi_i^{(t)}$ y $\Lambda = n_i \pi_i^{(t)}(1 - \pi_i^{(t)})$

3. El segundo paso se repite tantas veces como sea necesario hasta que la diferencia entre la matriz de coeficientes de regresión en dicha iteración y la matriz de la iteración previa, sea 0 ó practicamente 0 (por ejemplo 10^{-6}).

4. Una vez finalizadas las iteraciones, la inversa de la matriz informativa de la última iteración nos da los valores de varianzas y covarianzas de las estimaciones de los coeficientes de regresión.

   Es decir el error cuadrático estándar de cada coeficiente de regresión coincide con la raíz cuadrada del elemento respectivo de la diagonal principal, y por debajo de esta diagonal quedan las covarianzas de cada pareja de variables.

En la figura 2.11 mostramos un ejemplo para poder encontrar en esté caso el máximo utilizando el metodo de Newton-Rapson, como se puede ver iniciamos $x = 0$ con la finalidad que llegar al punto máximo.

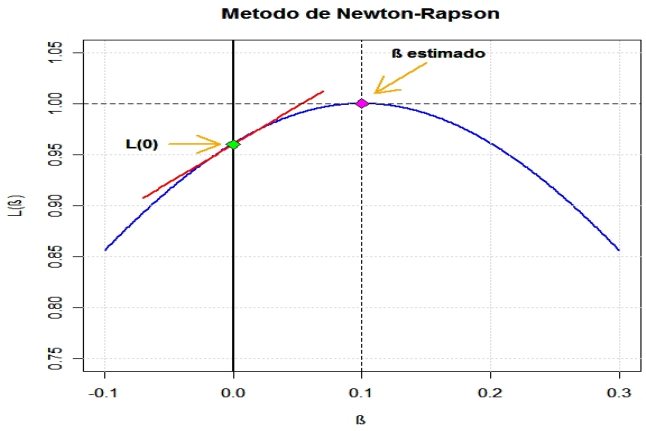

Figura 2.11: Metodo de Newton-Rapson

2.2.4. Regresión Logística con variables Cualitativas

Como la metodología empleada para la estimación del modelo se baso en la utilización de variables cuantitativas, al igual que en cualquier otro procedimiento de regresión, sin embargo es incorrecto pensar que en el no puedan intervenir variables cualitativas, ya sean nominales u ordinales.
Si la variable cualitativa tuviera más de dos categorías, para su inclusión en el modelo deberíamos realizar una transformación de la misma en varias variables cualitativas. Ya que la asignación de un número a cada categoría no resuelve el problema.

Las variables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia. La solución a este problema es crear tantas variables dicotómicas como número de respuestas menos una, estas nuevas variables, artificialmente creadas, reciben el nombre de "***dummy***" denominadas también como **variables internas, indicadoras,** ó **variables diseño**, de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la variable cualitativa posee n categorías, habrá que considerar $n-1$ variables auxiliares.

Supongamos una respuesta binaria Y tiene dos predictores independientes hallados, X y Z. Sea x y z, cada una toma valores de 0 y 1 para representar las dos categorías de cada variable explicativa. El modelo para $P(Y=1)$ resulta ser el siguiente:

$$logit[P(Y=1)] = \alpha + \beta_1 x + \beta_2 z$$

Por lo anterior x y z son las llamadas variables "***dummy***" ellas indican las categorías de los predictores. En el cuadro 2.4 se muestran los valores logit en las cuatro combinaciones de valores de las dos variables predictoras.

x	z	Logit
0	0	α
1	0	$\alpha + \beta_1$
0	1	$\alpha + \beta_2$
1	1	$\alpha + \beta_1 + \beta_2$

Cuadro 2.4: Variables predictoras "dummy"

Este modelo supone una ausencia de interacción, es decir, no existe relación entre el factor de estudio y la variable dependiente por lo que no se modifica según el valor de una tercer variable. El efecto de un factor es la misma en cada categoría del otro factor. En una categoría fija z de Z, el efecto sobre el logit de pasar de $x = 0$ y $x = 1$ es

$$= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1$$

La diferencia entre dos *logits* es igual a la diferencia de *log* de *momios*.

2.2.5. Ejemplo

Utilicemos la base de los "crabs" considerada en el Ejemplo 1 de análisis discriminante.

Las variables a considerar son las siguientes:

sp = Especies ("B" ó "O"), es decir, Blue=Azul ó Orange=Naranja.

FL = Tamaño del lóbulo Frontal.

RW = Anchura Trasera.

CL = Longitud del Caparazón

CW = Ancho del Caparazón

BD = Profundidad del Cuerpo.

sex = Sexo.

index = índice dentro de cada uno de los 4 grupos $(1 : 50)$.

A diferencia del ejemplo de análisis discriminante, consideraremos la variable *sex* ya que esta es una variable cualitativa con dos categorías de cangrejos; hembras y los machos. Comencemos observando cómo se ven los datos en el gráfico de la figura 2.12:

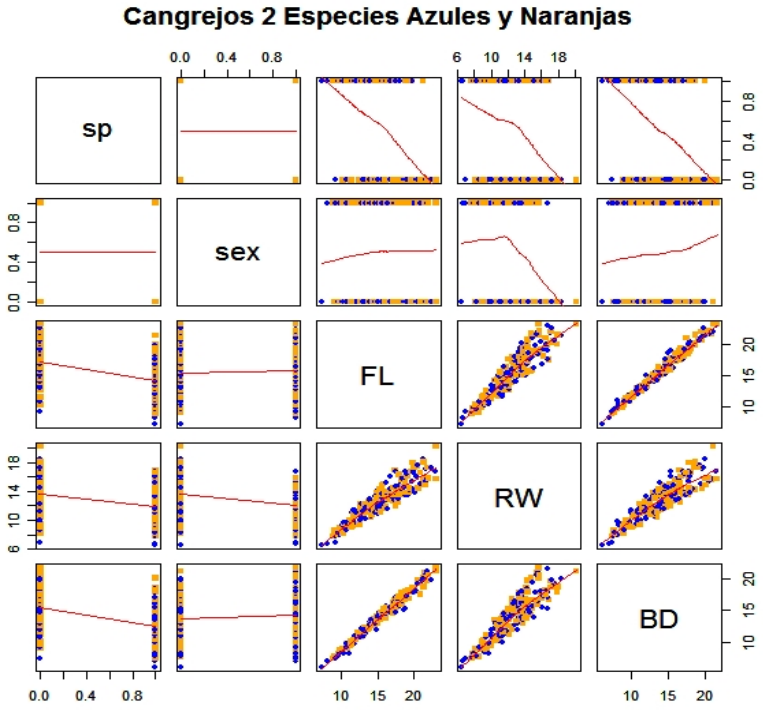

Figura 2.12: Gráfica de matriz de variables

De nuestra muestra de 200 cangrejos tomemos una submuestra de 120, para después clasificar los 80 cangrejos restantes y poder evaluar los errores de clasificación. Lo que buscamos es saber si mediante, las variables sex, FL, RW, BD, podemos decidir a que especie pertenece cada uno de los 80 cangrejitos restantes.

Entonces mediante la función de **R** que nos da una regresión para modelos generalizados, obtenemos la siguiente función de regresión logística. Considerando que los coeficientes dados son los llamados logit, la función resultante es:

$$log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = 1.92 + 2.14sex - 0.97FL + 1.34RW - 0.36BD$$

dado esto como lo que deseamos es obtener la función de probabilidad para saber como clasificar a nuestros cangrejitos y saber a que especie pertenece, entonces la $\pi(x)$ resulta ser:

$$\begin{aligned} \pi &= \frac{exp(1.92 + 2.14sex - 0.97FL + 1.34RW - 0.36BD)}{1 + exp(1.92 + 2.14sex - 0.97FL + 1.34RW - 0.36BD)} \\ &= \frac{6.80 * 8.48^{sex} * 0.38^{FL} * 3.81^{RW} * 0.70^{BD}}{1 + (6.80 * 8.48^{sex} * 0.38^{FL} * 3.81^{RW} * 0.70^{BD})} \end{aligned} \tag{2.64}$$

R nos arroja pruebas de significancia sobre los coeficientes que nos ayudan a verificar qué tan adecuado es el modelo de regresión logística, observemos qué nos dicen estas pruebas.

	Coeficiente	**Error Estandar**	**Prueba** z	$\mathbf{P}(>\|z\|)$
$\beta_\mathbf{0}$	1.92	1.6081	1.192	0.233367
sex	2.14	0.8473	2.522	0.011673
FL	-0.97	0.4811	-2.015	0.043889
RW	1.34	0.4059	3.293	0.000991
BD	-0.36	0.4518	-0.799	0.424016

Cuadro 2.5: Resumen del modelo de regresión logística

El cuadro 2.5 nos permite revisar cada uno de nuestros coeficientes por separado y probar $H_0 : \beta_i = 0$ para cada i. La última columna nos proporciona el p-valor de la prueba de significancia basado en la distribución $t - Student$. Si comparamos con un nivel de significancia $\alpha = 5\,\%$ entonces resulta qué para sex, FL, y RW se rechaza la hipótesis mientras que para BD y β_0 (intercepto) no se rechaza la hipótesis, debido a esto las podriamos excluir de nuestro modelo. El programa también nos proporciona el *índice de Akaike* AIC=127.17 este índice nos indica qué tan adecuado es nuestro modelo, buscamos un índice AIC chico a comparación que el de otro modelo. Lo que nos conduce a considerar un nuevo modelo.

Las predicciones para este modelo quedan como sigue:

B B B B B O B O O B O O O O O O O O B B B B B B B B B B B B B B B B B B O B B B B O O O O B O O O O O O O O O O O B O O B B B O B O B O O O O B B B B B B B B

Veamos que la clasificación no fue tan buena ya que tuvimos 29 errores de clasificación, pués de los 80 cangrejos que teniamos 40 de ellos tenian que ser azules y 40 naranjas, sin embrago, nos dice que de los 40 azules solo 28 perteencen a esta especie y los 12 restantes a la especie de los naranjas mientras que de los 40 naranjas 23 clasifica bien en esta clase y los restantes 17 los clasifica en la especie de azules, lo que nos lleva a concluir que la función de regresión no es la adecuada para poder hacer la clasificación. Si recordamos que las pruebas anteriores nos decian que BD y β_0 (intercepto) no rechazan la hipótesis, es decir, no son significativas. Entonces lo que nos resta hacer es verificar que ocurre si quitamos esta variable y el intercepto del modelo y observar como realiza la clasificación esa nueva función.

Observemos como nos queda la función de regresión logística si ahora no consideramos

la variable BD.

$$log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = 2.76sex - 1.49FL + 1.70RW$$

La probabilidad deseada es:

$$\begin{aligned} \pi &= \frac{exp(2.76sex - 1.49FL + 1.70RW)}{1 + exp(2.76sex - 1.49FL + 1.70RW)} \\ &= \frac{15.83^{sex} * 0.23^{FL} * 5.45^{RW}}{1 + (15.83^{sex} * 0.23^{FL} * 5.45^{RW})} \end{aligned} \tag{2.65}$$

Y el resumen de las pruebas nos queda como sigue:

	Coeficiente	**Error Estandar**	**Prueba** z	**P**$(>\|z\|)$
sex	2.76	0.6297	4.387	$1.15\text{x}10^{-5}$
FL	-1.49	0.2898	-5.138	$2.77\text{x}10^{-7}$
RW	1.70	0.3338	5.081	$3.76\text{x}10^{-7}$

Cuadro 2.6: Resumen del 2 ° modelo de regresión logística

Observemos en el cuadro 2.6 que efectivamente las pruebas se rechazan para las variables *sex*, FL y RW. El AIC= 126.97, aunque es poca la diferencia con respecto al modelo anterior, el AIC es menor por lo tanto esperaríamos que la clasificación sea mejor que en el modelo previo, veamos como nos queda la clasificación, con esta función de regresión logística.

B B B B B O B B B B O O O B B O O O B B B B B B B B O O B B B B B B B B B O B B B B O O O O B O O O O O O O O O O O B O O B O B O B O B O O O O O O O O B O O O

Con esta función clasifica un poco mejor ya que ahora solo tiene 19 errores en total, de los 80 cangrejos, 40 eran azules y 40 naranjas. Sin embargo de los 40 azules solo clasifica bién 30 y los restantes los clasifica como naranjas mientras que de los 40 naranjas clasifica bién 31 y los 9 restantes los clasifica como azules. Dado esto podemos decir que esta función de regresión logística es mejor que la anterior. Sin embargo, estas variables no nos proporcionan la información suficiente para hacer la separación de las especies.

Capítulo 3

Aplicación en "credit scoring"

En este capítulo haremos una aplicación de un problema al que las entidades de crédito se enfrentan día a día, como es el *"Credit Scoring"* (o traducido al español como "puntaje de crédito") usando la teoría tratada en el capítulo 2. Básicamente consiste en determinar si se concede ó se rechaza un crédito a un nuevo solicitante.

Utilizaremos la base Aprobación de Crédito de Australia, como su nombre lo dice es una base Australiana que consta de 690 datos, su fuente es confidencial, cuenta con 14 variables de las cuales 6 de ellas son de tipo numérico y 8 de tipo categórico, que miden si se concede o no una tarjeta de crédito.

De los 690 datos 307 pertenecen a la clase "1", es decir, son clientes a los cuales se le otorgó la tarjeta de crédito mientras que los 383 restantes (clase "0") son los clientes que se les negó la tarjeta de crédito.

Esta base se refiere a las solicitudes de tarjetas de crédito. Todos los nombres de sus atributos han sido cambiados para proteger los símbolos y la confidencialidad de los datos. Debido a esto sabremos que las variables aportan mucha información o que no aportan nada, pero no será posible decir exactamente a que se referia dicha variable, sin embargo, podremos hacer referencia a cada una de ellas como se denotan en la lista.

Las variables a analizar son las siguientes:

A_1 categórica	A_8 categórica	A_{15} clase (+ ó −)
A_2 continua	A_9 categórica	
A_3 continua	A_{10} continua	
A_4 categórica	A_{11} categórica	
A_5 categórica	A_{12} categórica	
A_6 categórica	A_{13} continua	
A_7 continua	A_{14} continua	

Comenzaremos conociendo la base mediante un análisis exploratorio de cada variable para ir observando el comportamiento de nuestra base de datos, para después continuar con la aplicación de la teoría analizada en el capítulo 2, es decir, intentar buscar un buen modelo para este problema.

Se utilizará Análisis Discriminante para las variables continuas y Regresión Logística para todas las variables categóricas o cualitativas y las variables continuas, después haremos la combinación de ambas técnicas y así llegar a una conclusión.

3.1. Análisis Exploratorio

En esta sección analizaremos cada una de las variables así como la base en general. Haremos diagramas de caja de las variables cuantitativas ya que de estos podemos obtener una primera visión de como se comporta cada una de ellas, intentaremos averiguar si estas variables se distribuyen normal. Veremos si los datos para cada variable se dispersan demasiado o no, su media así como cuales son las variables que mejor componen nuestro modelo [1]. Mediante un gráfico más general de estas variables intentaremos observar si es fácil distinguir dos grupos esperando que los grupos sean los de nuestro interes.

3.1.1. Gráficos

Antes de iniciar a analizar el comportamiento de las variables presentamos un diagrama de caja en donde señalamos el nombre de cada una de las partes como le llamaremos en nuestro análisis.

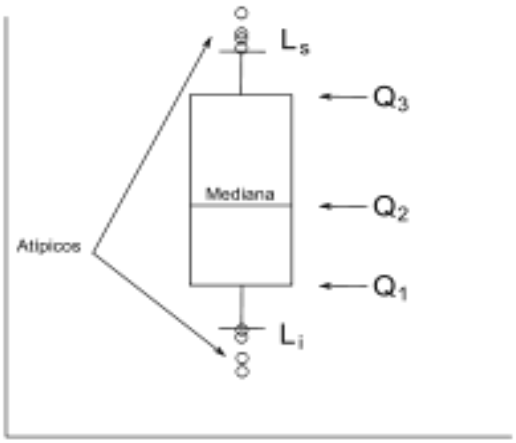

Figura 3.1: Diagrama de Caja

[1] Metodo de componentes principales

Comencemos observando cada uno de los diagramas de las variables cuantitativas.

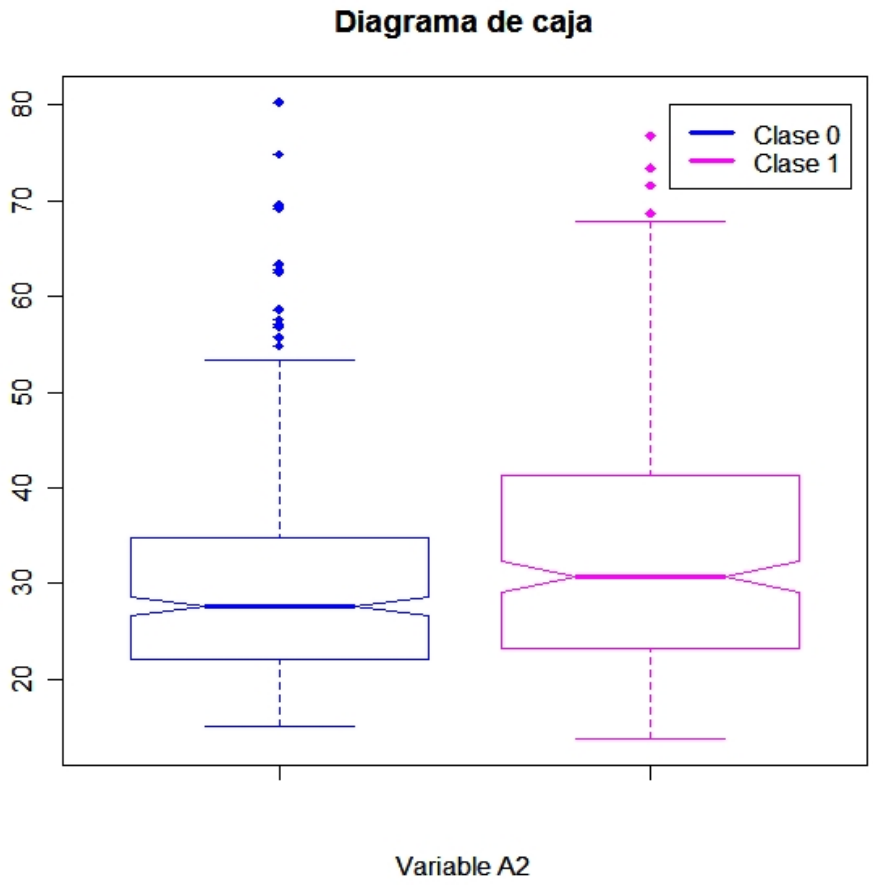

Figura 3.2: Variable A_2

En los diagramas de la figura 3.2 se muestran por separado el comportamiento de la variable A_2 para cada clase. Observemos que estos diagramas de caja son similares, el cuantil (Q_2) qué nos indica la mediana de cada diagrama, se aprecia que estan ubicadas aproximadamente en 29 y 34 para la clase "0" y "1" respectivamente. El diagrama para la clase "0" muestra que los datos estan más dispersos entre el 50 % y el 75 % que entre el 25 % y el 50 %, analogamente para la clase "1", por otro lado el bigote inferior en más corto que el superior lo que nos indica que los datos estan más concentrados entre el 75 % y el L_s de los datos, ambos presentan puntos atípicos.

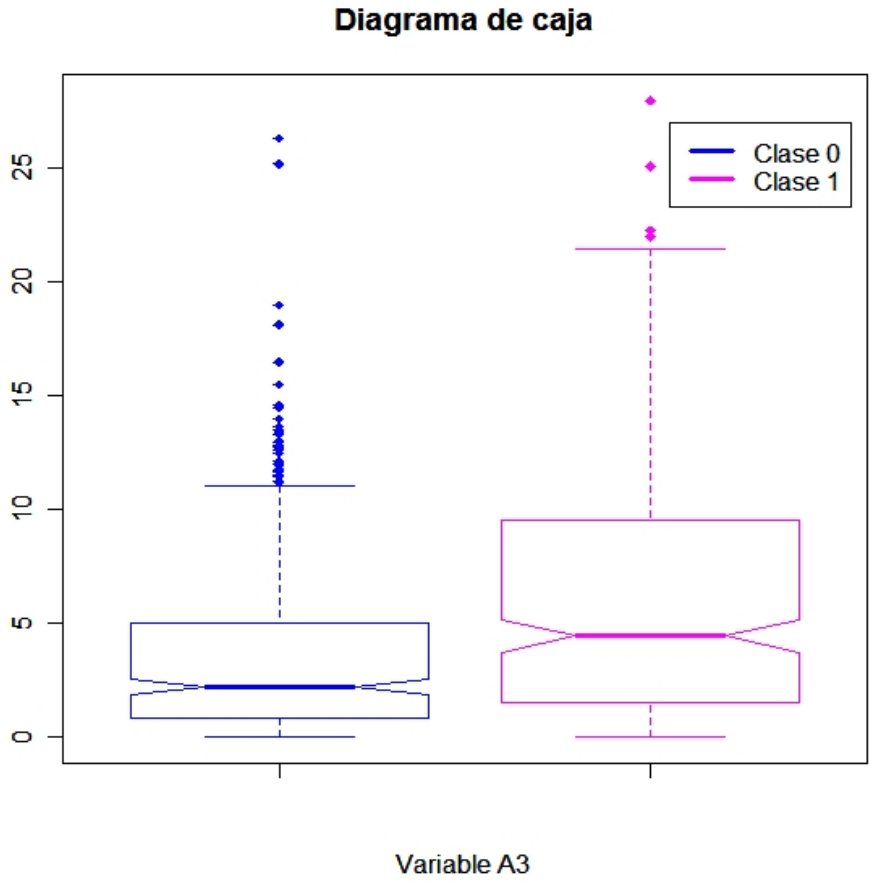

Figura 3.3: Variable A_3

En los diagramas de la figura 3.3 parece notarse una diferencia entre ellos pués el diagrama de caja de la clase "0" presenta mas dispersión en los datos entre el 50 %(Q_2) y el 75 %(Q_3) con una mediana de 4 aproximadamente, por otro lado el bigote superior es más largo que el inferior lo que nos indica que los datos se encuntran más concentrados entre Q_3 y L_s que entre el L_i y Q_1, por otro lado presenta muchos datos atípicos podríamos decir que los datos para esta clase estan dispersos. Mientras que para el diagrama que pertence a la clase "1" presenta una dispersión mayor entre Q_2 y el Q_3 a comparación que entre el Q_1 y el Q_2, como el bigote inferior es más pequeño en el superior esto nos dice que los datos se concentran más entre el Q_3 y el L_s, tambien presenta puntos atípicos pero en menos cantidad que el diagrama de la clase "0" estos diagramas parecen mostrar diferencias significativas entre ambas clases, éxiste la posibilidad que sea una variable candidata para poder hacer la separación.

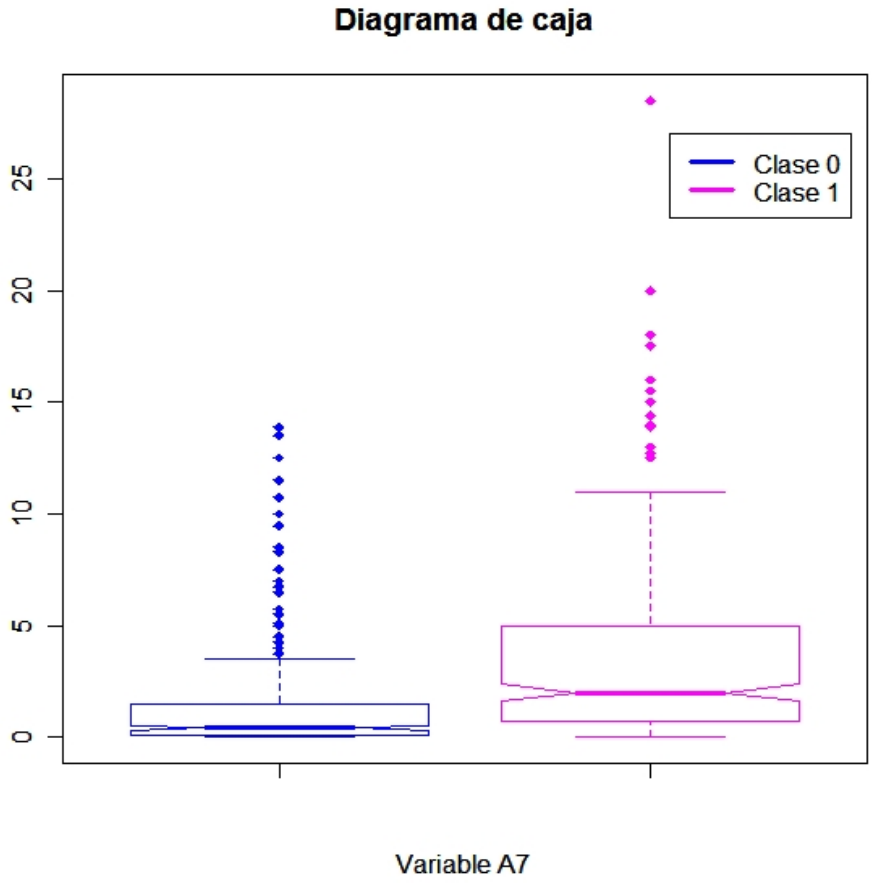

Figura 3.4: Variable A_7

Los diagramas de caja de la figura 3.4 muestran las diferencias que existen entre las clases para la variable 7, ya que en el dagrama de la clase "0" los datos se dispersan más entre el Q_2 y el Q_3 a comparación que entre el Q_1 y el Q_2 con una mediana aproximada de 1, por otro lado, se observa que el bigote superior es más largo que el inferior lo que nos indica que los datos se encuentran más concentrados entre el Q_3 y el L_s, presenta demasiados datos fuera del rango intercuantílico. Mientras que para el diagrama que pertenece a la clase "1" los datos se dispersan más entre el Q_2 y el Q_3, sin embargo se alojan más datos en el bigote superior que en el inferior ya que el bigote superior es más largo que el inferior, se aprecia que tiene una mediana 3.5 aproximadamente. Comparando ambos diagramas podemos decir que los dos presentan diferencias que nos ayudarían a poder realizar la distinción entre nuestro grupos. Sin embargo, es necesario tener cuidado con los datos atípicos que presentan ambas clases ya que podrian causar algun problema al momento de realizar la discriminación de clases.

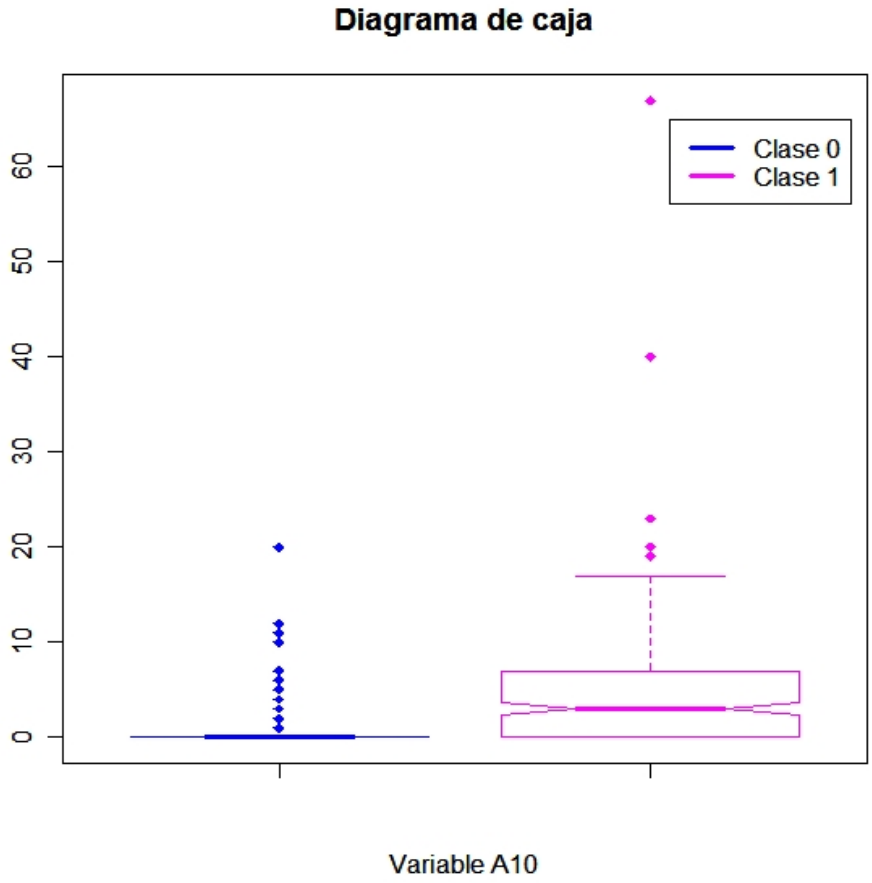

Figura 3.5: Variable A_{10}

En los diagrams de la figura 3.5 se observa que para la clase "0" todos los datos se concentran en el rango intercuantílico, tiene una mediana de se encuantra alrededor del 0, al igual que las variables anteriores presenta datos atípicos. Para el diagrama de la clase "1" los datos parecen acumularse simétricamente alrededor de la mediana, tiene una mediana de 5 aproximadamente, el bigote superior es considerablemente mayor que el bigote inferior lo que nos indica que los datos estan más concentrados entre el Q_3 y el L_s, también presenta datos atípicos muy separados del rango intercuantílico donde se agrupa la mayoría de los datos para esta clase. Las diferencias entre ambas clases son significativas, el cuantil Q_2 es distinto para ambas clases, aparenta que cada clase mide en distintos rangos, podría ser una candidata para realizar nuestra clasificación.

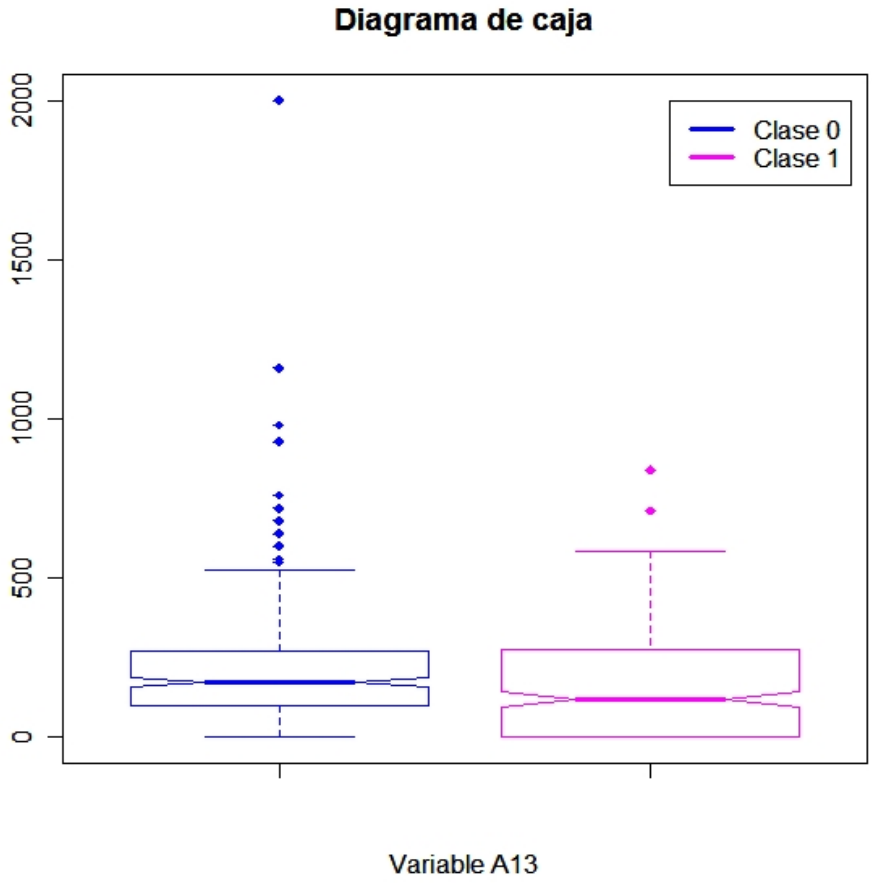

Figura 3.6: Variable A_{13}

Estos diagramas de la figura 3.6 reflejan diferencias entre los grupos, ya que para el diagrama de la clase "0" se observa la caja simétrica con respecto a la medina, tiene una mediana (Q_2) de 200 aproximadamente, por otro lado el bigote superior es mayor con respecto del inferior lo que nos dice que los datos se concenran más entre el Q_3 y el L_s, presenta datos atípicos, datos que se salen del rango intercuantílico. Mientras que para la clase "1" se observa que los datos se dispersan más entre el Q_2 y el Q_3, tiene una mediana de 165 aproximadamente y al igual que la clase "0" el bigote superior es más largo que el inferior indicandonos que los datos se concentrarón más entre el Q_3 y el L_s, tambien presenta datos atípicos. Si consideramos las diferencias pudiera ayudar a discriminar los grupos.

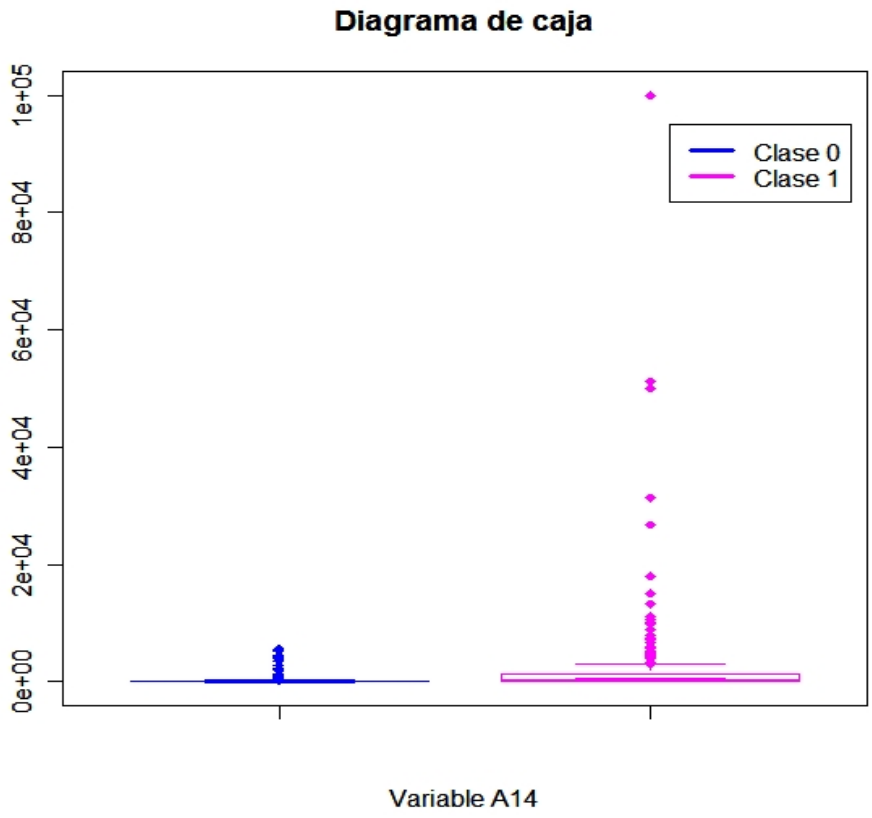

Figura 3.7: Variable A_{14}

Los diagramas de la figura 3.7 muestran el comportamiento de la variable 14, que para ambos diagramas los datos se concentran alrededor de Q_2, sin embargo, la escala en la que estan es muy diferente, indican que su concentración de datos esta en distintos intervalos ya que la clase "1" presenta bastantes datos atípicos, en esta variable parecen dispersarse demasiado los datos, será necesario tener cuidado con ya que puede causarnos algun problema al momento de realizar nuestra clasificación.

Como podemos observar los datos para la clase "0" se encuentran más dispersos entre el Q_2 y el Q_3 análogamente para la clase "1", esta presenta que el bigote superior es más largo lo que nos lleva a pensar que la tarjeta de crédito se le otorga a quienes tienen mayores cantidades en todas las variables y es lógico pensarlo pués se la otorgaríamos a quienes tengan mayores ingresos.

A continuación observemos como se ven las variables graficadas entre si, primero presentaremos el gráfico de todas las variables cuantitativas 3.8. Recodemos que estamos buscando distinguir dos grupos, entonces esperamos que de estos gráficos podamos hacer la distinción de dichos grupos. Ya que quisieramos que un grupo fuera el de aquellos individuos a los que si se les otorgó la tarjeta de crédito y el otro sea el grupo de indi-

viduos a los cuales se le negó dicha tarjeta de crédito.

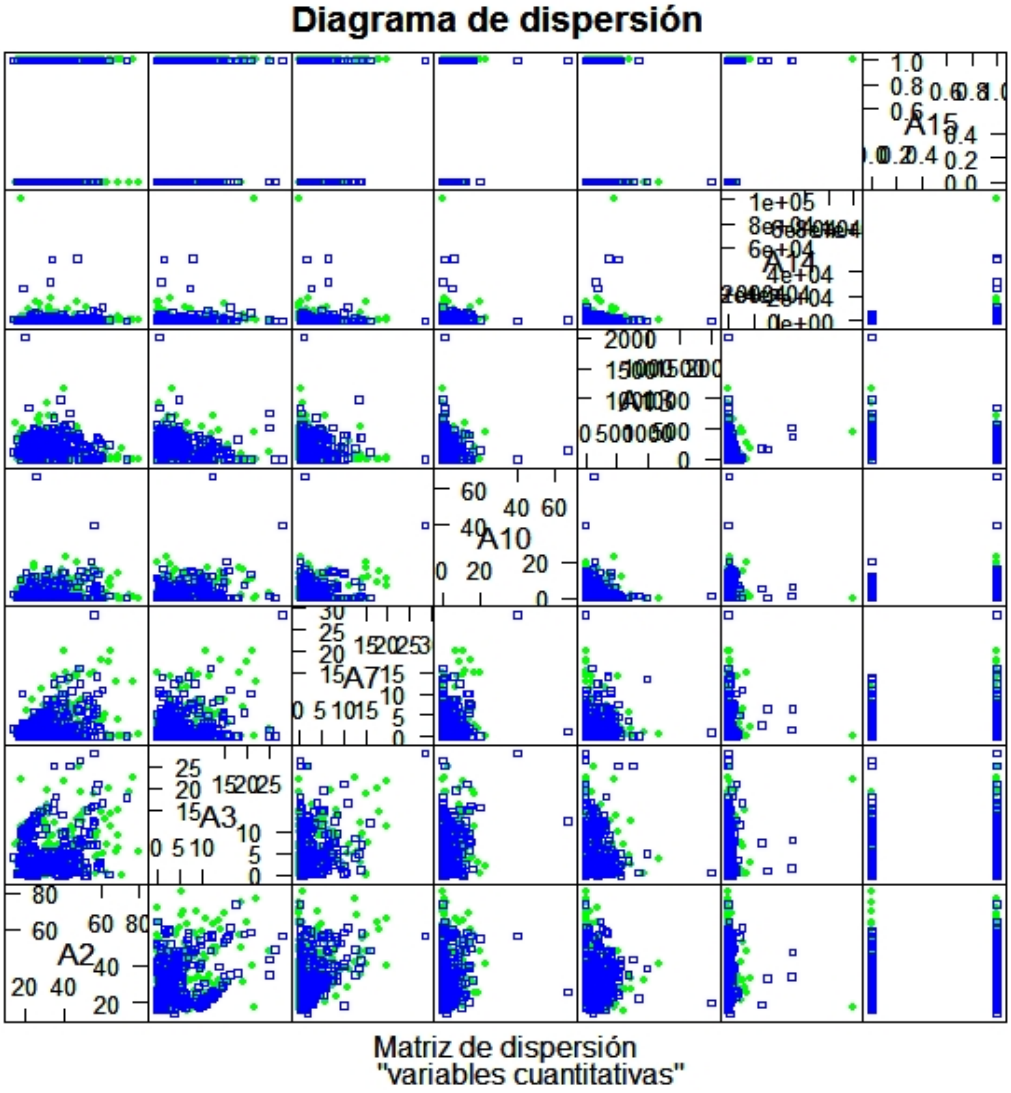

Figura 3.8: Diagrama de dispersión de las variables “Cuantitativas”

En la figura 3.8 no se distinguen los dos grupos esperados hay algunos datos que se dispersan de la acumulación de estos, pero pareciera que con todos los datos solo se puede hacer un grupo. Sigamos nuestro análisis a ver si logramos obtener la clasificación en dos grupos a los individuos que les otorgaremos la tarjeta de crédito y a los que se las negaremos.

Tenemos un gráfico que nos muestra estrellas de distintas especie.

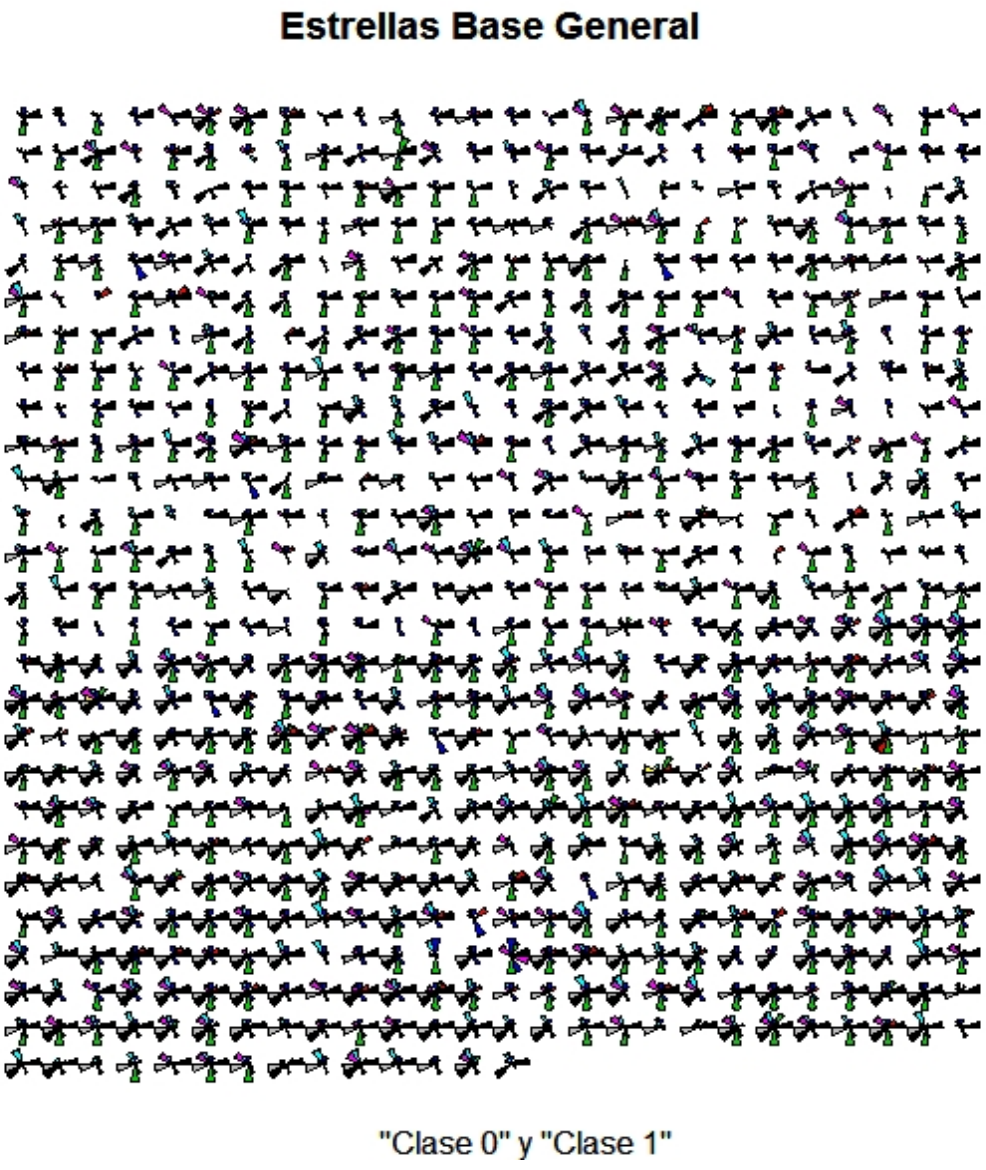

Figura 3.9: Estrellas de la clase 1 y clase 0

Aquí observamos estrellas de distintas especies, sería bueno observarlas e identificar la existencia de los dos grupos buscados los cuales son a los solicitantes que sí se les otorgó una tarjeta de credito mismos que pertencen a la clase "1" y a los que se les nego la tarjeta de crédito los cuales pertenecen a la clase"0".Estos datos están ordenados. Las primeras 383 estrellas pertenecen a los individuos a los cuales se les negó tarjeta de crédito y los 307 restantes son los individuos a los que sí se les otorgó la tarjeta de crédito. Y dada esta información sí es notoria la existencia de los dos grupos.

3.1.2. Matriz de Correlación

La matriz de correlación nos muestra si alguna de las variables está muy correlacionada con otra, si sucediera esto nos diría que podríamos excluir alguna de esas variables ya que dicha variable no nos estaría proporcionado información. Por otro lado también podemos observar cuales variables tienen correlación con la variables a explicar, en este caso podríamos pensar que estas variables son "más" significativas para elaborar nuestro modelo de análisis discriminante. Recordemos que esta variable es binaria, considera a los solicitantes que ya se les concedió una tarjeta de credito y a los que no se les concedió.

$$\begin{pmatrix} A_2 & A_3 & A_7 & A_{10} & A_{13} & A_{14} & A_{15} \\ 1.000 & 0.201 & \mathbf{0.393} & 0.186 & -0.077 & 0.019 & 0.162 \\ 0.201 & 1.000 & 0.299 & 0.271 & -0.222 & 0.123 & 0.206 \\ \mathbf{0.393} & 0.299 & 1.000 & \mathbf{0.322} & -0.076 & 0.051 & \mathbf{0.322} \\ 0.186 & 0.271 & \mathbf{0.322} & 1.000 & -0.120 & 0.064 & \mathbf{0.406} \\ -0.077 & -0.222 & -0.076 & -0.120 & 1.000 & 0.066 & -0.099 \\ 0.019 & 0.123 & 0.051 & 0.064 & 0.066 & 1.000 & 0.176 \\ 0.162 & 0.206 & \mathbf{0.322} & \mathbf{0.406} & -0.099 & 0.176 & 1.000 \end{pmatrix}$$

La matriz muestra que las variables "más" correlacionadas son las A_2 y A_7, así mismo tienen correlación la A_{10} y A_7, también podemos ver que las correlaciones de casi todas las variables son realmente pequeñas, es posible que no podamos hacer una reducción de variables. Mientras que las variables A_7 y A_{10} tienen alguna correlación con la variable que indica la distribución de los grupos, puede ser que estas variables sean las que nos proporcionen más información acerca de como separar a nuestros solicitantes de tarjetas de crédito.

3.2. Análisis Discriminante

En esta sección haremos la aplicación del análisis discriminante que vimos en el capítulo 2, en el contexto de *"Credit Scoring"*.

Entonces

- Tenemos 6 variables continuas (A_2, A_3, A_7, A_{10}, A_{13}, A_{14}) independientes entre si y una variable cualitativa (A_{15}).
- Tenemos dos categorías en la variable cualitativa, las cuales son: si le otorgamos una tarjeta de crédito a un solicitante (1) y si le negamos dicha tarjeta de crédito al solicitante (0).
- Necesitamos ver si las matrices de covarianza de cada grupo son aproximadamente iguales. Para verificar este supuesto consideremos la siguiente prueba de hipótesis.

3.2.1. Prueba de Hipótesis

A continuación presentamos la prueba de hipótesis de la estadística **Z**, si la estadística nos da un valor grande nos dejaría suponer que los grupos estan suficientemente separados en términos de su media. En este caso será mas probable que un nuevo solicitante de tarjeta de crédito pueda ser discriminado, con mejor éxito, en el grupo al que pertenezca.

Bajo la hipótesis $H_0 : \mu_1 = \mu_2$ y matriz de varianza-covarianza común el estadístico de prueba **Z** se distribuye como una distribución F con p y $n_1 + n_2 - p - 1$ grados de liberdad. Por lo que vimos en la sección de análisis discriminante tenemos lo siguiente:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T S^{-1}(\bar{x}_1 - \bar{x}_2) \qquad \Rightarrow \qquad \mathbf{Z} = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2$$

Para nuestros datos tenemos lo siguiente:

$$\bar{\mathbf{X}}_\mathbf{1} = \begin{pmatrix} 29.85 \\ 3.84 \\ 1.26 \\ 0.63 \\ 199.41 \\ 199.61 \end{pmatrix} \qquad \bar{\mathbf{X}}_\mathbf{2} = \begin{pmatrix} 33.71 \\ 5.90 \\ 3.43 \\ 4.61 \\ 164.80 \\ 2039.86 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 137.030 & 9.926 & 13.533 & 6.921 & -124.662 & -609.315 \\ 9.926 & 23.762 & 3.877 & 4.542 & -173.135 & 2256.768 \\ 13.533 & 3.877 & 10.049 & 3.117 & -25.473 & -92.547 \\ 6.921 & 4.542 & 3.117 & 19.770 & -66.384 & -195.283 \\ -124.662 & -173.135 & -25.473 & -66.384 & 29385.240 & 74708.988 \\ -609.315 & 2256.768 & -92.547 & -195.283 & 74708.988 & 2.635 * 10^7 \end{pmatrix}$$

Entonces

$$T^2 = \frac{383 * 307}{383 + 307} \begin{pmatrix} -3.852 \\ -2.065 \\ -2.170 \\ -3.974 \\ 34.608 \\ -1840.254 \end{pmatrix}^T \begin{pmatrix} 137.030 & 9.926 & 13.533 & 6.921 & -124.662 & -609.315 \\ 9.926 & 23.762 & 3.877 & 4.542 & -173.135 & 2256.768 \\ 13.533 & 3.877 & 10.049 & 3.117 & -25.473 & -92.547 \\ 6.921 & 4.542 & 3.117 & 19.770 & -66.384 & -195.283 \\ -124.662 & -173.135 & -25.473 & -66.384 & 29385.240 & 74708.988 \\ -609.315 & 2256.768 & -92.547 & -195.283 & 74708.988 & 2.635 * 10^7 \end{pmatrix}^{-1} \begin{pmatrix} -3.852 \\ -2.065 \\ -2.170 \\ -3.974 \\ 34.608 \\ -1840.254 \end{pmatrix}$$

$$\Rightarrow \qquad T^2 = 206.2678$$

$$\therefore \qquad \mathbf{Z} = \frac{383 + 307 - 6 - 1}{(383 + 307 - 2)6} \quad 206.2678 = 34.12813$$

Para poder comparar nuestra estadística Z y así saber si aceptamos o rechazamos nuestra hipótesis nula, necesitamos el cuantil de la distribución F:

$$F_{0.95;(6,683)} = 2.111836 \qquad \Rightarrow \qquad 34.12813 > 2.111836$$

$\therefore$ rechazamos la hipótesis nula, que quiere decir que el vector de medias para cada grupo no son parecidos, por lo que no podemos afrimar que tenga matriz de varianza–covarianza común para los dos grupos. No podríamos suponer que los grupos estan suficientemente separados.

3.2.2. Aplicación: Utilizando rutina de R

Vamos aplicar la función discriminante que nos proporciona **R** para analizar la clasificación. Ocupemos las 6 variables continuas para intentar explicar un poco más de la proporción de varianza, entonces comenzaremos aplicando la función discriminante con todas las variables.

Tomemos una submuestra de 483 de los 690 datos que tenemos, para después clasificar los 207 restantes, de los cuales 115 pertenecen a la clase "0" y 92 a la clase "1" (aproximadamente el 30 % de cada clase) para evaluar el error de clasificación. Consideramos que cada individuo tiene la misma probabilidad de pertecer a cada grupo, esto debido a que no se tiene información al respecto, sin olvidar que en *Credit Scoring* el costo de conceder un crédito a un candidato con mal riesgo de crédito debiera ser signaficativamente mayor que el costo de denegar un crédito a un candidato con un buen riesgo de crédito.

La función discriminante arrojada por **R** es la que se presenta a continuación:

$$\begin{aligned} A_{15} \;=\; & 2.74\text{x}10^{-3}A_2 + 1.46\text{x}10^{-2}A_3 + 1.38\text{x}10^{-1}A_7 + 1.55\text{x}10^{-1}A_{10} \\ & -6.71\text{x}10^{-4}A_{13} + 6.58\text{x}10^{-5}A_{14} \end{aligned}$$

Observemos que la variable A_{14} es la que tiene el coeficiente más pequeño al igual que la variable A_{13} lo que nos indica que estas variables son las que menos información le aportan al modelo.

Presentamos los histogramas de la figura 3.10 que nos representan la separación de los grupos bajo el modelo de análisis discriminante:

De la figura 3.10 es muy claro que con una alta probabilidad se hará una clasificación erronea, por lo que es muy probable que le otorguemos una tarjeta de crédito a una persona con un mal riesgo crediticio o le neguemos el crédito a una persona que tenga un riesgo crediticio bueno. Veamos cómo clasifica, los 207 datos restantes.

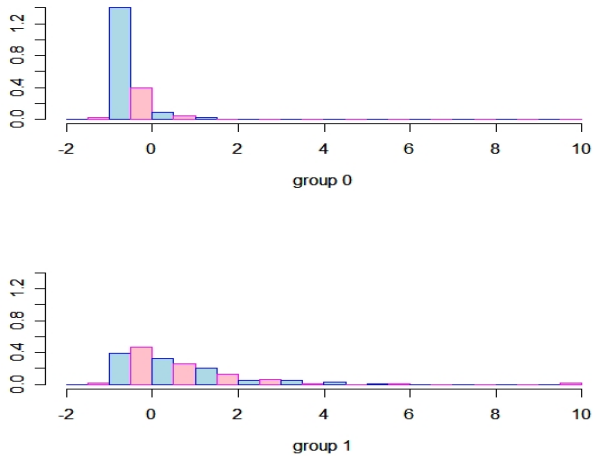

Figura 3.10: Grupos de Clasificación

Clasificación

0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0
1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 0 1 0 1 1 0 0 1 1 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 0 0 0 1 1 1 1 1
1 1 1 1 1 0 1 0 0 0 1 0 1 1 1 1 0 1 0 0 1 0 0 1 1 1 0 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 0 0 1 1 0

Clasifica 13 de la clase "0" en la clase "1" así mismo clasifica 35 de la clase "1" en la clase "0". Es decir de 115 de la clase "0" solo clasifica 102 bien y de 92 de la clase "1" clasifica bien 57. Por lo tanto con una mayor probabilidad negaremos una tarjeta de crédito a solicitantes con un buen riesgo de crédito mientras será un poco menos probable otorgarle una tarjeta de crédito a una persona con un historial crediticio malo.

A continuación ajustaremos una vez más el modelo quitando la variable A_{14} ya que como habíamos visto nos aporta menos información. Así mismo haremos el análisis con distintas variables, quitando de manera iterativa aquella variable que menos información nos aporta, para ver si encontramos una función que clasifique con menos error. Los resultados se muestran en el cuadro 3.1 en el que se observan los coeficientes[2] y los errores cometidos expresados en números y en porcentajes con respecto al total de errores que se cometen.

Dado el cuadro 3.1 nos podemos quedar con todas las variables excepto la variable A_{14}

[2]Son los eigen-vectores asociados al eigen-valor más grande

Variables	**Coeficientes**	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
$\mathbf{A_2}$	0.0024668742					
$\mathbf{A_3}$	0.0246295426					
$\mathbf{A_7}$	0.1436205750	34	13	36.95	11.3	22.70
$\mathbf{A_{10}}$	0.1630534364					
$\mathbf{A_{13}}$	-0.0004874283					
$\mathbf{A_2}$	0.002691838					
$\mathbf{A_3}$	0.028120474					
$\mathbf{A_7}$	0.143459902	34	13	36.95	11.3	22.70
$\mathbf{A_{10}}$	0.164483491					
$\mathbf{A_3}$	0.02865701					
$\mathbf{A_7}$	0.14685322	34	13	36.95	11.3	22.70
$\mathbf{A_{10}}$	0.16485036					

Cuadro 3.1: Errores en diferentes modelos por grupo(porcentajes, totales)

para conservar más información ya que quitándola nos mejora el error de clasificación.

Lo que nos dice la tabla es que el modelo de análisis discriminente bajo las variables indicadas en el cuadro 3.1 comete pocos errores de otorgarle una tarjeta de crédito a un solicitante con un mal riesgo de crédito sin embargo; no le otorga la tarjeta de crédito a solicitantes con un buen riego de crédito.

Ahora presentamos el gráfico en el que se contemplan todas las variables excepto la A_{14} este es el que presenta menos error de clasificación.

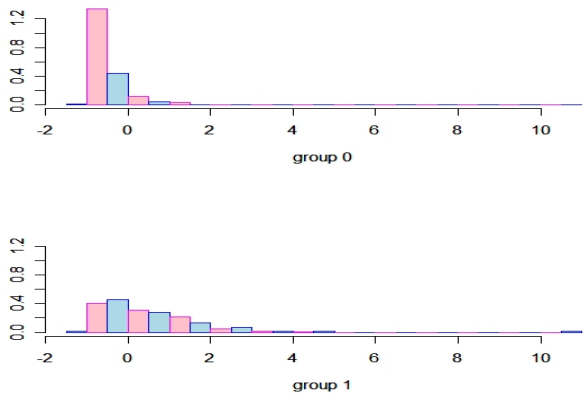

Figura 3.11: Clasificación de Grupos

En la figura 3.11 podemos apreciar cómo los grupos se empalman claramente en un

intervalo considerable lo que hace que con un error grande se realice una mala clasificación de los individuos a los que se les concede o se les niega la tarjeta. Razón por la cual se comete un error grande de clasificación.

Por lo tanto la función discriminante queda como sigue:

$$\begin{aligned} A_{15} &= 0.0024668742A_2 + 0.0246295426A_3 + 0.1436205750A_7 \\ &\quad +0.1630534364A_{10} - 0.0004874283A_{13} \end{aligned}$$

Podríamos decir que esto hace que la entidad financiera se asegure de que sus beneficios se maximicen y no pone en riesgo su capital. Por que puede estar otorgando una tarjeta de crédito a personas que sí se financiarían con la tarjeta y que son poco riesgosos. Una persona que nunca se financie con la tarjeta no le va a dar nada de rentabilidad, entonces prefiere negársela. Podemos decir que esta entidad financiera decide si le otorga o no se lo otorga deacuerdo a la base del rendimiento que espera obtener. Sería más eficaz el hecho de conceder o de rechazar una tarjeta de crédito si conociéramos la probabilidad que tiene cada solicitante de presentar problemas de incumplimiento con los pagos.

3.3. Regresión Logística

En esta sección consideremos un modelo de regresión logística. La base consta de 8 variables cualitativas y para estas variables es necesario usar esta técnica.

Las variables cualitativas a tratar son: A_1, A_4, A_5, A_6, A_8, A_9, A_{11}, A_{12} las cuales se usaran para poder explicar la variable A_{15} que es la que se refiere a si le concedemos la tarjeta de crédito a un nuevo solicitante o se la rechazamos.

- Las variables A_1, A_8, A_9, A_{11}, son variables con dos categorías es decir son variables dicotómicas

- Mientras que las variables A_4, A_5, A_6, A_{12} tienen 3, 14, 8 y 3 categorías respectivamente lo que nos conduce a incluir variables ***"dummy"***

    1. Para las variables A_4 y A_{12} que tienen 3 cateogías incluiremos 2 variables ***"dummy"***
    2. Para la A_5 se incluiran 13 variables ***"dummy"***
    3. Y para la variable A_6 incluiremos 7 variables ***"dummy"***

de esta manera haremos que cada categoría esté de forma individual en nuestro modelo.

3.3.1. Modelo de Regresión: Utilizando R

Como ya lo mencionamos aplicaremos la rutina de **R** que nos calcula una regresión para modelos generalizados, así mismo la estimación de los parámetros[3] para un modelo de regresión logística.

Tomemos la misma muestra que en el caso de análisis discriminante para nuestro modelo de regresión logística.

Hay que tener en cuenta que solo estamos considerando las variables cualitativas.

Considerando todo lo anterior y que los coeficientes dados por la rutina son los llamados *logit*, la función resultante es:

$$\begin{aligned} log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = \; & - \; 5.506 - 0.243A_1 + 0.828A_{4.1} + 31.606A_{4.2} - 11.429A_{5.1} - 11.153A_{5.2} \\ & - \; 10.989A_{5.3} - 30.973A_{5.4} - 11.033A_{5.5} - 9.709A_{5.6} - 10.476A_{5.7} \\ & - \; 10.102A_{5.8} - 9.361A_{5.9} - 10.868A_{5.10} - 11.475A_{5.11} - 9.603A_{5.12} \\ & - \; 8.387A_{5.13} + 11.528A_{6.1} + 31.129A_{6.2} + 12.009A_{6.3} + 11.826A_{6.4} \\ & + \; 16.397A_{6.5} + 12.874A_{6.6} + 10.743A_{6.7} + 4.050A_8 + 1.207A_9 \\ & + \; +0.079A_{11} + 0.080A_{12.1} + 3.328A_{12.2} \end{aligned}$$

$$\pi(x) = \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}},$$

donde:

$$\mathbf{X} = \begin{pmatrix} 1 \\ A_1 \\ A_{4.1} \\ \vdots \\ A_{12.1} \\ A_{12.2} \end{pmatrix} \qquad \beta = \begin{pmatrix} -5.506 \\ -0.243 \\ 0.828 \\ \vdots \\ 0.080 \\ 3.328 \end{pmatrix}$$

Es interesante mencionar que el algoritmo utilizado para encontrar los coeficientes realizó 16 iteraciones para llegar a las estimaciones anteriores.

Analicemos en los siguientes gráficos el comportamiento de la regresión logística, para poder apreciar qué tan adecuado es el modelo que se ha obtenido.

En el gráfico de la figura 3.12 se aprecian los residuos del modelo. Lo que esperábamos

[3]Mediante iteraciones de mínimos cuadrados reponderados ("glm.fit")

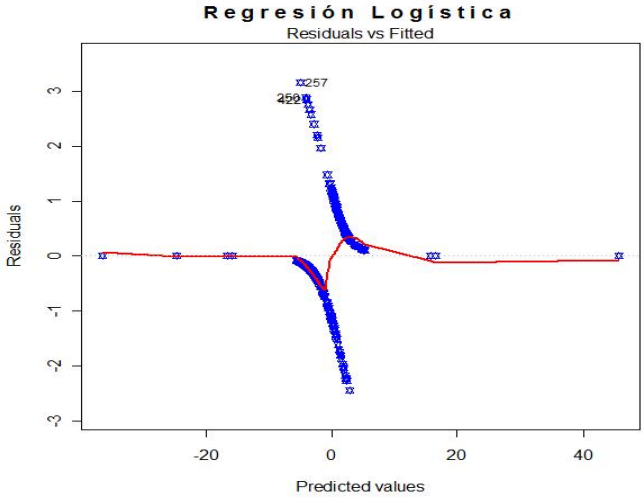

Figura 3.12: Residuos

es que estos se pegaran a las bandas, sin embargo del lado derecho tenemos datos que se dispersan de los demás, podría ser un buen modelo si los residuos se encontraran más cerca del cero. Posiblemente el modelo no es el mejor.

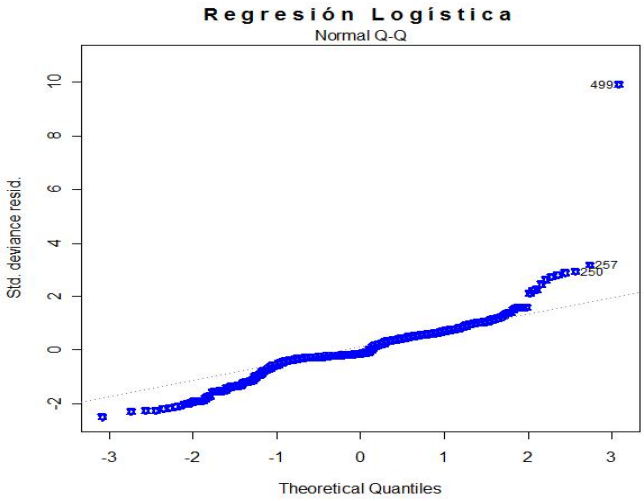

Figura 3.13: QQ-plot

En el gráfico de la figura 3.13 se muestran los cuantiles teorícos de la regresión logística de este podemos apreciar que alrededor del cero se ajustan a la línea, sin embargo no se ajustan a la línea de 45 ° para ambos extremos estos se salen considerablemente de las bandas, hubieramos preferido que para todo cuantil se ajustara, ya que esto nos ayudaría a determinar un buen ajuste. Además se observa un dato muy disperso con respecto a los demás. Una prueba más para que sigamos pensando que algo anda mal

en el modelo. Posiblemente este modelo no sea el adecuado para hacer la clasificación entre nuestros solicitantes de tarjetas de crédito.

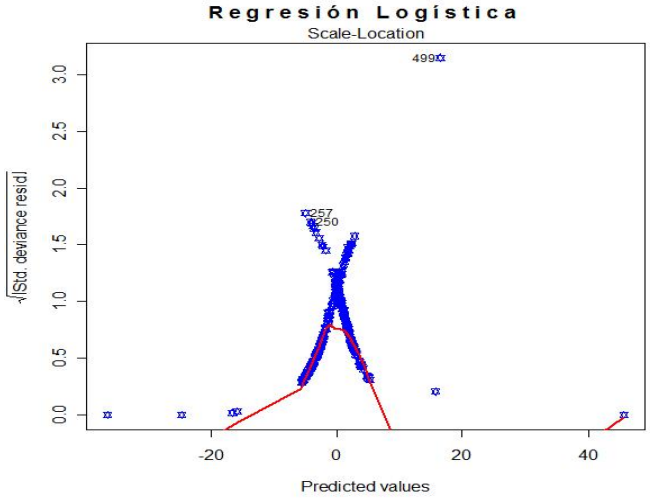

Figura 3.14: Devianza

En el gráfico de la figura 3.14 se observan los predictores lineales contra la raíz cuadrada de la devianza estandar de los residuales. Lo que esperabamos era que las bandas se cerraran alrededor del cero y que los predictores del modelo se ajustaran dentro de las bandas. Sin embargo, algunos predictores se salen de las bandas además de que presenta un dato atípico, como las bandas se concentran alrededor del cero nos indica que la devianza esta alrededor del cero si no tuvieramos los datos atípicos esto nos llevaría a obtener una mejor clasificación.

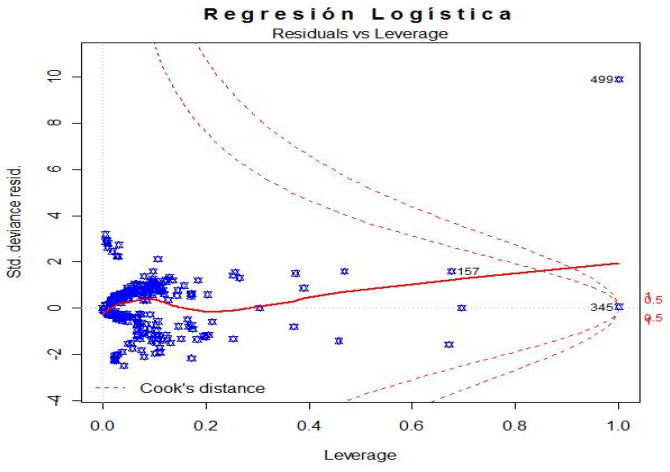

Figura 3.15: Leverage

En el gráfico de la figura 3.15 se observa que existe una separación no tan clara dos grupos, hay datos que si presentan características diferentes a los de la otra clase, sin embargo la mayoria de los datos se concentran alrededor del cero.

Por otro lado **R** nos arroja pruebas de significancia las cuales nos indican si el modelo ajustado es bueno o no:

	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)
β_0	-5.506	0.98619	-5.583	$2.36\text{x}10^{-8}$
A_1	-0.243	0.34137	-0.712	0.476419
$A_{4.1}$	0.828	0.37451	2.212	0.026971
$A_{4.2}$	31.606	2780.81596	0.011	0.990932
$A_{5.1}$	-11.429	1385.61535	-0.008	0.993419
$A_{5.2}$	-11.153	1385.61529	-0.008	0.993578
$A_{5.3}$	-10.989	1385.61503	-0.008	0.993672
$A_{5.4}$	-30.973	2151.06936	-0.014	0.988512
$A_{5.5}$	-11.033	1385.61520	-0.008	0.993647
$A_{5.6}$	-9.709	1385.61536	-0.007	0.994409
$A_{5.7}$	-10.476	1385.61519	-0.008	0.993968
$A_{5.8}$	-10.102	1385.61525	-0.007	0.994183
$A_{5.9}$	-9.361	1385.61580	-0.007	0.994610
$A_{5.10}$	-10.868	1385.61517	-0.008	0.993742
$A_{5.11}$	-11.475	1385.61622	-0.008	0.993392
$A_{5.12}$	-9.603	1385.61536	-0.007	0.994470
$A_{5.13}$	-8.387	1385.61547	-0.006	0.995171
$A_{6.1}$	11.528	1385.61646	0.008	0.993362
$A_{6.2}$	31.129	2151.06887	0.014	0.988454
$A_{6.3}$	12.009	1385.61501	0.009	0.993085
$A_{6.4}$	11.826	1385.61513	0.009	0.993190
$A_{6.5}$	16.397	1385.61605	0.0126	0.990558
$A_{6.6}$	12.874	1385.61503	0.009	0.992586
$A_{6.7}$	10.743	1385.61617	0.008	0.993814
A_8	4.050	0.39799	10.176	$< 2\text{x}10^{-16}$
A_9	1.207	0.32645	3.696	0.000219
A_{11}	0.079	0.31434	0.252	0.800888
$A_{12.1}$	0.080	0.59552	0.135	0.892797
$A_{12.2}$	3.328	1.22837	2.709	0.006747

Cuadro 3.2: Resumen de las pruebas de significancia

El cuadro 3.2 nos permite revisar cada uno de nuestros coeficientes por separado y probar la hipótesis nula $H_0 : \beta_i = 0$ para cada i. Recordemos que las variables A_4,

A_5, A_6 y A_{12} tienen más de dos categorías, debido a esto se observan más variables y cada una se representa con el número de la variable y el número de la categoría. La última columna nos proporciona el p-valor de la prueba. Si comparamos con un nivel de significancia $\alpha = 5\,\%$ entonces resulta que para A_1, A_4, A_5, A_6, A_{11} y A_{12} no se rechaza la hipótesis de que su coeficiente sea 0, mientras que para las demás variables se rechaza la hipótesis, debido a esto las podríamos excluir de nuestro modelo. Aunque observamos que A_4 tiene una categoría en la cuál se rechaza la hipótesis, sin embargo es necesario eliminar toda la variable, podríamos no excluirla del modelo y ver si al excluir las demás está se puede incluir en el modelo.

Por otro lado también nos proporciona el *índice de Akaike* (AIC) [4]=356.93. Esté índice nos indica qué tan adecuado es nuestro modelo comparado con algún otro, buscamos un índice menor, debido a que el modelo óptimo es el que suele tener valores más altos para la verosimilitud y menos parámetros.

Las clasificaciones para este modelo quedan como sigue:
0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 0 1
0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 11 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1

En esta clasificación de los solicitantes sabemos que los primeros 115 pertenecían a la clase "0" clasifica 24 individuos incorrectamente, por lo que tiene un 21 % de error. Mientras que de 92 que pertenecen a la clase "1", 7 son mal clasificados, es decir, de los 92 individuos a los cuales se les otorgo el crédito indica que a 85 si se les debe otorgar dicha tarjeta de crédito y a 7 no se le debe conceder la tarjeta por estos datos se clasifica con un 7.6 % mal.

Como podemos ver tiene mayor error al momento de decidir sobre a quién rechazar una tarjeta, ya que con mayor probabilidad le otorga tarjetas de crédito a solicitantes con un riesgo crediticio malo y esto significara pérdida para la entidad financiera. Por otro lado está perdidendo clientes que sí le darían rendimientos a la entidad. Recordemos que el costo de otorgarle una tarjeta de crédito a alguien que es muy riesgoso o que no devolvera la cantidad otorgada es considerablemente mayor a no concederle dicha tarjeta a un solicitante que no sea nada riesgoso.

Después de ver que este modelo realmente no es el que mejor nos clasifica a los individuos busquemos si existe otro modelo que nos realice una mejor clasificación de dichos solicitantes.

[4] AIC=-2(log(verosimilitud)– número de parámetros en el modelo)

3.3.2. Otros modelos de regresión logística

Se utilizó la rutina de **R** primero excluyendo las variables A_1, A_5, A_6 A_{11} y A_{12} que eran las variables no significativas y obtuvimos los siguientes resultados:

2° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)	AIC
β_0	-3.6390	0.4010	-9.074	$< 2\text{x}10^{-}16$	
$A_{4.1}$	0.8661	0.3331	2.600	0.00932	
$A_{4.2}$	18.2050	624.1940	0.029	0.97673	358.82
A_8	3.6351	0.3164	11.489	$< 2\text{x}10^{-}16$	
A_9	1.1668	0.2785	4.189	$2.8\text{x}10^{-}5$	

Cuadro 3.3: Pruebas de significancia del 2° Modelo

En el cuadro 3.3 se muestran las pruebas y el índice de AIC, una vez más nos indica que para la variable A_4 no se rechaza la hipótesis y para las demás variables sí se rechaza la hipótesis $H_0 : \beta_i = 0$, como vemos el AIC es mayor que el anterior pero ¿Qué tal clasificará este modelo?

Se hizo una vez más la clasificación para este modelo y desafortunadamente se obtuvieron casi los mismos resultados de clasificación ya qué se incrementó el error en la clase "1" pero no se disminuyo el error en la clase "0".

Una vez más se utilizó la rutina de **R** ahora excluyendo la variable A_4. Se obtuvierón los siguientes resultados:

3° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)	AIC
β_0	-2.8374	0.2729	-10.396	$< 2e-16$	
A_8	3.5103	0.3003	11.691	$< 2e-16$	373.44
A_9	1.1757	0.2720	4.322	$1.55e-05$	

Cuadro 3.4: Pruebas de significancia del 3° Modelo

En el cuadro 3.4 se rechaza la hipótesis de que sus coeficientes sean "0" para todas las variables utilizadas con el nivel de significancia que sugerimos; sin embargo, el AIC no lo mejora en lugar de disminuirlo lo incrementa lo que puede hacer que este modelo no sea mejor que el anterior. A pesar de esto se realizó la clasificación y mejora en poco ya que ahora solo tiene 6 errores en la clase "1", sin embargo es poco significante.

Por lo tanto el modelo que adoptamos como un "mejor modelo" es:

$$\begin{aligned} log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = \; & - \; 5.506 - 0.243A_1 + 0.828A_{4.1} + 31.606A_{4.2} - 11.429A_{5.1} - 11.153A_{5.2} \\ & - \; 10.989A_{5.3} - 30.973A_{5.4} - 11.033A_{5.5} - 9.709A_{5.6} - 10.476A_{5.7} \\ & - \; 10.102A_{5.8} - 9.361A_{5.9} - 10.868A_{5.10} - 11.475A_{5.11} - 9.603A_{5.12} \\ & - \; 8.387A_{5.13} + 11.528A_{6.1} + 31.129A_{6.2} + 12.009A_{6.3} + 11.826A_{6.4} \\ & + \; 16.397A_{6.5} + 12.874A_{6.6} + 10.743A_{6.7} + 4.050A_8 + 1.207A_9 \\ & + \; +0.079A_{11} + 0.080A_{12.1} + 3.328A_{12.2} \end{aligned}$$

Escogimos este modelo como el "mejor", por que, todas las variables aportan información para la clasificación de acuerdo a las pruebas que se realizarón, el AIC disminuye un poco a comparación del primer modelo.

3.4. Análisis Discriminante y Regresión Logística

Efron (1975) demostró que cuando los datos son normales multivariantes y se estiman los parámetros en la muestra, la función discriminante lineal de Fisher funciona mejor que la regresión logística. Sin embargo, no es nuestro caso ya que cuando hicimos el análisis exploratorio en los histogramas no se nota normalidad en nuestras variables.

Como hemos visto en las secciones anteriores de este capítulo la clasificación no ha sido buena sin embargo el análisis discriminante sólo lo utilizamos con las variables continuas que teníamos en nuestra base de datos y la regresión logística sólo para las variables cualitativas. Esto con la finalidad de análizar toda la información qué se tiene. Entonces, hagamos algunas observaciones sobre las clasificaciones que hicieron ambos modelos, intentando observar las diferencias de ambos quizá con la unión de los dos métodos podríamos llegar a un mejor resultado. A continuación presentaremos la clasificación de los métodos utilizados así como los errores de clasificación y los porcentajes de error.

Clasificación obtenida del análisis discriminante:

0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 0 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 0 1 0 1 1 1 1 1 1 0 0 1 0 0 1 1 1 0 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 0 0 0 0 1 1 0

Variables	**Coeficientes**	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
$\mathbf{A_2}$	0.0024668742					
$\mathbf{A_3}$	0.0246295426					
$\mathbf{A_7}$	0.1436205750	34	13	36.95	11.3	22.70
$\mathbf{A_{10}}$	0.1630534364					
$\mathbf{A_{13}}$	-0.0004874283					

Cuadro 3.5: Errores de clasificación utilizando Análisis Discriminante

Clasificación obtenida de la regresión logística:
0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 0 1
0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
11 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1

Variables	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
Todas las categóricas	7	24	7.6	20.86	15

Cuadro 3.6: Errores de clasificacion utilizando Regresión Logística

Como podemos ver los porcententajes de clasificación son diferentes hay que tener encuenta que no se utilizaron las mismas variables para hacer las clasificaciones pero sí la misma muestra. En el cuadro 3.5 podemos observar que el análisis discriminante tuvo más error de clasificación en la clase "1" ya que de 92 que pertenecian a esta clase solo clasifico 58 bién lo que provoca un error de clasificacion de casi el 37 %. Por otro lado en el cuadro 3.6 se observa que la regresión logística tuvo mayor error de clasificación en la clase "0" ya que de 115 clasifica 91 bien y los restantes 24 mal provocando un error de clasificación del 21 %.

Observemos que en general la regresión logística para las variables utilizadas es la que menos se equivoca. Sin embargo, si unimos ambos métodos para tomar la desición podríamos ser un poco más eficaces. Es decir, usar la regresión logística para aceptar créditos y el análisis discriminante para rechazar las tarjetas de crédito y así se toma en cuenta toda la información.

En general podemos decir que ninguno de los dos métodos supera al otro de manera uniforme ya que los resultados dependeran de la base de datos que se tenga.

3.5. Regresión Logística: Todas las variables

En esta sección utilizaremos regresión logística como método de decisión. Ya que podemos utilizar tanto variables cuantitativas como cualitativas debido a esto haremos uso de todas las variables, primero con la misma muestra de entrenamiento utilizada en la sección anterior y después clasificacando todos los datos.

3.5.1. Muestra de entrenamiento

Como ya mencionamos haremos uso de todas las variables, observaremos la forma de clasificación que se obtiene si utilizamos todas las variables, así mismo veremos si el modelo es bueno y si no lo es intentaremos mejorarlo, excluyendo las variables que no nos ayuden a realizar una buena clasificación, es decir, no sean significantes.

Se tomó una submuestra de 483 de los 690 datos que tenemos, para después clasificar los 207 restantes y verificar que clasifique bien, de los cuales 115 pertenecen a la clase "0" y 92 a la clase "1" (aproximadamente el 30 % de cada clase).

Recordemos que haremos uso de la rutina de **R** para ajustar la regresión logística. Así mismo recordemos que esta rutina nos arroja los llamados "logit". La función resultante es la mostrada a continuación.

$$\begin{aligned} log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = & - 5.35 - 0.26A_1 + 0.01A_2 - 0.04A_3 + +0.87A_{4.1} + 31.64A_{4.2} \\ & - 11.33A_{5.1} - 11.21A_{5.2} - 10.99A_{5.3} - 30.76A_{5.4} - 10.90A_{5.5} \\ & - 9.43A_{5.6} - 10.44A_{5.7} - 9.85A_{5.8} - 9.18A_{5.9} - 10.75A_{5.10} \\ & - 11.75A_{5.11} - 9.13A_{5.12} - 7.77A_{5.13} + 11.85A_{6.1} + 30.60A_{6.2} \\ & + 12.26A_{6.3} + 11.98A_{6.4} + 16.74A_{6.5} + 13.29A_{6.6} + 10.32A_{6.7} \\ & + 0.01A_7 + 4.006A_8 + 0.61A_9 + 0.12A_{10} + 0.15A_{11} - 0.04A_{12.1} \\ & + 3.16A_{12.2} - 0.003A_{13} \end{aligned}$$

$$\pi(x) = \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}},$$

A continuación presentaremos algunos gráficos para intentar apreciar qué tan adecuado es el modelo que tenemos.

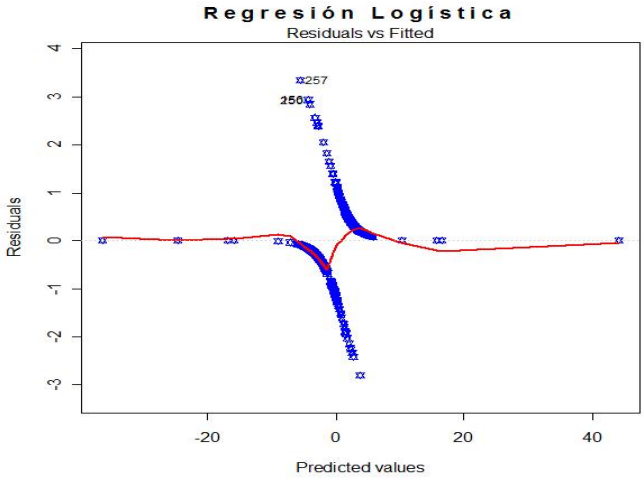

Figura 3.16: Residuos del modelo

En el gráfico de la figura 3.16 se observan los residuos del modelo de regresión logística, lo que se observa es que estos se tienen datos atípicos que se salen de la bandas y que de los datos que se aprecian se concentra la mayoría alrededor del cero, los datos atípicos que se aprecian dentro del modelo pueden influir en el modelo, es posible que estos dato esten perjudicando un poco el modelo.

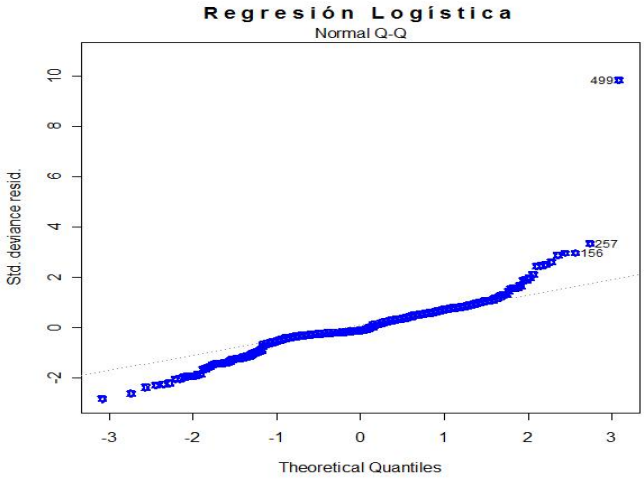

Figura 3.17: QQ - plot

En el gráfico de la figura 3.17 observamos los cuantiles teorícos del modelo, lo que podemos apreciar que ambas colas estan pesadas ya que de ambos extremos no se ajustan a la línea de 45 °, pero en general son pocos los datos que no se ajustan a esta línea, ya que la mayoría de los datos se concentran en la parte media. Una vez más se pueden apreciar los datos atípicos que aparecieron en el gráfico anterior.

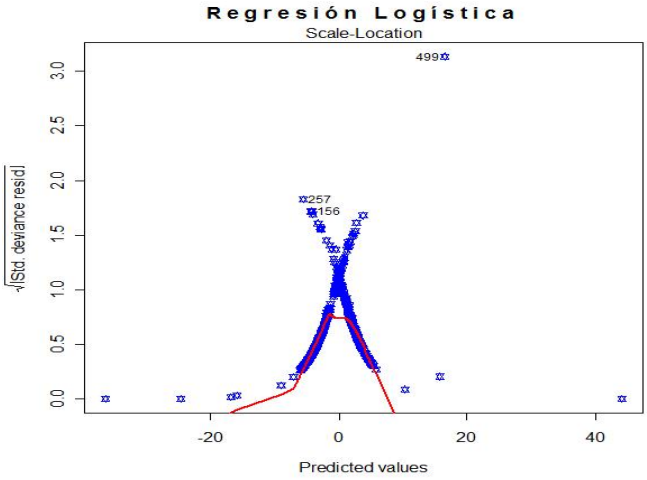

Figura 3.18: Devianza

El gráfico de la figura 3.18 nos muestra los predictores contra la devianza, como podemos observar los datos no se salen de las bandas, pero de ambos lados hay datos que se alejan de los demás. Por otro lado en la parte central no se ve un ajuste en las bandas donde se intersectan. Lo que nos hubiera agradado era que la línea se ajustara a ambas bandas para poder tener una mayor evidencia de que es un buen modelo.

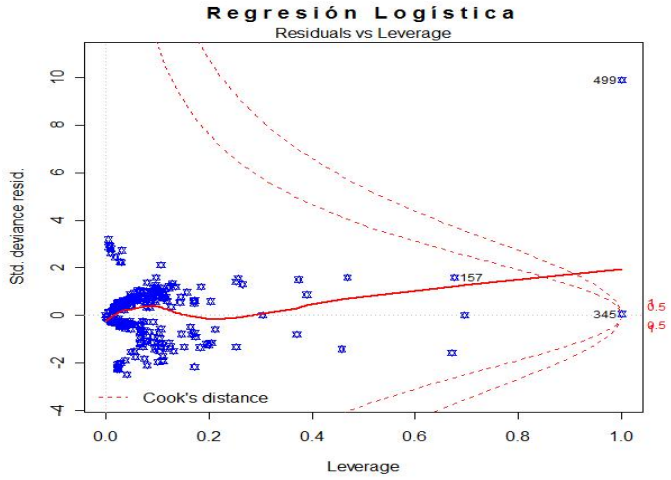

Figura 3.19: Palanca

En el gráfico de la figura 3.19 se ve una separación en los grupos, sin embargo alrededor del cero los datos se agrupan de tal manera que parece ser sólo un grupo. Existen datos que presentan diferencias para poder realizar la clasificación sin embargo no son suficientes para poder ajustar un "buen" modelo.
R arroja pruebas para evaluar el ajuste, que observaremos a continuación. En el cuadro

	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)
β_0	-5.35	1.192	-4.491	$7.1\text{x}10^{-6}$
A_1	-0.26	0.3534	-0.730	0.46537
A_2	0.01	0.01542	0.640	0.52221
A_3	-0.04	0.03429	-1.059	0.28968
$A_{4.1}$	0.87	0.03918	2.217	0.02664
$A_{4.2}$	31.64	2806	0.011	0.99100
$A_{5.1}$	-11.33	1416	-0.008	0.99362
$A_{5.2}$	-11.21	1416	-0.008	0.99369
$A_{5.3}$	-10.99	1416	-0.008	0.99381
$A_{5.4}$	-30.76	2166	-0.014	0.98867
$A_{5.5}$	-10.90	1416	-0.008	0.99386
$A_{5.6}$	-9.43	1416	-0.007	0.99469
$A_{5.7}$	-10.44	1416	-0.007	0.99412
$A_{5.8}$	-9.85	1416	-0.007	0.99445
$A_{5.9}$	-9.18	1416	-0.006	0.99483
$A_{5.10}$	-10.75	1416	-0.008	0.99394
$A_{5.11}$	-11.75	1416	-0.008	0.99338
$A_{5.12}$	-9.13	1416	-0.006	0.99483
$A_{5.13}$	-7.77	1416	-0.005	0.99563
$A_{6.1}$	11.85	1416	0.008	0.99333
$A_{6.2}$	30.60	2166	0.014	0.98873
$A_{6.3}$	12.26	1416	0.009	0.99310
$A_{6.4}$	11.98	1416	0.008	0.99325
$A_{6.5}$	16.74	1416	0.012	0.99057
$A_{6.6}$	13.29	1416	0.009	0.99252
$A_{6.7}$	10.32	1416	0.007	0.99419
A_7	0.01	0.0545	0.207	0.83631
A_8	4.006	0.4271	9.380	$< 2\text{x}10^{-16}$
A_9	0.61	0.4445	1.381	0.16721
A_{10}	0.12	0.06746	1.755	0.07928
A_{11}	0.15	0.3346	0.435	0.66337
$A_{12.1}$	-0.04	0.5970	-0.061	0.95121
$A_{12.2}$	3.16	1.228	2.572	0.01011
A_{13}	-0.003	0.001	-2.979	0.00289

Cuadro 3.7: Pruebas de significancia utilizando todas las varables

3.7 se muestran los coeficientes del modelo de regresión logística, el error estandar, y las pruebas de significancia para $H_0 : \beta_i = 0$ para $i = 1, \ldots, 13$. La última columna nos proporciona el p-valor de la prueba. Si comparamos con un nivel de significancia $\alpha = 5\,\%$, observamos que para las variables A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_9, A_{10}, A_{11} y A_{12} no se rechaza la hipótesis nula de que el coeficiente de cada una de ella sea 0, mientras que para las demás variables β_0 *(intercepto)*, A_8 y A_{13} se rechaza la prueba, es decir, su coeficiente es distinto de cero, por lo tanto podemos decir que dichas variables son significantes.

Por otro lado también obtenemos el índice de Akaike AIC = 351.43, recordemos que buscamos un índice pequeño a comparación del de otro modelo ya que nos indicaría que es un mejor modelo por que su verosimilitud es mayor y por lo tanto podremos decir que el modelo es *parsimonioso*. Podemos hacer otro modelo sin las variables que no rechazaron la prueba y ver si el AIC es menor, pero observemos que para la variable A_{10} no se rechaza la prueba con una diferencia poco significantiva de nuestro p-valor, para no deshacernos de tanta información dejaremos esas variables esperando que mejoren en otro modelo.

A continuación se muestra la clasificación que realizó el modelo, recordemos que los primeros 115 pertenecen a la clase "0" y que los 92 restantes a la clase "1".

Clasificación

0 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 1 1 0 0 0 0
1 0 0 1 0 1 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0
0 0 1 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1

La clasificación presenta 32 errores en total que representa un 15.45 % de la muestra de prueba total de los cuales 25 errores pertenecen a la clase "0" y los 7 restantes a la clase "1". Es decir, de los 115 que pertenecian a la clase"0", a 90 clasifica bien y el 21.73 % mal, mientras que de 92 de la clase"1" clasifica 85 bien y el 7.6 % mal.

Podemos ver que tiene más error al momento de rechazar una tarjeta de crédito, acción que perjudica a la entidad financiera ya que estaría otorgandole la tarjeta a quienes pueden no pagar el crédito que utilicen. Por otro lado rechaza algunas tarjetas a quienes podrían darle rentabilidad a la entidad financiera. Sería bueno que buscáramos un modelo que disminuyera el error en otorgarle una tarjeta de crédito a quienes no le daran rentabilidad a la entidad y le causen pérdidas.

Otros modelos de regresión logística

Después de observar que el modelo presenta errores intentaremos realizar una vez más el modelo esperando que mejore la clasificación. En particular buscamos que el modelo disminuya el error en la clase "0" ya que en el modelo anterior existe una mayor probabilidad de equivocarse al momento de otorgarle una tarjeta de crédito a quien tiene un mal riesgo de crédito.

Se utilizó la rutina de **R** primero excluyendo las variables A_1, A_2, A_3, A_4, A_5, A_6 A_7 y A_{11} que eran con las que no se rechaza la hipótesis nula de que su coeficiente sea 0. Como para A_{10} no se rechaza la prueba con una diferencia mínima al nivel de significancia $\alpha = 5\,\%$ que se eligio para tomar la decisión, no sera excluída.

En las pruebas de significancia se obtuvieron los siguientes resultados:

2° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)	AIC
β_0	-2.50	0.5676341	-4.403	$1.07\text{x}10^{-5}$	
A_8	3.60	0.3253954	11.056	$< 2\text{x}10^{-16}$	
A_9	0.77	0.3846553	2.009	0.04449	
A_10	0.11	0.0624668	1.776	0.07575	363.34
$A_{12.1}$	-0.35	0.5322954	-0.655	0.51268	
$A_{12.2}$	3.11	1.0617938	2.933	0.00336	
A_13	-0.0009	0.0007925	-1.173	0.24071	

Cuadro 3.8: Pruebas de significancia del 2° Modelo

En el cuadro 3.8 se muestran las pruebas de significancia que se obtuvieron, de la que podemos apreciar que las variables A_{10}, A_{12} y A_{13} no son significativas para el modelo bajo el nivel de significancia propuesto al momento de realizar la distinción de los grupos.

Por otro lado el AIC que presenta es mayor al del modelo anterior, debido a esto podemos decir que este modelo no es "mejor", sin embargo, se observaron las clasificaciones de este modelo y se obtuvieron los mismos resultados que en el modelo anterior. Recordemos que buscamos un modelo que mejore la clasificación en cuanto a no conceder tarjetas de crédito a quienes presenten un mal historial crediticio, ya que en esta decisión tiene mayor error. Debido a esto realizaremos un modelo adicional para revisar si hay una mejora.

Se utilizó una vez más la rutina ahora excluyendo las variables A_{12} y A_{13} que no es significativa, se obtuvieron los siguientes resultados.

3 ° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P(> |z|)	AIC
β_0	-2.7564	0.2697	-10.221	¡2e-16	
A_8	3.3933	0.3029	11.202	¡2e-16	370.07
A_9	0.6166	0.3685	1.673	0.0943	
A_{10}	0.1259	0.0622	2.024	0.0430	

Cuadro 3.9: Pruebas para el 3 ° Modelo

Ahora en este modelo del cuadro 3.9 observamos que se rechaza la hipótesis nula $H_0 : \beta_i = 0$ para casi todas las variables excepto para la variable A_9, quizá esta variable esta provocando error al momento de hacer la clasificación. En cuanto al AIC= 370.07 se incrementó a comparación del modelo anterior.

Sin embargo, se hizo la clasificación para este modelo y se mejoró en dos observaciones, ahora solo presenta 30 errores de la muestra de 207 solicitantes, es decir, presenta un error del 14 % , de los cuales 6 pertecen a la clase "1" y los 24 restantes a la clase "0". Como podemos ver la clasificación que mejora es en otorgar un crédito a personas que posiblemente le darán una buena rentabilidad a la entidad financiera. Pero con una mayor probabilidad sigue concediendo tarjetas de crédito a personas que no le darán rentabilidad a la entidad si no que le ocasionaran una pérdida. Es "mejor" nuestro modelo por que mejora la clasificación en un 1.45 % aun que el AIC aumenta, pero no significativamente.

Como ya observamos en el cuadro 3.9 la variable A_9 no es significativa para un $\alpha = 5\,\%$, una vez más haremos un nuevo modelo ahora excluyendo la variables A_9.

Los resultados son los siguientes:

4 ° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P(> |z|)	AIC
β_0	-2.646	0.25753	-10.274	¡ $2\text{x}10^{-16}$	
A_8	3.428	0.30257	11.328	$< 2\text{x}10^{-16}$	370.86
A_{10}	0.197	0.05077	3.888	0.000101	

Cuadro 3.10: Pruebas obtenidas para el 4 ° modelo

Como podemos ver en el cuadro 3.10 ya se rechaza la hipótes $H_0 : \beta_i = 0$ para todas las variables. En este aspecto el modelo ya está mejor. Mientras que el AIC se incrementó a comparación del modelo anterior e igual que el primer modelo. Sin embargo, la diferencia del AIC con respecto al modelo anterior es muy pequeña, quizá poco significativa.

Se hicieron de nuevo las clasificaciones y se obtuvieron los mismos resultados de clasificación que el modelo anterior. Por lo que sí es un mejor modelo a comparación del primer modelo.

Por lo tanto el modelo que consideramos como un "buen" modelo es el siguiente:

$$log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = -2.646 + 3.428A_8 + 0.197A_{10}$$

Por lo tanto la probabilidad deseada es:

$$\begin{aligned}\pi(x) &= \frac{exp(-2.646 + 3.428A_8 + 0.197A_{10})}{1 + exp(-2.646 + 3.428A_8 + 0.197A_{10})} \\ &= \frac{0.07095606 * 30.80390494^{A_8} * 1.21820424^{A_{10}}}{1 + 0.07095606 * 30.80390494^{A_8} * 1.21820424^{A_{10}}}\end{aligned}$$

Concluimos que este es el "mejor" modelo debido a que era el modelo con todas las variables significativas, a pesar de que el AIC no disminuyó a comparación del modelo anterior. Además de que este modelo conserva la información necesaria para hacer una "mejor" clasificación a comparación del primer modelo.

3.5.2. Clasificación con todos los datos

Aquí haremos el modelo con todos los datos y así mismo la clasificación. Es decir, clasificaremos los 690 datos que tenemos de nuestra base de datos esperando que como ya sabemos clasifique 383 que pertenecen a la clase"0" y los 307 a la clase "1". Para poder lograr esto intentaremos buscar un buen modelo que nos dé la mejor clasificación.

Como ya hemos estado mencionando haremos uso de rutinas de **R**. La función obtenida es:

$$\mathbf{X} = \begin{pmatrix} 1 \\ A_1 \\ A_2 \\ \vdots \\ A_{12.2} \\ A_{13} \end{pmatrix} \qquad \beta = \begin{pmatrix} -5.699 \\ 0.163 \\ 0.008 \\ \vdots \\ 3.426 \\ -0.002 \end{pmatrix}$$

$$
\begin{aligned}
log\Big(\frac{\pi(x)}{1-\pi(x)}\Big) = \; & - \; 5.69 - 0.16A_1 + 0.009A_2 - 0.14A_3 + +0.76A_{4.1} + 18.68A_{4.2} \\
& + \; 3.93A_{5.1} + 3.24A_{5.2} + 3.20A_{5.3} - 0.93A_{5.4} + 3.41A_{5.5} \\
& + \; 3.92A_{5.6} + 3.88A_{5.7} + 4.55A_{5.8} + 5.45A_{5.9} + 3.97A_{5.10} \\
& + \; 4.47A_{5.11} + 5.05A_{5.12} + 6.21A_{5.13} - 2.92A_{6.1} + 2.43A_{6.2} \\
& - \; 1.86A_{6.3} - 2.21A_{6.4} - 0.01A_{6.5} - 1.60A_{6.6} - 4.15A_{6.7} \\
& + \; 0.05A_7 + 3.80A_8 + 0.71A_9 + 0.13A_{10} - 0.33A_{11} - 0.22A_{12.1} \\
& + \; 3.42A_{12.2} - 0.002A_{13}
\end{aligned}
$$

Sin embargo la probabilidad buscada es la que se obtiene de la siguiente expresión:

$$\pi(x) = \frac{e^{(\beta^T X)}}{1 + e^{(\beta^T X)}},$$

A continuación presentaremos unos gráficos que nos muestran un poco acerca del comportamiento de nuestro modelo.

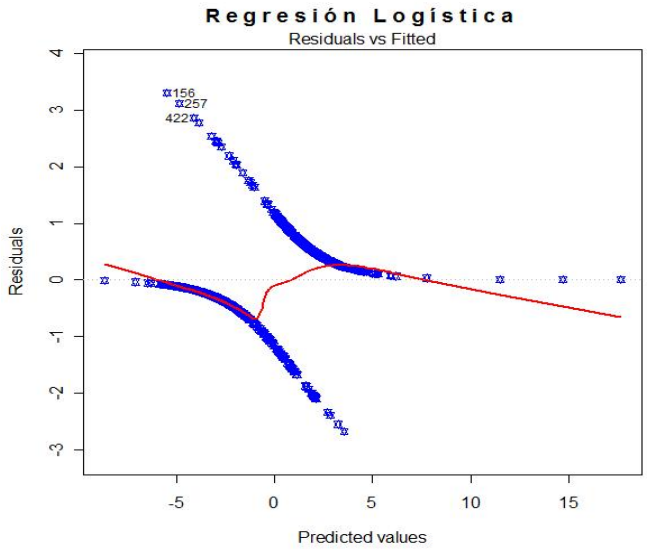

Figura 3.20: Residuos del modelo

En el gráfico de la figura 3.20 se observan los residuos del modelo de regresión logística. No hay que olvidar que el modelo involucra todos los datos, observemos que los residuos se salen de las bandas pero más de la banda superior porque se observan unos puntos que se alejan de los demás, pudiera ser que sean datos atípicos. Pero es poco lo que se salen, mientras que de la banda inferior se comporta mejor.

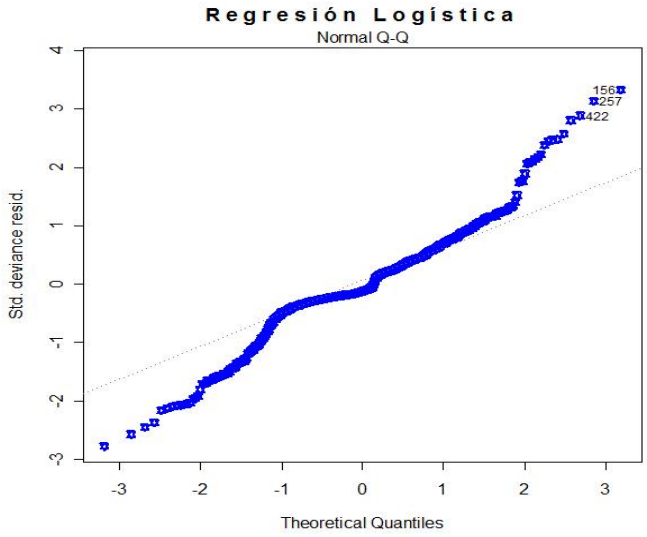

Figura 3.21: QQ-PLOT

El gráfico de la figura 3.21 muestra los cuantiles teóricos del modelo, en el que se aprecia que el modelo tiene colas muy pesadas ya que ambos extremos se alejan considerablemete de la línea, mientras que de la parte central se ajusta bien a la línea de 45°. Nos hubiera gustado que nuestros datos se ajustaran a esta línea ya que eso nos indicaría que sería un buen modelo, pero como los extremos no se ajustan dudamos que sea un buen modelo.

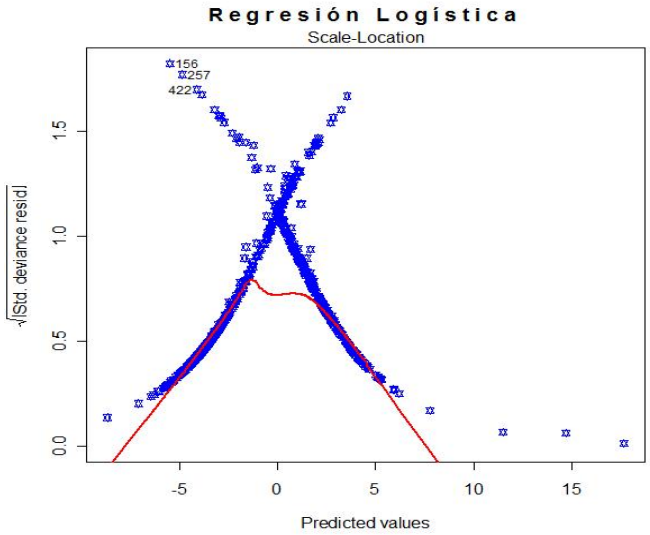

Figura 3.22: Devianza

El gráfico de la figura 3.22 nos muestra los predictores del modelo contra la raíz cuadra-

da de la devianza de los residuos, lo que se observa es que del lado derecho decrece más que la banda y esto por que se observan unos datos que no se agrupan con todos los demás, pareciera que hay presencia de datos atípicos dentro del modelo.

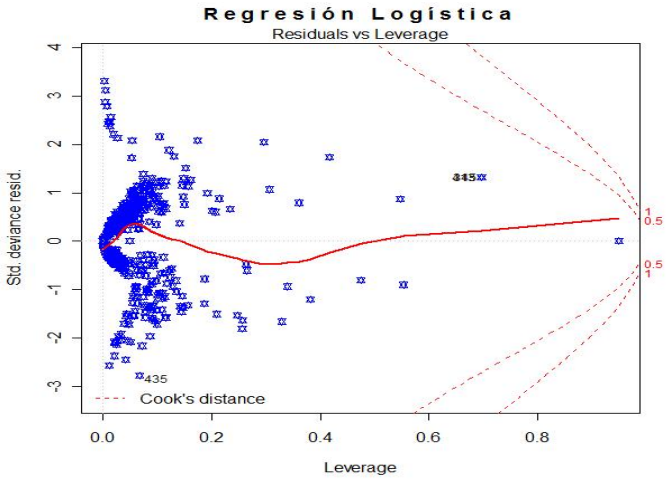

Figura 3.23: Palanca

En el gráfico de la figura 3.23 se observa que alrededor del cero los puntos parecen unirse demasiado, sin embargo presenta puntos dispersos que provocan que sí se pueda distinguir la presencia de los dos grupos, pero la gran mayoría indican la existencia de un sólo grupo.

Como lo hemos visto en los modelos anteriores ahora presentaremos el cuadro 3.11 que nos muestra las pruebas de significancia para observar si nuestro modelo es adecuado o si hay qué hacer pruebas adicionales para mejorarlo.

El cuadro 3.11 muestra los coeficientes del modelo, el error estandar así como la prueba $H_0 : \beta_i = 0$. Así mismo la última columna nos da el p-valor de cada coeficiente los cuales si los comparamos con un $\alpha = 5\,\%$ de confiabilidad nos indica que las variables A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{11}, A_{12} y A_{13} no son significativas, mientras que las restantes A_8, A_{10} y el β_0(intercepto) son significantes.

Por otro lado presenta un índice de AIC = 478.65 grande a comparación de los que nos habían estado apareciendo pero esto se debe a que estamos utilizando más datos. Pero observemos que para las variables A_9 y A_{12} no se rechaza la prueba con una diferencia mínima a nuestro valor de significancia.

Ahora observemos la clasificación que realizó con todos los datos:

	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)
β_0	-5.699	1.044	-5.459	$4.8\text{x}10^{-8}$
A_1	-0.163	0.2.954	-0.552	0.580627
A_2	0.008	0.01264	0.684	0.494065
A_3	-0.015	0.02915	-0.513	0.607705
$A_{4.1}$	0.759	0.3199	2.375	0.017572
$A_{4.2}$	18.69	911.6	0.020	0.983646
$A_{5.1}$	3.614	2.179	1.659	0.097211
$A_{5.2}$	3.237	2.124	1.524	0.127560
$A_{5.3}$	3.210	2.025	1.585	0.112967
$A_{5.4}$	-0.926	2.111	-0.439	0.660825
$A_{5.5}$	3.407	2.088	1.632	0.102666
$A_{5.6}$	3.926	2.127	1.845	0.064975
$A_{5.7}$	3.885	2.066	1.881	0.060029
$A_{5.8}$	4.550	2.098	2.169	0.030082
$A_{5.9}$	5.449	2.302	2.367	0.017922
$A_{5.10}$	3.966	2.078	1.909	0.056265
$A_{5.11}$	4.477	2.889	1.549	0.121300
$A_{5.12}$	5.054	2.153	2.347	0.018903
$A_{5.13}$	6.215	2.203	2.821	0.004794
$A_{6.1}$	-2.916	2.659	-1.097	0.272795
$A_{6.2}$	2.436	2.119	1.150	0.250158
$A_{6.3}$	-1.867	1.941	-0.962	0.336041
$A_{6.4}$	-2.214	2.000	-1.107	0.268280
$A_{6.5}$	-1.179	2.288	-0.005	0.995890
$A_{6.6}$	-1.599	1.961	-0.816	0.414766
$A_{6.7}$	-4.150	2.558	-1.623	0.104653
A_7	0.046	0.048	0.948	0.343325
A_8	3.798	0.342	11.088	$< 2\text{x}10^{-16}$
A_9	0.706	0.365	1.932	0.053307
A_{10}	0.138	0.057	2.401	0.016329
A_{11}	-0.3313	0.027	-1.201	0.229838
$A_{12.1}$	0.2160	0.486	0.444	0.657122
$A_{12.2}$	3.426	0.971	3.528	0.000419
A_{13}	-0.002	0.956	-2.598	0.009363

Cuadro 3.11: Pruebas de significancia utilizando todos los datos y variables

0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 1 1 0
0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
0 1 0 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1
0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 0 1 1 0 1 1 1 1 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0

Variables	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
TODAS	26	52	8.46	13.57	11.30

Cuadro 3.12: Errores de clasificaión

La clasificación que observamos en el cuadro 3.12 presenta errores, pero realmente son pocos ya que de los 690 datos de la muestra total tiene 78 errores. Entonces, de los 383 que pertenecian a la clase "0" se les nego a 331 y a los 52 restantes se les otorgó el crédito cuando no se les debía conceder, mientras que de 307 que pertenecen a la clase "1" a 281 individuos se les concedió la tarjeta de crédito solicitada y a los 26 restantes se les negó dicha tarjeta.

Observemos que la clase "0" presenta más errores que la clase "1", es decir, con una mayor probabilidad concederemos una tarjeta de crédito a una persona con un historial crediticio malo. Lo cual provoca una pérdida financiera para la entidad. Por otro lado estamos negando créditos a solicitantes que presentan un buen historial de crédito y que podrían darle rentabilidad a la entidad.

Hicimos algunas pruebas en las que se observó que el modelo tenía dificultades, así mismo se observaron las clasificaciones y estas nos indicaron que el modelo presenta errores, por esta razón el modelo anterior no es un modelo adecuado para realizar la clasificación. Sin embargo, podemos revisar otro modelo para verificar si mejoran los resultados.

Se hizo un nuevo modelo considerando que en las pruebas de significancia del modelo anterior, excluyendo las variables A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_{11}, A_{12} y A_{13}.

A continuación presentaremos el siguiente modelo:

2° Modelo de Regresión Logística					
	Coeficiente	Error Estandar	Prueba z	P($>\|z\|$)	AIC
β_0	-2.88768	0.23648	-12.211	$< 2\text{x}10^{-16}$	
A_8	3.43234	0.25941	13.231	$< 2\text{x}10^{-16}$	506.33
A_9	0.68558	0.30980	2.213	0.02690	
A_{10}	0.13833	0.05117	2.704	0.00686	

Cuadro 3.13: Pruebas de significancia 2° modelo

En el cuadro 3.13 se pueden apreciar las pruebas de significancia así como el índice de Akaike (AIC). Se excluyeron las variables indicadas y se obtuvo que todas las variables que se incluyeron en este modelo son significativas. Es decir estas variables nos proporcionan información al modelo para poder realizar nuestra clasificación. También se obtuvo el índice de Akaike AIC = 506.33 el cual es mayor que el del modelo anterior. En principio podríamos decir que este no es un mejor modelo que el anterior y que es necesario tener más variables para poder tener más información.

Sin embargo, las clasificaciones en lugar de mejorar, empeoraron. Observemos el cuadro 3.14 que contiene los resultados obtenidos:

Variables	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
β_0 A_8 A_9 A_{10}	23	77	7.5	20.10	14.5

Cuadro 3.14: Errores de clasificaión en el 2° modelo

En la clasificación del modelo anterior que se observa en el cuadro 3.14 se obtuvieron

sólo 52 errores en la clase "0" y en el nuevo modelo se obtuvieron 77 errores, causa de pérdida para la entidad financiera. Mientras que en el modelo anterior tuvo 26 errores en la clase "1" y en el actual tiene 23, lo que causa que se le niegue crédito a más personas que sí le podran dar rentabilidad a la entidad financiera. Lo anterior provoca una pérdida mayor comparada a otorgarle tarjetas de crédito a solicitantes con un historial crediticio malo.

En el primer modelo se obtuvo un 11.3 % de error lo que significa que fue mejor la clasificación un 3.2 % aproximadamente en el primer modelo ya que en el segundo modelo se tuvo un 14.6 % de error en clasificación. Lo que nos conduce a adoptar como "buen" modelo, en términos de su clasificación, el modelo que obtuvimos primero.

En todo el análisis que se realizó de regresión logística fue claro ver que el modelo comete mayor error cuando decide que debe otorgar tarjetas de crédito a personas que no le daran rentabilidad a la entidad. Es cierto que en algunos casos es mejor otorgarle tarjetas de crédito a personas que son poco riesgosas ya que dará más rentabilidad una persona que es poco riesgosa ya que si se financia con la tarjeta a diferencia de una persona que nunca se financie con la tarjeta y que a consecuencia no le dará rentabilidad a la entidad. Pero tampoco es recomendable poner mucho capital en riesgo, ya que se corre el riesgo de llegar a un desfalco financiero muy desfavorable. Por esta razón sería mejor usar este método en combinación con otro, por ejemplo, con análisis discriminante y revisar a detalle a los solicitantes que alguno de los métodos decidan no otorgarle el crédito para que sean sometidos a otro análisis. A pesar de que recordemos que estos métodos no funcionan de manera automática si no solo arrojan sugerencias para la toma de decisiones.

Capítulo 4

Métodos no paramétricos de clasificación en "credit scoring"

En los capítulos anteriores hemos estudiado dos técnicas multivariadas: análisis discriminante y regresión logística, sin embargo observamos que la clasificación de nuestros individuos realmente no es muy buena. Con una probabilidad considerable podemos equivocarnos al momento de tomar la decisión de otorgar o negar un crédito, razón que nos lleva a recurrir a otros métodos propuestos en la literatura de aprendizaje automatizado "machine learning": "K – nearest-neighbor classifiers" y "Learning Vector Quantization". Basandonos en [Hastie, 2008] ya que este autor menciona que estos métodos funcionan bien para los casos de relativa o baja dimensionalidad con decisión complicada. Estos métodos hacen uso de prototipos[1] donde cada uno de estos tiene una etiqueta de la clase asociada y la clasificación de un punto x se realiza con la busqueda en un conjunto de prototipos con los K prototipos más próximos. Donde "Cercano" se define generalmente por la distancia euclídeana en el espacio de características, después de que cada variable se haya estandarizado. Es decir, necesitamos restarle su media a cada variable y dividirla por su desviación estandar de tal manera que logremos obtener una media aproximada a 0 y varianza aproximadamente igual a 1 en cada una de las variables. La distancia euclidiana es apropiada para las características cuantitativas.

Estos métodos pueden ser muy eficaces si los prototipos están bien colocados para conocer la distribución de cada clase. Los límites irregulares de la clase pueden ser representados con prototipos suficientes en los lugares correctos en el espacio de características. El principal desafío consiste en averiguar cuántos prototipos hay que utilizar y dónde se tienen que colocar; ya que es importante lograr una separación de los grupos para no adquirir costos de mala clasificación, debido a que si clasificamos a un individuo como rentable y no se financia con la tarjeta de crédito a consecuencia no dará ganancia a la entidad financiera, sin embargo, será más costoso otorgar un crédito a una persona

[1]Los métodos prototipo representan los datos de entrenamiento por un conjunto de puntos en función de un espacio de prototipos.

que no regrese el crédito concedido por la entidad financiera y a consecuencia traera pérdidas financieras. Estos métodos difieren según el número y la forma en que los prototipos sean seleccionados.

4.1. Clasificadores de K-vecinos más cercanos (KNN)

El método KNN (K nearest neighbors Fix y Hodges, 1951) es un método de clasificación supervisada[2] que sirve para estimar la función de densidad $F(x/C_j)$ de las variables predictoras x_i para cada clase C_j. Se aplicó en "*Credit Scoring*" por primera vez por Chatterjee y Barcun (1970) y más tarde por Henley y Hand (1996). La idea es elegir una métrica en el espacio de datos de las características. Después, con una muestra de los solicitantes que se consideran como una representación estandar y posteriormente un nuevo solicitante es clasificado como bueno o malo en función de la proporción de prototipos de los K vecinos más cercanos.

Este es un método de clasificación no paramétrico que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C_j, a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

KNN es un algoritmo algo diferente de los anteriores en el sentido de que los propios datos proporcionan el "modelo". El costo de aprendizaje es 0, todo el costo es para el cálculo de la predicción. Para predecir un nuevo registro se encuentran los vecinos más próximos al calcular la distancia euclidiana y luego realizando una media ponderada o mayoría de votos para obtener la predicción final.

En general se usa la distancia Euclidiana la cual esta dada por:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

Funciona bien para los casos de relativa o baja dimensionalidad con decisión complicada.

Los empates en las distancias pueden ocurrir con los datos de precisión finita (o si la distribución subyacente es discreta). Una solución es incluir a todos los prototipos lo más cerca o más cerca que el K vecino más cercano, que no tengan una mayoría de votos entre ellos. Por lo que los empates se rompen de forma aleatoria.

[2]El aprendizaje se hace de tal manera que a cada conjunto se le indica a que concepto pertenece, es decir, se tienen característica que distinguen unos conjuntos de otros; estimación basada en un conjunto de entrenamiento y prototipos

En el reconocimiento de patrones, el algoritmo KNN es usado como método de clasificación de objetos (elementos) basado en un entrenamiento mediante un elemento cercano en el espacio de los elementos. KNN es un tipo de "Lazy Learning", donde la función se aproxima sólo localmente y todo el cómputo es diferido a la clasificación.

Bajo la idea de "*Credit Scoring*" se necesitan tres parámetros para poder ejecutar este enfoque como son: la métrica, cuantos candidatos constituyen el conjunto de K vecinos más cercanos y que proporción de estos deben de ser buenos para que el solicitante sea clasificado como bueno. En general, si la mayoría de los K vecinos son buenos, el solicitante se clasifica como bueno de lo contrario, se clasificará como malo.

4.1.1. Algoritmo para *K*-vecinos más cercanos

Supongamos que tenemos un conjunto de datos de entrenamiento D formado por muestras de entrenamiento. Entonces el espacio es particionado en regiones por localizaciones y etiquetas de elementos de entrenamiento. Un punto en el espacio es asignado a la clase C_j si esta es la clase más frecuente entre los k elementos de entrenamiento más cercanos. La fase de entrenamiento del algoritmo consiste en almacenar los vectores característicos y las etiquetas de las clases de los elementos de entrenamiento. Calcularemos la distancia euclideana de cada muestra de entrenamiento a todas las demás que tenemos almacenadas en el vector del punto anterior y de las que conocemos la clase a la que corresponden, quedandonos con las K muestras más cercanas y clasificando la nueva muestra de entrenamiento en la clase más frecuente a la que pertenecen los K vecinos obtenidos.

La fase de clasificación es para diseñar el clasificador. Se calcula el porcentaje de clasificación sobre los ejemplos de este conjunto (desconocidos en la tarea de aprendizaje) para conocer la evaluación del elemento.

En general, primero se tiene que estandarizar cada una de las características para que se obtenga media 0 y varianza 1, ya que es posible que las variables hallan sido medidas en unidades diferentes. En el caso de que no se estandarizaran nos podrían provocar un error de clasificación, debido a que el método es sensible a escalas.

Un método simple es elegir K constante a lo largo de los nuevos solicitantes más cercanos de K y cero en otro lugar. Esto en realidad no define una densidad, pero sugiere una estimación de la distribución *a posteriori* como las proporciones de la clase más cercana entre los K datos. Esta es una función a trozos constante en el espacio y da un clasificador que se conoce como la regla del K vecinos más próximo. La densidad del vecino para cada clase, es como elegir la vecindad de los K puntos más próximos de cualquier clase. Si las probabilidades *a priori* son conocidas y las proporciones de las clases en el conjunto de entrenamiento no son proporcionales, las proporciones entre los

vecinos deben ser ponderadas.

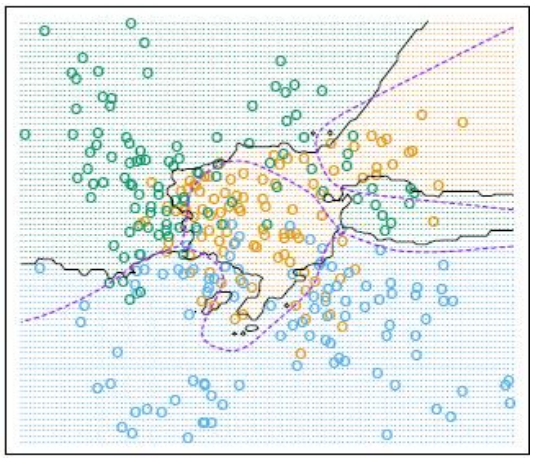

Figura 4.1: 15 Vecinos más cercanos

En la figura 4.1 se muestra un ejemplo donde se realizó la clasificación de grupos bajo el método de KNN haciendo uso de 15 prototipos.

Regla KNN La regla de clasificación por vecindad más general es la regla de clasificación de los K vecinos más cercanos o simplemente KNN. Se basa en la suposición de que los prototipos más cercanos tienen la misma probabilidad a posteriori.

La versión de este algoritmo con $K = 1$ es a menudo bastante éxitosa. El límite de decisión es bastante suave en comparación con el KNN, donde se utiliza un clasificador (1-NN). Hay una estrecha relación entre el vecino más cercano y métodos de prototipo: en 1-clasificador de un vecino más cercano cada punto es un prototipo. Debido a que utiliza sólo el punto más cercano hasta el punto de consulta, el sesgo de la estimación de 1-vecino más cercano a menudo, es escasa, pero la diferencia es alta.

Un importante resultado de Cover & Hart (1967) muestra que asintóticamente la tasa de error del clasificador 1-vecino más cercano no es más que el doble de la tasa de error de la regla de Bayes. La idea aproximada de la prueba es el siguiente (utilizando el error al cuadrado): suponemos que el punto de consulta coincide con uno de los puntos de entrenamiento, de modo que el sesgo es cero. Esto es cierto asintóticamente si la dimensión del espacio de características es fijo y los datos de entrenamiento llenan el espacio de una manera densa.

Entonces el error de la regla de Bayes es la varianza de una variable aleatoria Bernoulli, mientras que el error de la regla de 1-vecino más cercano es el doble de la varianza de

una variable aleatoria Bernoulli como se indica a continuación:

$$\begin{aligned} \text{Error de la regla de Bayes} &= 1 - p_{k^*}(x) \\ \text{1- vecino más cercano} &= \sum_{k=1}^{K} p_k(x)(1 - p_k(x)) \geq 1 - p_{k^*}(x) \end{aligned}$$

donde:
k^* es la clase dominante y
$p_k(x)$ la probabilidad condicional de la clase k

Así mismo Cover & Hart dieron un resultado sobre el comportamiento en una muestra grande con respecto de la regla del vecino mas cercano. Hay que tener encuenta que la tasa del error esperado es siempre limitada por debajo de E^*, por la optimización de la regla de Bayes. De lo que surge la siguiente proposición.

Proposición 1. *Sea E^* la tasa de error de la regla de Bayes en un problema de K-clases. Entonces la tasa de error de la regla del vecino más cercano, como promedio durante sesiones de entrenamiento, converge en L_1 como el tamaño de los incrementos de formación establecido, a un valor E_1 acotado por arriba por*

$$E^*\left(2 - \frac{K}{K-1}E^*\right)$$

Demostración. Sea X_1 el vecino más cercano a X, el patrón de una muestra aleatoria de la clase C. El argumento de C.J. Stone (1977) y Devroye (1981) muestra que:

$$E|p(k|X_1) - p(k|X)| \longrightarrow 0$$

ahora

$$P(C_1 \neq C|X = x) = \sum_{i \neq j} p(i|x)E[p(j|X_1)|X = x]$$

y así

$$E|P(C_1 \neq C|X) - \sum_{i \neq j} p(i|X)p(j|X)| \to 0$$

por lo tanto $E_1 = Ee_1(X)$ donde $e_1 = \sum_{i \neq j} p(i|x)p(j|x) = 1 - \sum_i p(i|x)^2$

El riesgo condicional de Bayes $r^*(x)$ es decir, $1 - \text{máx}_i\, p(i|x) = 1 - p(k|x)$, así que por la desigualdad de Cauchy-Schwarz.

$$\begin{aligned} (K-1)\sum_{i \neq j} p(i|x)^2 &\geq \left[\sum_{i \neq k} p(i|x)\right]^2 = r^*(x)^2 \\ (K-1)\sum_{i} p(i|x)^2 &\geq r^*(x)^2 + (K-1)(1 - r^*(x)^2) \end{aligned}$$

y

$$1 - \sum_i p(i|x)^2 \leq 2r^* - \frac{K}{K-1} r^*(x)^2$$

Al tomar esperanza obtenemos:

$$E_1 \leq 2E^* - \frac{K}{K-1} E[r^*(X)^2] \leq 2E^* - \frac{K}{K-1}(E^*)^2$$

$\Longrightarrow$

$$E_1 \leq E^* \left(2 - \frac{K}{K-1} E^* \right)$$

□

La regla de clasificación por vecindad más simple es la regla de clasificación del vecino más cercano o simplemente 1-NN. Esta regla se basa en la suposición de que la clase del patrón a etiquetar, x, es la del prototipo más cercano en el conjunto de referencia. Cabría esperar que la regla 2-NN no tendría ninguna mejora con respecto a la regla 1-NN, ya que lograriamos una mayoría de votos o un empate que suponemos se rompe al azar. Se muestra también el siguiente resultado.

Proposición 2. *Supongamos que hay dos clases, y sea E_k la tasa de error asintótico de la regla k-nn con vínculos rotos al azar y $E_k^{'}$ si los vinculos se reportaron como una incertidumbre. Entonces*

$$E_2^{'} \leq E_4^{'} \leq \cdots \leq E_{2k}^{'} \nearrow E^* \swarrow E_{2k} = E_{2k-1} \leq \cdots \leq E_2 = E_1 = 2E_2^{'}$$

Demostración. Se observa en [Ripley, 1996] □

C. J. Stone (1977) demostró que éxiste otra versión asintótica para que el número de prototipos aumente con el tamaño del conjunto de entrenamiento (n) menciona: que cuando la $K \to \infty$ y $K/n \to 0$, la regla KNN converge en probabilidad al riesgo de Bayes. Y por otro lado, si $K/log(n) \to \infty$, se fortalece la convergencia casí segura. Los mismos métodos permiten que la varianza en las series de entrenamiento para K fijo sean estimadas.

Todas estas medidas de desempeño son asintóticas y no se aplican a muestras finitas, por ejemplo la tasa de error no disminuye monótonamente con K impares. Observe que los resultados no dependen de la métrica, mientras que en la práctica la elección de métricas a menudo es muy importante. Hay pocos resultados para muestras finitas.

4.1.2. Elección de K

La mejor elección de K depende fundamentalmente de los datos; generalmente, valores grandes de K reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas.

Si $K_i(x)$ para $i = 1, 2, \ldots, j$ es el número de prototipos de la clase C_i que se encuentran en una vecindad de x, el valor del estimador:

$$\widehat{p}(x|C_i)$$

puede calcularse como la proporción de los N_i prototipos de la clase que se encuentran en esa vecindad:

$$\widehat{p} = \frac{K_i(x)}{N_i}$$

El problema que se plantea es el de si existe un valor óptimo para K o en su defecto si se pueden dar algunas reglas para fijar este valor. Observamos que la elección de K está determinado por la densidad de los datos y debería hacerse en relación a N_i.

Si K toma un valor grande se corre el riesgo de hacer la clasificación de acuerdo a la mayoría (y no al más parecido), y si el valor es pequeño puede haber imprecisión de la clasificación a causa de los pocos datos seleccionados como instancias de comparación. Un buen valor de K puede ser seleccionado mediante una optimización de uso. El caso especial en que la clase es predicha para ser la clase más cercana al nuevo dato de entrenamiento (cuando $K = 1$) es llamado Algoritmo del vecino más cercano (Nearest Neighbor Algorithm) que no es otra cosa más que usar como instancia de comparación al primer vecino más cercano encontrado.

Tambien el valor de K puede hallarse mediante el uso de diagramas de Voronoi. La exactitud de este algoritmo puede ser severamente degradada por la presencia de ruido o características irrelevantes, o si las escalas de las características no son consistentes con lo que uno considera importante.

Diagrama de Voronoi: Son polígonos de Thiessen nombrados en honor al meteorólogo estadounidense Alfred H. Thiessen; son una construcción geométrica que permite construir una partición del plano euclídeano. Estos mismos objetos también fueron estudiados por Georgy Voronoi de donde toma el nombre de diagramas de Voronoi. Para poder definir un diagrama de Voronoi lo primero que nos hace falta es un conjunto de puntos, las regiones de Voronoi de un conjunto de puntos son las regiones que están más cerca a uno de los puntos que a cualquiera de los otros puntos.

Es uno de los métodos de interpolación más simple basado en la distancia euclidiana siendo especialmente apropiada cuando los datos son cualitativos. Se crean al unir puntos entre sí, trazando las mediatrices de los segmentos de unión. En la figura 4.2 se observa un diagrama de Voronoi que se realizó en una esfera con 100 puntos arbitrarios.

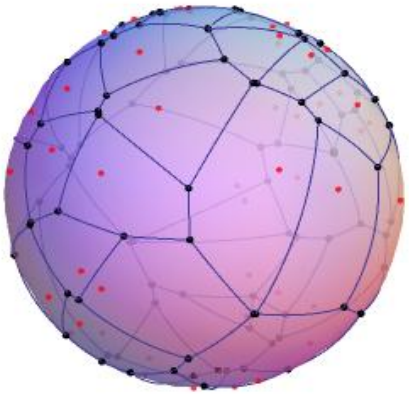

Figura 4.2: Diagrama de Voronoi

Sea cual sea el valor de K la búsqueda requiere explorar todo el conjunto de referencia. Esto significa que el costo de la busqueda depende linealmente del tamaño del conjunto de referencia (N). Ahora debemos considerar el costo del cálculo de cada distancia. El cálculo de la distancia Euclídeana tiene un costo lineal respecto a la dimensionalidad de los datos (d) (costo global: $O(Nd)$) mientras que el costo de una distancia de Mahalanobis es cuadrático respecto a d (costo global: $O(Nd^2)$). Adicionalmente debemos considerar el espacio de almacenamiento requerido.

Como podemos ver no hay un mecanismo para decidir el valor optimo para K ya que siempre dependerá de cada conjunto de datos.

4.1.3. Distancias o Métricas

Los K vecinos más cercanos son seleccionados en base a la métrica. En la práctica, es natural usar la distancia euclídeana, pero después de obtener una escala adecuada de las variables. Tal vez tenga sentido utilizar la distancia de Mahalanobis, si la distribución dentro de la clase es normal y con matriz de covarianza similar. Para dos clases, Short & Fukunaga (1980, 1981) observaron una métrica con el objetivo de minimizar el error cuadrático medio entre el riesgo finito de la muestra y el riesgo asintótico.

Fukunaga y Flick (1984) sugieren la elección de una métrica cuadrática con el mismo objetivo (con dos clases). Su métrica es la distancia de Mahalanobis y así elegir la

covarianza inversa A para obtener:

$$d(x, z) = \sqrt{(x - z)^T A(x - z)}$$

Se calcula el valor de A que aproximadamente minimice la tasa de error y luego se estiman las cantidades correspondientes (la densidad subyacente) para estimar su densidad en KNN.

Uno de los pasos más importantes en la elección de una métrica puede ser la de excluir las características que tienen poca o ninguna relevancia. Las características seleccionadas por el método de componentes principales o por arboles de particinamiento recursivo (arboles de clasificación) puede ser una guía muy útil [Thomas, 2002].

Una idea atractiva es combinar las características de los métodos del kernel y k-vecino más cercano "métodos de distancia ponderada" de las clases de los vecinos para tomar una decisión. Esto ha sido objeto de controversia, ya que asintóticamente para una k fija la "distancia ponderada" no ayudan a una buena elección de la métrica. Sin embargo, esta no es la base correcta para la comparación, ya que la "distancia ponderada" puede permitir que un mayor valor de k pueda utilizarse para un determinado tamaño del conjunto de entrenamiento.

Una característica importante e interesante de KNN es que el método puede cambiar radicalmente sus resultados de clasificación sin modificar su estructura, solamente cambiando la métrica utilizada para hallar distancia. La gran ventaja de poder variar métricas es que para obtener diferentes resultados el algoritmo general del método no cambia, únicamente el procedimiento de medida de distancias.

A continuación sólo mostraremos los modelos matemáticos generales de cada métrica, sin detalles, ya que no es nuestro propósito profundizar sobre este tema; simplemente deseamos recordar que adémas de la distancia euclidiana existen otras métricas alternativas.

- Distancia euclídeana

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ir} - x_{jr})^2}$$

- Distancias de Chebychev

$$d(x_i, x_j) = \max_{r=1,\ldots,n} |x_{ir} - x_{jr}|$$

- Distancia promedio

$$d(x_i, x_j) = \frac{1}{p} \sum_{r=1}^{p} (x_{ir} - x_{jr})^2$$

- Distancia euclidiana cuadrada

$$d(x_i, x_j) = \sum_{r=1}^{p} (x_{ir} - x_{jr})^2$$

- Distancia de Minkowski ($m \geq 2$)

$$d(x_i, x_j) = \left(\sum_{r=1}^{p} |x_{ir} - x_{jr}|^m \right)^{1/m}$$

- Distancia de Manhattan

$$d(x_i, x_j) = \sum_{r=1}^{p} |x_{ir} - x_{jr}|$$

- Distancia de Mahalanobis

$$d^2(x, y) = (x - y)^T S^{-1} (x - y)$$

4.1.4. Ejemplo

Utilizaremos la base de iris que se puede localizar en **R** la cual contiene 3 distintas especies de flores: virginica(v), versicolor(c) y setosa (s). Iris es una base que cuenta con 150 datos de los cuales 50 pertenecen a cada una de las distintas especies y 5 variables las cuales son:

- Longitud del Sepalo.
- Ancho del Sepalo.
- Longitud del Petalo.
- Ancho del Petalo.
- Especies.

Presentaremos la figura 4.3 donde se pueden observar las diferentes especies de la base iris.

En la Figura 4.3 se muestra la existencia de las 3 especies, es claro apreciar que no se observan diferencias entre la especie virginica y versicolor ya que aparentan una similitud debido a qué no se aprecian totalmente separadas, a simple vista se observa solo la existencia de dos grupos no de tres como se esperaba. Ahora observemos que puede hacer el método de K-vecinos más cercanos principalmente con estas especies.

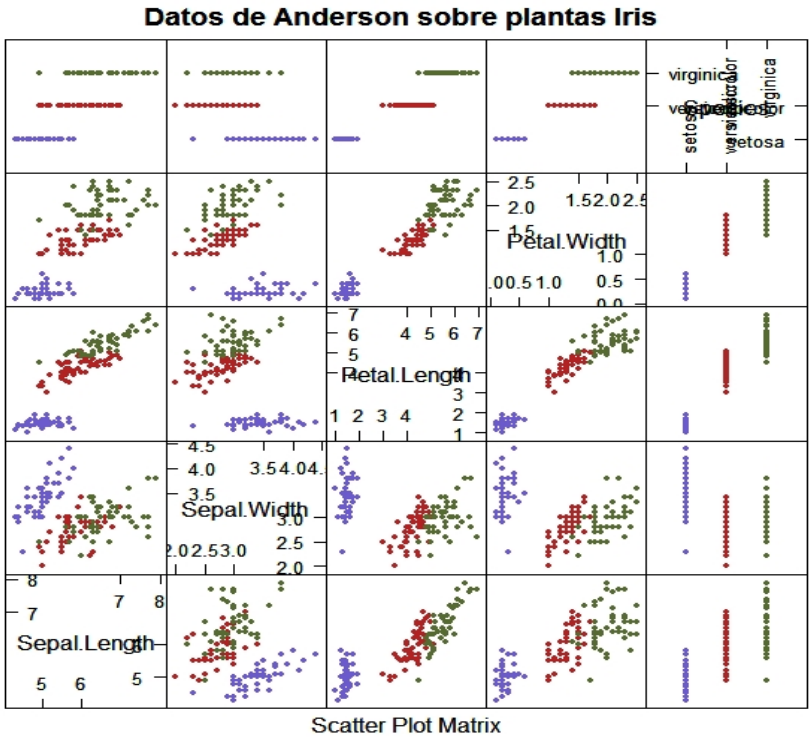

Figura 4.3: Especies de iris

Se tomaron aleatoriamente 25 datos de cada especie para hacer la muestra de entrenamiento es decir 75 datos en total, para después clasificar los 75 restantes y verificar qué tan bueno es el método no paramétrico del K-vecino más cercano.

Utilizaremos la rutina de **R**, llamada "knn" ubicada en la libreria "class"; no sera necesario estandarizar la base de datos ya que todas la variables fueron medidas bajo la misma escala, además de que utilizaremos K=3.

A continuación se muestra la clasificación obtenida por el método:

Clasificación:
s s s s s s s s s s s s s s s s s s s s s s s s s c c v c c c c c v c c c c c c c c c c c c c c c **v c c v v v v v v v v v v c v v v v v v v v v v v**

Como podemos observar existen errores en la clasificación y justo en las especies que se notaban menos separadas en el gráfico, ya que de 25 observaciones de la especie de versicolor (c) se equivoca e indica que dos de ellas son de la especie de virginica (v) mientras que de las 25 observaciones virginicas (v) indica que 3 pertenecen a la especie de versicolor (c) y en la especie setosa (s) no presenta errores de clasificación.

Especie	versicolor (c)	setosa (s)	virginica (v)
versicolor (c)	23	0	2
setosa (s)	0	25	0
virgnica (v)	3	0	22

Cuadro 4.1: Clasificación mediante KNN

Como se puede apreciar en el cuadro 4.1 tiene errores la clasificación, sin embargo, son pocos lo que nos lleva a creer que el método de KNN es un buen clasificador para estos datos.

4.2. "Learning Vector Quantization"

El enfoque adoptado en el aprendizaje de cuantización vectorial (LVQ) es construir un conjunto de formación modificada interactivamente. Siguiendo a Kohonen(1989), se utiliza la formación modificada para establecer un "codebook" que son valores usados para almacenaje o transmisión, estos son retrasladados a valores cercanos de los datos originales. Este procedimiento trata de representar los límites de decisión en lugar de las distribuciones de clase. Una vez más la métrica en el espacio es importante, así que suponemos que las variables se han reducido de tal manera que la distancia euclídeana sea adecuada (al menos localmente), ya que la métrica se usa para seleccionar la clase más cercana al vector de entrada.

En esta técnica, los prototipos se colocan estratégicamente con respecto a las fronteras de decisión. El aprendizaje por cuantización vectorial (LVQ) es una técnica usada para reducir la dimensionalidad de los datos, supone una extensión del aprendizaje competitivo donde los prototipos están etiquetados y las clases están predefinidas. El objetivo es determinar un conjunto de prototipos que representan lo mejor de cada clase. Ahora, además de considerar la cercanía de un prototipo se puede evaluar la clase de éste e imponer correcciones de acercamiento o alejamiento.

Es una versión bajo la supervisión de la cuantización de vectores que se pueden utilizar cuando se han etiquetado los datos de entrada. Esta técnica de aprendizaje utiliza la información de la clase para reposicionar los vectores de Voronoi [3] ligeramente, a fin de mejor var la calidad de las regiones de clasificación.

Esto es particularmente útil para los problemas de clasificación de patrones. El primer paso es la selección de características, la identificación sin supervisión de un conjunto razonablemente pequeño de características en las que se concentra el contenido de la

[3]Son conjuntos de puntos más cercanos a un representante frente a otro.

información esencial de los datos de entrada. La idea es que los puntos de entrenamiento atraigan los prototipos de la clase correcta y rechazacen otros prototipos. El segundo paso es la clasificación en función de los dominios, asignando a las clases individuales.

Un inconveniente de los métodos de aprendizaje de cuantización vectorial es el hecho de que son definidos por algoritmos, en lugar de optimización de algunos criterios fijos, lo cual hace difícil entender sus propiedades. En las siguientes subsecciones hablaremos del algoritmo para crear el conjunto de iniciación y de los algoritmos de corrección sobre dicho conjunto. En la figura 4.4 se observa un ejemplo de clasificación bajo el método de LVQ. En el cual se posicionarón tres prototipos logrando una separación de los grupos existentes.

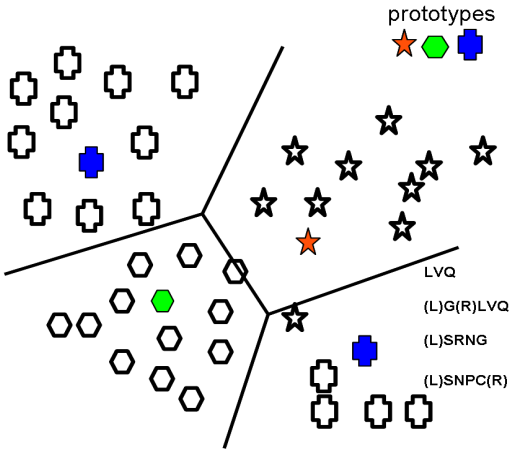

Figura 4.4: Ejemplo de LVQ

A continuación definiremos la notación a considerar en esta sección:

S : Conjunto de entrenamiento.

$S(0)$: Conjunto de iniciación.

$z(t)$: Un prototipo del conjunto S.

N_p : Número inical de prototipos.

N_{pi} : Número de prototipos por clase.

$\alpha(t)$: Valor de desplazamiento.

m_k : Vector de pesos inicial ("codebook").

m_c : Vector de pesos actualizado.

π_i : Funcion de densidad de la proporción de la clase i del conjunto de entrenamiento.

p_i : Función de densidad de la clase i.

$\alpha_c(t)$: Ritmo de aprendizaje en el tiempo t

Cuantización vectorial

Fue originalmente desarrollado por Linde (1980) y Gray (1984) como una herramienta de compresión de datos.

Cuantización vectorial es un método clásico de procesamiento de señales para producir una aproximación a la distribución de una sola clase por un "codebook". Cada señal de entrada se asigna al vector "codebook" más próximo y al vector enviado en vez de la señal original. Una forma de elegir el "codebook" es minimizar alguna medida del error de aproximación de media sobre la distribución de las señales (en la práctica los patrones de entrenamiento de esa categoría):

$$\begin{aligned} m_c &\leftarrow m_c + \alpha(t)[x - m_c] \qquad \text{si } m_c \text{ esta cerca de } x \\ m_i &\leftarrow m_i \qquad \text{para el resto del codebook} \end{aligned} \tag{4.1}$$

Es necesario tener en cuenta que este no es un buen algoritmo de k-medias.

Max (1960) y Zador (1982) han señalado que la elección de la media $r - ésima$ de la distancia como la medida del error de aproximación ascendió a elegir la densidad $(p(x))$ de los vectores de "codebook" para aproximar la densidad de probabilidades de éxito real en la dimensión d. Así, para d grande el código de procedimiento de $k - medias$ es una aproximación a la densidad$(p(x))$. Este es un resultado asintótico para "codebook" grandes, pero sí indica que un "codebook" producido por la cuantización de vectores para cada clase puede ser una buena reducción inicial de un conjunto de entrenamiento muy grande.[Ripley, 1996]

4.2.1. Iniciación

LVQ elige los vectores del "codebook" inicial de entre el conjunto de vectores de entrenamiento. Es deseable que los vectores iniciales se encuentren dentro de los límites de Bayes para su clase, de lo contrario terminan representando "vecindades" en la clasificación inducida por la regla $1 - NN$ basado en el conjunto de entrenamiento que no son más que ruido. Por lo tanto los candidatos para la iniciación se analizan para asegurarse de que están clasificadas por una regla KNN para K moderada

El proceso consiste en la construcción de $S(0)$, o sea, se trata de determinar el valor inicial de los N_p prototipos que constituirán $S(0)$ a partir del conjunto de entrenamiento (S). Sobre este punto debemos aclarar dos cuestiones:

- Aunque conocemos N_p, el número total de prototipos de referencia de $S(0)$, ¿Cual será el número de prototipos por clase, N_{p_i} para $i = 1, 2, \ldots, j$?

  Hay dos alternativas que pueden adoptarse:

  a. Que N_{pi} sea proporcional a N_i. Esto implica que se consideran diferentes probabilidades a *priori*.

  b. Que N_{pi} sea el mismo para todas las clases. Implica considerar que todas las clases tienen iguales probabilidades a *priori*.

- Una vez establecidos los valores de N_{pi}, ¿Cómo se seleccionan los prototipos de $S(0)$?

El procedimiento recomendado es la selección de prototipos con ciertas restricciones que aseguran que éstos se encuentren dentro del agrupamiento correspondiente a su clase. Se puede asegurar que un prototipo está dentro de un agrupamiento si su clasificación mediante la regla *knn*, (con $k = 5$ ó $k = 7$, por ejemplo) es correcta. En otro caso, ese prototipo estará cerca de la frontera de decisión o inmerso en otro agrupamiento. Esta estrategia de selección acelera la convergencia del aprendizaje.

Así, el procedimiento empleado es el siguiente:

- Para cada clase, se procesan secuencialmente sus prototipos.

- Se añaden a $S(0)$ si la clasificación *knn* es correcta.

En cualquier caso, una conclusión importante y que debe tenerse en cuenta es que las clases se caracterizan a partir de un número fijo y relativamente bajo de prototipos.

El valor del parámetro de entrada que corresponde al número de prototipos debe ser suficientemente grande aunque el valor adecuado es desconocido a *priori*. No obstante, el proceso de aprendizaje puede estabilizarse antes por lo que es posible establecer una condición de convergencia adicional, como hemos indicado en el algoritmo. En nuestra aplicación de "*Credit Scoring*", por simplicidad, no utilizaremos ningún criterio de convergencia adicional.

Finalmente, la eficacia conseguida en la clasificación y el tiempo requerido por el aprendizaje dependerán de dos factores:

- El número de prototipos asignados a cada clase y sus valores iniciales.

- El algoritmo de aprendizaje seleccionado y en particular, los valores de los parámetros utilizados.

Otra cuestión es cuántos vectores "codebook'" se deben utilizar, y cómo deben dividirse entre las clases. La literatura parece suponer que mientras el número sea más grande es mejor, ya que esto permitirá una mejor aproximación lineal a trozos a los límites de decisión del clasificador de Bayes. Sin embargo, este argumento supone que los algoritmos iterativos pueden hacer un buen trabajo de distribuir el "codebook".

No está claro cuántos vectores de "codebook" deben ser seleccionados por cada clase, ya que el número necesario depende tanto de lo bien que se emplean como de la proporción de la clase.

4.2.2. Algoritmo de LVQ

Dada una secuencia de datos de entrada, se selecciona un grupo de vectores de referencia m_k ("codebook"). Cada uno de los procedimientos mueve vectores "codebook" para intentar una mejor clasificación del individuo del conjunto de entrenamiento por la regla $1 - nn$ basado en el "codebook". En cada iteración, se selecciona un vector de entrada x_i y se actualiza el vector para ajustar x_i de la mejor manera. El algoritmo LVQ trabaja de la siguiente manera:
A cada clase k se le asocia un vector de peso m_k. En cada iteración el algoritmo selecciona un vector de entrada, x_i, y lo compara con cada vector de pesos, m_k, usando la distancia euclídeana $||x_i - m_k||$; el vector m_c será el ganador si es el más cercano a x_i, por lo que c será la clase asignada:

$$||x_i - m_i|| = \min_k\{||x_i - m_k||\}$$

Las clases compiten entre ellas para encontrar el vector de entrada más parecido, para que el ganador sea el que menor distancia euclídeana tenga respecto al vector de entrada. Sólo la clase ganadora podrá modificar el vector de pesos usando un algoritmo de aprendizaje reforzado, positivo o negativo, dependiendo de que la clasificación sea correcta o no. De este modo, si la clase ganadora pertenece a la misma clase que el vector de entrada (la clasificación ha sido correcta), se incrementará el peso, acercándose ligeramente al vector de entrada (premio). Por el contrario, si la clase ganadora es diferente a la clase del vector de entrada (la clasificación no ha sido correcta), se decrementará el peso, alejándose ligeramente del vector de entrada (castigo).

A continuación estudiaremos con detalle los métodos de corrección LVQ. [Ripley, 1996]

4.2.3. Metodo LVQ1

El procedimiento original LVQ1 utiliza la siguiente regla de actualización. Un vector de entrada x_i se presenta. El próximo vector "codebook", m_c, se actualiza por:

$$\begin{aligned} m_c &\leftarrow m_c + \alpha(t)[x_i - m_c] && \text{Si } x_i \text{ es clasificado correctamente por } m_c \\ m_c &\leftarrow m_c - \alpha(t)[x_i - m_c] && \text{Si } x_i \text{ es clasificado incorrectamente} \end{aligned} \quad (4.2)$$

que se interpretan:

- Premio: Si la clase del prototipo de referencia más cercano, m_c, coincide con la del patrón de aprendizaje, x_i, entonces m_c se acerca a x_i.
- Castigo: En otro caso, m_c se aleja de x_i.

En cualquier caso, la dirección de la corrección está determinada por el vector $x_i - m_c$ y el valor del desplazamiento depende de t.

Los vectores "codebook" no se modifican. Inicialmente $\alpha(t)$ es elegido inferior a 0.1 (0.03 generalmente en LVQ) y se reduce linealmente a cero en un número fijo de iteraciones, de tal manera que este cerca de su propia clase, y lejos de las otras clases. Aquí"Cercanos" puede cubrir regiones muy grandes; como el "codebook" normalmente será pequeño y, en todo caso cubrirá el espacio. Kohonen (1990) da pie a la cuantización vectorial que se aplica a la función $|\pi_1 p_1(x) - \pi_2 p_2(x)|$ para las dos clases (o las dos clases que son más relevantes a nivel local).

El algoritmo LVQ1 tiende a mover los prototipos hacia prototipos de aprendizaje de su misma clase y a alejarlos de los de otra clase. De esta manera, aproxima las funciones de densidad de probabilidad de las clases o recíprocamente, reduce la densidad de los prototipos alrededor de las fronteras de decisión entre clases.

Si se han utilizado todos los prototipos de una clase y no se ha terminado el aprendizaje se vuelve a visitar de nuevo, presentando de forma aleatoria o cíclica los prototipos de esa clase. Es importante destacar que no es tan importante si el conjunto de prototipos inicial es de buena calidad.[Ripley, 1996]

4.2.4. Metodo LVQ1 optimizado (OLVQ1)

Como su nombre indica, se trata de una mejora sobre el aprendizaje LVQ1. La diferencia radica en que cada prototipo del conjunto de referencia tiene su propia razón de aprendizaje.

Una variante, OLVQ1, establece el ritmo de aprendizaje $\alpha_c(t)$ para cada vector de "codebook", con una regla de actualización de los tipos de aprendizaje de

$$\alpha_c = \frac{\alpha_c(t-1)}{1 + (-1)^{I(\text{clasificación es incorrecta})}\alpha_c(t-1)} \tag{4.3}$$

donde el signo del denominador es:

- **Positivo** Si la clase (m_c) = Clase $(z(t))$. Este signo hace que en la expresión 4.3, t decrezca cuando $z(t)$ se clasifique correctamente. En los prototipos situados en los centros de los agrupamientos su aprendizaje decrece rápidamente, en posteriores correcciones no se modifican perceptiblemente.
- **Negativo** Si la clase $(m_c) \neq$ Clase $(z(t))$. Este signo hace que en la expresión 4.3, t crezca cuando $z(t)$ se clasifique incorrectamente. En los prototipos cercanos a las fronteras de apendizaje crece rápidamente, lo que significa que se alejan muy pronto hacia el centro del agrupamiento.

Para evitar un crecimiento ilimitado se establece un máximo (0.3). La experiencia práctica demuestra que la convergencia suele ser rápida y LVQ utiliza tantas iteraciones como vectores de "codebook". En todo momento los vectores "codebook" son una combinación lineal del conjunto de vectores de entrenamiento.

Sea $s(t) = (-1)^{I(\text{clasificación es incorrecta})}$ por lo que podemos rescribir 4.2 como

$$\begin{aligned} m_c(t+1) &= m_c(t) + s(t)\alpha(t)[x(t) - m_c] \\ &= [1 - s(t)\alpha(t)]m_c(t) + s(t)\alpha(t)x(t) \\ &= [1 - s(t)\alpha(t)][1 - s(t-1)\alpha(t-1)]m_c(t-1) \\ &\quad +[1 - s(t)\alpha(t)]s(t-1)\alpha(t-1)x(t-1) + s(t)\alpha(t)x(t) \end{aligned}$$

Supongamos ahora que $x(t-1) = x(t)$ y el mismo vector de "codebook" es el más cercano en ambos momentos (para $s(t-1) = s(t)$). Si le pedimos que el multiplicador de $x(t)$ sea al mismo en ambos términos, encontramos:

$$[1 - s(t)\alpha(t)]\alpha(t-1) = \alpha(t)$$

Esta elección de adaptación de la tasa parece funcionar bien.

En este tipo de aprendizaje puede establecerse la siguiente condición de convergencia adicional: si no hay clasificaciones incorrectas en un número fijo de iteraciones, terminar el aprendizaje.[Ripley, 1996]

4.2.5. Otras versiones de LVQ (LVQ2.1 y LVQ3)

Otras versiones propuestas del aprendizaje LVQ son los llamados métodos LVQ2.1 y LVQ3. La principal diferencia respecto a LVQ1 es que modifican más de un prototipo de referencia en cada paso de aprendizaje. Sin embargo, requieren más parámetros que especificar. A continuación describiremos brevemente la fase de aprendizaje de estos métodos, ya que la inicialización es la usual.

El procedimiento **LVQ2.1** (Kohonen, 1990) esta más aproximada a la regla de Bayes

por los ajustes por pares de los vectores de "codebook". Supongamos que m_s, m_t, son los dos vecinos más cercanos a x_i. Se actualizan de forma simultánea, siempre que el vector (m_s) sea de la misma clase como x_i y la clase del vector (m_t) sea diferente, y x_i cae en una "vecindad" (δ) cerca del punto medio de m_s y m_t. En concreto, debemos tener :

$$\text{mín}\left(\frac{\delta(x,m_s)}{\delta(x,m_t)},\frac{\delta(x,m_t)}{\delta(x,m_s)}\right) > \frac{1-\omega}{1+\omega} \tag{4.4}$$

donde ω es el "ancho" relativo de la vecindad usualmente $\omega \approx 0.25$ podemos interpretar esta condición geométricamente: si x_i es proyectada sobre el vector que une m_s y m_t, debe caer por lo menos (1-ω) / 2 de la distancia de cada extremo.

Dado un patrón de aprendizaje, sí se cumplen las condiciones:

- $m_s(t)$ y $m_t(t)$ son los dos prototipos más cercanos a $z(t)$,
- $z(t)$ está dentro de la vecindad definida por $m_s(t)$ y $m_t(t)$,
- $m_t(t)$ es de su misma clase, y,
- $m_s(t)$ es de diferente clase

si todas estas condiciones se han satisfecho los dos vectores son actualizados por:

$$\begin{aligned} m_s &\leftarrow m_s + \alpha(t)[x - m_s] \qquad \text{premio a } m_s \\ m_t &\leftarrow m_t - \alpha(t)[x - m_t] \qquad \text{castigo a } m_t \end{aligned} \tag{4.5}$$

Esta estrategia de corrección tiene un serio problema: se "desestabiliza" para valores altos con respecto al número de "codebook" ya que aleja los prototipos y no aproxima las funciones de densidad de probabilidad. Por lo tanto se recomienda que LVQ2.1 sólo se utilice para un pequeño número de iteraciones (30 a 200 veces).

La regla de **LVQ3** intenta superar el exceso de corrección mediante el uso de LVQ2.1 esta estrategia de corrección es la siguiente: Dado un prototipo de aprendizaje, $z(t)$, sean $m_s(t)$ y $m_t(t)$ los dos prototipos más cercanos a $z(t)$, si uno de ellos es de la misma clase que $z(t)$ y el otro no, y $z(t)$ está en la vecindad, entonces se aplica LVQ2.1. Si los dos vectores "codebook" son de la misma clase que $z(t)$, se premia a ambos:

$$m_i \leftarrow m_i + \epsilon\alpha(t)[x - m_i] \tag{4.6}$$

para ϵ alrededor de 0.1 a 0.5, para cada uno de los dos vectores más cercano del vector "codebook" si son de la misma clase que x. (La vecindad sólo se utiliza si los dos vectores codebook son de diferentes clases.) Esto introduce un segundo elemento en la iteración, para garantizar que los vectores "codebook" no se conviertan en demasiado o poco representativos de su distribución por clase. El procedimiento recomendado es ejecutar OLVQ1 hasta la convergencia (por lo general rápido) y luego un número moderado de nuevas medidas de LVQ1 (preferiblemente) y / ó LVQ3 sobre el conjunto de prototipos resultante, con un valor de prototipos moderado.[Ripley, 1996]

4.2.6. Ejemplo

Consideremos la base de datos de iris de la paquetería de **R-project** que, como ya mencionamos en la sección anterior contiene 3 distintas especies de flores: virginica,versicolor y setosa. Es una base que cuenta con 150 datos de los cuales 50 pertenecen a cada una de las diferentes especies y 5 variables.

- Longitud del Sepalo.
- Ancho del Sepalo.
- Longitud del Petalo.
- Ancho del Petalo.
- Especies.

Como ya habiamos observado las especies virginica (v) y versicolor (c), muestran mucha similitud entre ellas ya que en la figura 4.3 no se notan la separacion en estas especies.

Una vez más tomamos 75 datos aleatoriamente de 150 totales de los cuales 25 pertenecian a cada una de las distintas especies, con estos 75 haremos la muestra de entrenamiento para después clasificar los 75 restantes verificando que tal funciona el algoritmo de LVQ.

Como en todos los casos anteriores haremos uso de las rutinas de **R** para aplicar los métodos vistos anteriormente como son olvq1, lvq1, lvq2, lvq3 ubicados en la librería *class*, no será necesario estandarizar la base de datos ya que todas las variables fueron medidas bajo la misma escala.

Mediante la función de lvqinit construiremos el conjunto inicial de codebook de la muestra de entrenamiento utilizando 10 prototipos, después utilizando lvqtest, clasificaremos los elementos restantes, el método utiliza la rutina $1 - NN$ haciendo uso del conjunto de "codebook" inical. La clasificación obtenida es la siguiente:

Clasificacion:
s s s s s s s s s s s s s s s s s s s s s s s s s c c v c c c c c v c c c c c c c c c c c c c
c c **v v v v v v v v v v v v v v v v v v v v v v v v v**

En la tabla siguiente se muestran las clasificaciones:

Como podemos observar en el cuadro 4.2 tiene errores de clasificación, en la especie de versicolor, ya que de las 25 especies clasifica 2 como virginicas, en las otras especies no presenta errores, a comparación de KNN es mejor la clasificación resultante.

Especie	versicolor	setosa	virginica
versicolor	23	0	2
setosa	0	25	0
virgnica	0	0	25

Cuadro 4.2: Clasificación, LVQ conjunto de iniciación

Por otro lado es bueno observar cómo funciona el método de olvq1, la literatura sugiere utilizar este método hasta su convergencia, por esta razón primero haremos uso de esta rutina mediante olvq1 y luego clasificaremos mediante la misma rutina anterior. Haremos uso de los "codebook" de iniciación para después mejorar el algoritmo por olvq1 y después hacer la clasificación.

Clasificación:
s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s s c c c c c c c c c v c c c c c c c c c c c c c c c c c c c c c c **v v v v v v v v v c v v v v v v v v v v v v v v v v v v v v v v**

En la tabla siguiente se muestran las clasificaciones:

Especie	versicolor	setosa	virginica
versicolor	24	0	1
setosa	0	25	0
virgnica	1	0	24

Cuadro 4.3: Clasificacion mediante OLVQ1

Como podemos ver en el cuadro 4.3 una vez más presenta errores de clasificación en las mismas dos especies ya que de las 25 versicolor una de ellas la clasifica como virginica, analogamente hace lo mismo para la especie virginica, pero a pesar de esto la clasificación sigue siendo buena.

Verifiquemos si haciendo uso de los métodos lvq1, lvq2.1 y lvq3 la clasificación mejora, se mantiene igual o empeora. Presentaremos las clasificaciones en el orden que se fueron mencionadas. Ahora en lugar de utilizar el conjunto de "codebook" inicial haremos uso del conjunto que se obtuvo mediante el método de olvq1.

En el cuadro 4.4 las primeras 3 columnas representan las clasificaciones mediante lvq1 las siguientes 3 de lvq2.1 y las restantes de lvq3. Como podemos observar una vez más

	lvq1			lvq2.1		
Especie	versicolor	setosa	virginica	versicolor	setosa	virginica
versicolor	24	0	1	24	0	1
setosa	0	25	0	0	25	0
virgnica	3	0	22	1	0	24

	lvq3		
Especie	versicolor	setosa	virginica
versicolor	24	0	1
setosa	0	25	0
virgnica	3	0	22

Cuadro 4.4: Clasificaciones mediante LVQ1, LVQ2.1 y LVQ3

presenta errores de clasificación y principalmente en las especies virginica y versicolor, el método lvq1, 25 de la especie de versicolor solo clasifica bien 24 y 1 como virginica para la especie virginica 22 clasifica bien y 3 los coloca como 3 versicolor. Mediante el método lvq2.1 deja la clasificación como olvq1 y lvq3 clasifica como lvq1. Podemos ver que el método olvq1 tiene una buena convergencia, para hacer una mejor clasificación. ¿Podremos mejorar la clasificación si utilizamos un mayor número de prototipos? Esta pregunta la discutiremos cuando hagamos la aplicación a nuestra base de datos de "*Credit Scoring*". Este ejemplo lo dejaremos hasta aquí ya que solo es para ilustrar el método.

Capítulo 5

Aplicación en "credit scoring" metodos no paramétricos

En el capítulo 3 presentamos la base a estudiar así como un análisis exploratorio de los datos y la aplicación en dos métodos de clasificación supervisada como fueron análisis discriminante y regresión logística. Sin embargo, los resultados de clasificación en estos métodos no fueron favorables, debido a esto se decidió implementar los siguientes métodos con la finalidad de mejorar la clasificación.

Recordemos que la base a explorar esta titulada Aprobación de Crédito de Australia, como su nombre lo dice es una base Australiana que consta de 690 datos, su fuente es confidencial, cuenta con 14 variables de las cuales 6 son de tipo numérico y 8 de tipo categórico. De los 690 datos 307 de ellos pertenecen a la clase "1", es decir, son clientes a los cuales se le otorgó la tarjeta de crédito mientras que los 383 restantes son los clientes que se les nego la tarjeta de crédito.

Los métodos a análizar son los vistos en el capítulo 4, "K-Nearest Neighbour" (KNN) y "Learning Vector Quantización"(LVQ). Recordemos que estos métodos trabajan con variables cuantitativas, por lo tanto haremos uso sólo de las 6 variables de tipo numérico y de la variable categórica que nos indica a que grupo pertenece cada individuo, para corroborar si los métodos no paramétricos estudiados hacen una buena o una mala clasificación. Como ambos métodos son sensibles a escalas, ya que utilizan el concepto de "cercano" mediante la distancia euclidiana, estandarizaremos nuestra base de datos de la siguiente manera.

Para estos métodos se necesita estandarizar característica por característica, de tal manera que cada una tenga media 0 y varianza 1. Es decir, necesitamos restarle su media y dividir entre su desviación estandar a cada variable, para después poder aplicar cada uno de los métodos explicados en el capítulo 4.

A continuación presentaremos un gráfico que muestra cómo se observan las características cuantitativas una vez que ya se estandarizaron.

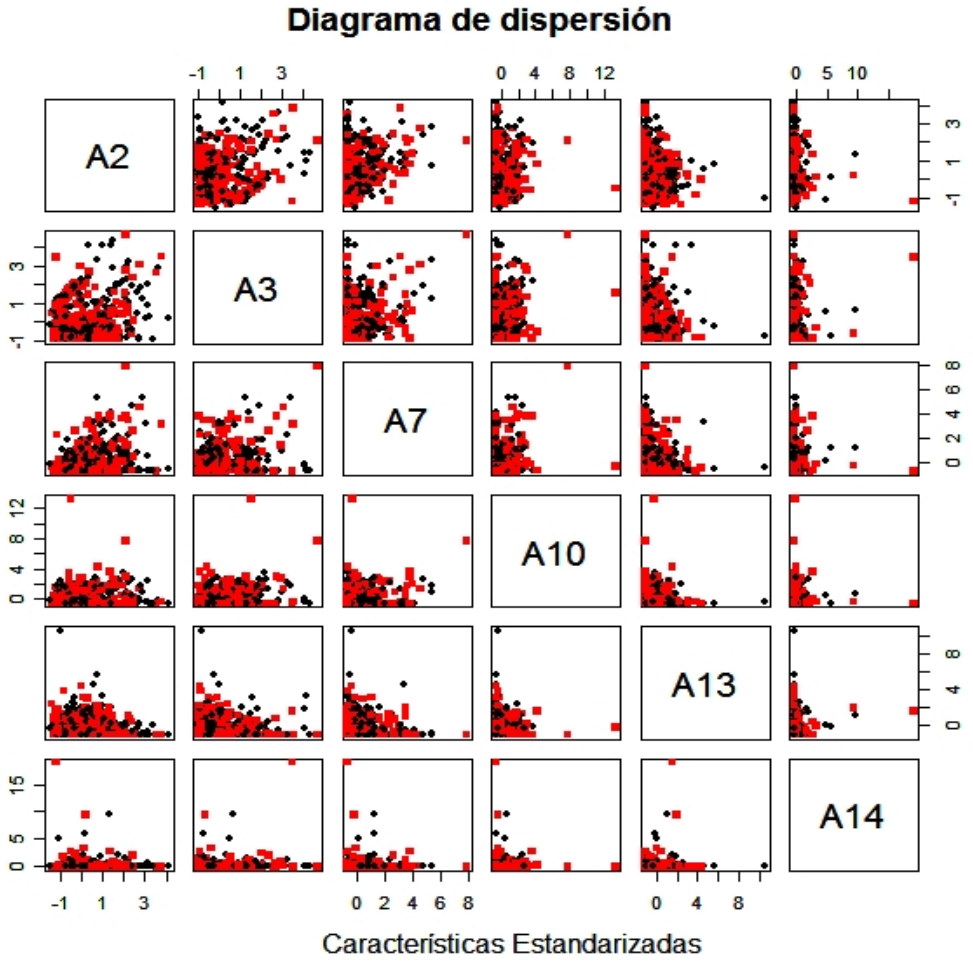

Figura 5.1: Características Estandarizadas

Como podemos ver, no se aprecia una clara separación en dos grupos de las variables presentadas, sin embargo, sí se alcanza a distinguir que existen individuos con características distintas a los demás.

Observemos que las variables no tienen un buen comportamiento, es posible que estas variables no esten midiendo lo que se buscaba, o sea, distinguir fácilmente a un individuo que si se le otorgue la tarjeta de crédito con respecto a otro que no se le otorgue un crédito.

Aplicaremos los métodos estudiados KNN y LVQ en las siguientes secciones.

5.1. K-Vecinos más cercanos

En esta sección haremos uso del método no paramétrico KNN, intentaremos ver si este método realiza una mejor clasificación que los métodos estudiados en el capítulo 3. Como se mencionó necesitamos tener las características que más información nos proporcionen. Anteriormente realizamos un estudio de la base de datos en la que observamos que la variable 14 era una de las variables que más problema nos causaba y aunque las características ya se estandarizaron, no usaremos esa variable, para ver si se obtiene una mejor clasificación.

5.1.1. Aplicación en R

El programa que hemos estado utilizando, tiene entre sus rutinas, un método que realiza este tipo de clasificaciones, la cual se llama "knn" y se encuentra en la libreria "*class*", por lo tanto haremos uso de rutinas de **R − project**.

Consideraremos para la muestra de entrenamiento a 153 individuos que pertenecen a la clase "1" y 192 individuos de la clase "0", aproximadamente el 50 % de cada clase, para después clasificar los 345 individuos restantes y evaluar el desempeño de los métodos. Para comenzar utilizaremos 5 prototipos.

Observemos la clasificación que se obtuvo utilizando 5 variables y 5 prototipos:

n n n n n n n n n n n n s n n s n s n s s n n n s n s n s n n n
n n n n n n s n n s n n n n n n n n n n n n n n n n n s n s n n n
n n s n n n n n n n n n n s n s s n n n n n n n n n n n n n n n
n n n s n n n n n s n n n n s n n n n n n n s n n n s n n s n s s n
n n n n n n n n s n n n s s n n n n n n n n n n n n n n n n s n n
n n n n s n s n n n n s n n n n n n n n n s s n s s s n s n n s n s n s n n
s n s n s s n n s s s s s s s s s n s s s n s s s s s s s n n n n n n s n n s n n s n n
s s s s s s s s n s s n n n s s n s s n s s n s n s s n n s n n n n s s s n n n s s s
s n s s s n s n s n n s s s n s s n n n n n n s s s s s n n n s s n s n s n n s n s s s n n

Como se puede ver, en la clasificación anterior se obtuvieron nuevamente errores, ya que tenía que clasificar los primeros 191 como individuos a los cuales se les deberá negar el crédito, sin embargo sólo a 158 individuos les niega la tarjeta de crédito y a los restantes 33 decide otorgarles la tarjeta de crédito provocando un error de 17.27 %, los siguiente 154 eran individuos a los que sí se le tenía que otorgar el crédito y de estos solo clasifica correctamente a 86 y a los restantes 68 decide negarle la tarjeta de crédito, provocando un error del 44 % un error bastante considerable en esta clase. De esta manera este método arroja un error total del 29 %.

Este método produce mayor error para los individuos a los cuales sí les debe otorgar la tarjeta, pero eso no quiere decir que el error que presenta al tomar la decisión de otorgar créditos a individuos que no se les debe otorgar sea despreciable, este error es el que nos gustaría disminuir principalmente, ya que otorgar una tarjeta de crédito con un monto considerable y que no sea regresado a la entidad financiera sería una causa de pérdidas financieras para ésta. Hay que tener en cuenta que el hecho de otorgar una tarjeta a alguien que no regresara el crédito que se les brinde es más costoso que no otorgarle crédito a individuos que sí lo regresaran, también implica pérdida financiera.

Como vimos esta no es una buena clasificación, se exploró con la disminución o aumento en el número de prototipos para revisar si mejora el desempeño, observemos el cuadro 5.1 que considera clasificaciones con diferentes números de prototipos y los porcentajes de errores.

# Prototipos	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
1	60	47	38.96	24.6	31.01
2	61	46	39.61	24.08	31.01
3	61	38	39.61	19.89	28.69
5	68	33	44	17.27	29
10	71	26	46.10	13.61	28.11
15	77	24	50	12.56	29.27
25	86	16	55.84	8.37	29.56
45	86	16	55.84	8.37	29.56
225	139	1	90.25	0.52	40.57

Cuadro 5.1: Errores de clasificación con variación en el número de prototipos

En el cuadro 5.1 observamos que sí se aumenta el número de prototipos se incrementa el error en la clase "1" y se disminuye en la clase "0", es decir, decide no otorgar créditos a personas que sí son rentables para la entidad financiera lo que nos provoca pérdidas financieras. Los porcentajes de error son considerables ya que ambos le provocan pérdidas financieras a la entidad de crédito. Algo que se puede ver es que mientras más sea el número de prototipos lo que hace es identificar un solo grupo y es por eso que decide no otorgarle créditos a niniguno de los solicitantes.

Si nos fijamos lo que ocurre cuando se le proponen 225 prototipos, se observa que el método alcanza a distinguir un solo grupo ya que de los 345 individuos no le concede el crédito a 329, y sabíamos que no le tenía que conceder el crédito sólo a 191 individuos pero de los 154 individuos que pertenecen a la clase "1" sólo decide otorgarle crédito a 15 individuos y esto con una probabilidad de 0.52. Lo que nos hace pensar que las variables no estan midiendo lo que se proponían ó que alguna de las variables no nos

ayuda a hacer la clasificación. Entonces lo que haremos es volver a aplicar la rutina pero ahora sin la variable 13.

Por lo tanto ahora excluiremos las variables 13 y 14, presentamos la tabla en donde aparece la información que se obtuvo:

# Prototipos	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
1	63	41	40.9	21.46	30.14
2	61	49	39.61	25.65	31.88
3	63	35	40.9	18.32	28.40
5	68	30	44.15	15.70	28.40
10	70	21	45.45	10.99	26.37
15	76	20	49.35	10.47	27.82
25	76	21	49.35	10.99	28.11
45	81	18	52.59	9.42	28.69
225	132	1	85.71	0.52	38.55

Cuadro 5.2: Clasificación excluyendo las variables A_{13} y A_{14}

En el cuadro 5.2 observamos que no se obtuvo una gran mejoría, sólo se disminuyó el porcentaje de error ya que con todas las variables el menor porcentaje total que se obtenía era de 28 % con 10 prototipos y ahora se logró disminuir a 26 % con el mismo número de prototipos. Se disminuyó en un 3 % el error en la clase "0", lo que indica que otorga menos créditos a individuos que no regresaran el crédito otorgado. A pesar de que se mejoró un poco el porcentaje de error sigue siendo considerable.

Excluyendo una variable más, la pérdida de información sería mayor, arrojando resultados menos favorables ya que en lugar de mejorar los errores de clasificación los empeora. Se consideraron combinaciones de variables para ver si se mejoraba la clasificación, quitando las variables 2 y 14 se logro mejorar la clasificación y los datos se muestran en el cuadro 5.3:

En los datos del cuadro 5.3 se mejora en muy poco la clasificación ya que del 26 % de error ahora se tiene un 25 % de error en general. Los datos de la clase "0" tienen un error de 15.5 % pero con solo 5 prototipos mientras que en el análisis anterior se logro un error del 11 %, pero la mejoría se debe a que en la clase "1" se tenía un error del 45 % y ahora se tiene del 38 %.

# Prototipos	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
1	53	49	34.41	25.65	29.56
2	57	51	37.01	26.70	31.30
3	60	36	38.96	18.84	27.82
5	58	30	37.66	15.50	25.50
10	67	27	43.50	14.13	27.24
15	69	24	44.80	12.56	26.95
25	76	20	49.35	10.47	27.82
45	81	15	52.59	7.85	27.82
225	144	1	41.73	0.52	42.02

Cuadro 5.3: Clasificacion excluyendo las variables A_2 y A_{14}

Como lo hemos estado mencionando, en una entidad de crédito la pérdida financiera será mayor si otorga créditos a personas que no devolverán el crédito otorgado y la pérdida será menor si no otorgan créditos a personas que sí son rentables. El problema radica en que no se conocen las probabilidades de incumplimiento de cada cliente, ya que con esta información se podrían tomar mejores decisiones. Entonces en los datos expuestos anteriormente se mejora la clasificación pero la probabilidad de otorgar un crédito a una persona que no devolverá el crédito es mayor respecto a la clasificación que se realizó con 10 prototipos. Lo anterior llevaría a la compañia a tener más pérdidas y esto no es conveniente. Esto no quiere decir que no nos interese mejorar el error que se tiene en la clase "1", sí queremos mejorarlo y sería mejor ya que la compañia tendría mayores ganancias al otorgar créditos a personas rentables, pero como hemos visto en este método mejorar los dos errores conjuntamente es poco probable.

Las clasificaciones obtenidas son poco favorables. Las variables no marcan diferencias que nos permita distinguir los dos grupos. Se presenta un gráfico en donde se aprecia más de cerca dos variables, para ver más claramente que no se distinguen los grupos que buscamos.

Apreciamos en el gráfico 5.2 que no se distinguen los grupos, llegando a la conclusión que las variables no estan midiendo lo que se pretendia medir, que era distinguir dos clases de individuos.

Hemos visto que a pesar de que este método es eficiente para problemas de difícil decisión, la base de datos no mide lo necesario para poder hacer una buena clasificación. Intentamos ver que ocurre cuando se implementa las técnicas de "Learning Vector Quantization", que se presenta en la siguiente sección.

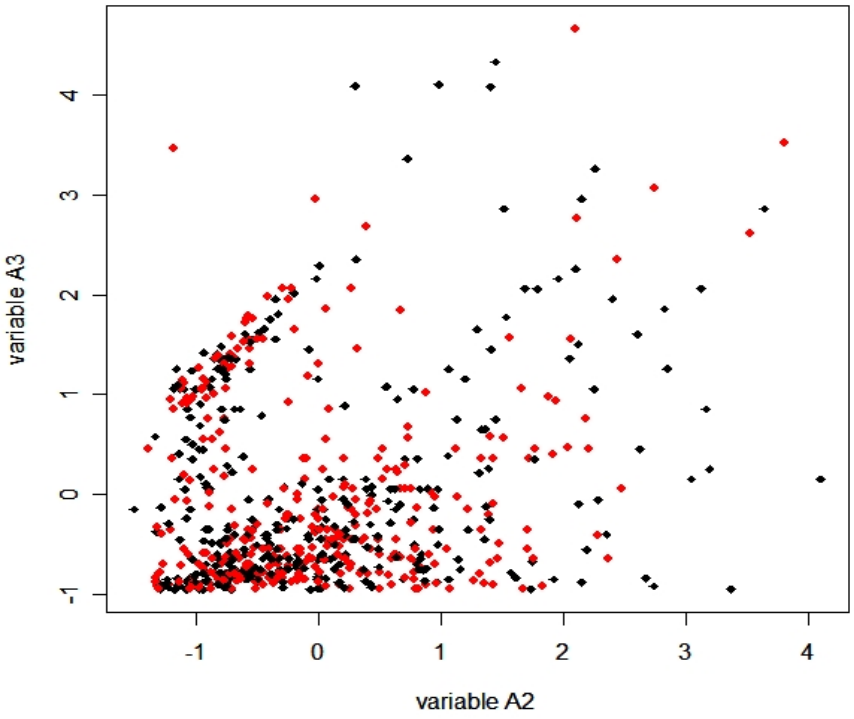

Figura 5.2: Gráfico de A_2 vs A_3

5.2. "Learning Vector Quantization"

Hemos estudiado que esta técnica es para problemas de baja dimensionalidad; nuestra base de datos es una base de solo 690 elementos, pero utilizamos 345 datos para hacer la muestra de entrenamiento y los restantes 345 elementos para hacer la muestra de prueba de la técnica, pero si vemos una base de datos de un banco o de una entidad financiera de crédito no solo tiene esta cantidad de elementos en su base de datos, esta base es una base truncada y de baja dimensionalidad.

Hemos observado que el método de KNN no hizo mucho ya que la base de datos no muestra diferencias entre sus individuos, entre los que sí se les debe otorgar tarjeta de crédito con respecto de los individuos que no se les debería otorgar una tarjeta de crédito.

5.2.1. Aplicación de LVQ: utilizando R

Utilizaremos la muestra anterior para fines de comparación al final de este análisis, ya que nos interesa evaluar si este método no paramétrico es mejor que KNN o vicever-

sa. Algo que comentamos es que la última clasificación mejoraba en muy poquito los errores de clasificación pero incrementaba en casí un 5 % el error en la clase "0" lo que nos llevaría a tener mayores pérdidas en una entidad financiera. Utilizamos los datos que se obtuvieron mediante 10 prototipos y con 4 variables, es decir, sin la variable 13 y 14.

Como el título menciona vamos a hacer uso de rutinas de **R − project**. En este programa se encuentran rutinas que realizan este tipo de técnicas mediante las funciones olvq1, lvq1, lvq2, lvq3 localizadas en la librería "*class*".

Mediante la función lvqinit construiremos el conjunto inicial de codebook para la muestra de entrenamiento que consta de 345 individuos que se compone aproximadamente del 50 % de individuos de cada clase. Primero utilizaremos 15 prototipos, después utilizaremos lvqtest para clasificar los restantes 345 individuos, hay que señalar que el método utiliza la rutina 1-NN para encontrar el conjunto de iniciación de este método. Haciendo uso del conjunto de codebook inical, se obtuvieron los siguientes resultados.

Clasificación incial:

n n n n n n n n s n n n n n n n n n n n s n n n n s n n n n n n
n n n n n n n s s n n n n n n n s n n n n n n n n n n n s n n n n n
n n n n n n n n n n n n n n n n n n s n n n n n n n n n s n n n n n
n n n n s n n n n s s n n n n n n n n n n n n n n n n n n s s n n n
n n n n n n n n n n n n s n n n n n n n n n n n n n n n n n n n n n
n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n s s s n s n n s n s n s n
n s n s n n n n n s n s s s s s s s n s s s n s s s s s s n n s n n s s s n n n n n n n
s n s s n n n s n s s s n n n s n n s s s s n s n s n n n s s n s n n s s s n s s s s s s
n n n n n s n s n s s n s n n n s n n s s s s s n n n n s n s n n s n s n s n n n n s n n n

Los resultados anteriores muestran el conjunto de iniciación, que la clase "0" sólo clasifica bien 176 elementos y a los restantes 15 indica que sí se les debería conceder un crédito provocando así un porcentaje de error del 7.85 %, mientras que en la clase "1" clasifica bien a solo 75 individuos y a los restantes 79 decide no concederles el crédito teniendo un porcentaje de error del 51.29 % y conjuntamente un error del 27.24 %. Observamos que este método se inclina más por no conceder créditos, el error obtenido en general es un error considerable, sin embargo, presenta un error menor en la clase "0". Es decir, se equivoca menos al momento de otorgar créditos a personas que no son rentables y que con una probabilidad mayor le traeran pérdidas a la entidad financiera, es el error que intentabamos mejorar, sin embargo, el error de la clase "1" sigue siendo muy grande lo que nos dice que no estamos otorgandole créditos a personas que sí serán más rentables ya que exite una mayor posibilidad de que sí se financie con la tarjeta y además devuelva el crédito otorgado.

Seguiremos haciendo nuestro análisis, ya que existe la posibilidad de que al aplicar las rutinas de olvq, lvq1, lvq2 y lvq3 mejore la clasificación, primero haremos uso de la rutina de olvq1 ya que la literatura sugiere aplicar este método después de la iniciación ya que tiene una mejor convergencia. Debido a esto presentamos a continuación la clasificación que se obtuvo:

Clasificación con olvq 1:

n n n n n n n s n n n n s n n n n n n s s n n n s n n n s n n n
n n n n n n s s n n n n n n n s n n n n n n n n n n n n n n n n
n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n
n n n s n n n n n s n n n n n n n n n n n s n n n n n s n n n n n
n n s n n n n n n n n n s n n n n n n n n n n n n n n n n n n n
n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n n s s s n s n n s s s n s s n
s n s n n n n s s s s s s s s s s n s s s n n s s s s n s n n n n n n s s n n n n n n
s n s s n n n s n s s n n n n s n s s s s s n s n s s n n s s n s n n s s s n n n s s s s
n n n n s s n s n s n n s n n s s n s n n n s s s s s n n n n n s n s n n s n s s n s n n n s n n n

Lo que podemos observar es que los errores de clasificación se mantienen, aunque la clasificación no sea exactamente la misma que la anterior ya que algunos elementos que en la clasificación de iniciación decide no otorgar créditos como debería de ser ahora decide otorgarle el crédito, analogamente ocurre para la clase "1". Ahora verifiquemos si haciendo uso de los otros métodos sugeridos lvq1, lvq2.1 y lvq3 la clasificación mejora, se mantiene igual o empeora. Lo que nos agradaría es que mejorara ya que eso sería satisfactorio para una entidad de crédito aunque es de esperarse que su mejoria sea mínima ya que observamos que la clasificación no mejoró en nada cuando se obtuvo el conjunto de iniciación a la clasificación obtenida por olvq1.

Haremos uso de los siguientes métodos que nos llevarán a observar si la clasificación mejora, solo hay que aclarar que ahora en lugar de usar el conjunto de "codebook" de iniciación emplearemos el que se obtuvo mediante olvq1. Los resultados se muestran en el cuadro 5.4:

	lvq1		lvq2.1		lvq3	
Decisión	***s***	***n***	***s***	***n***	***s***	***n***
n	19	172	16	175	25	166
s	80	74	77	77	71	83

Cuadro 5.4: Clasificación mediante LVQ

En el cuadro 5.4 se observan los errores de clasificación, para los métodos lvq1 y lvq2.1 se tiene un error total de clasificación del 26.95 % mientras que el método lvq3 presenta un error del 27.82 % los primeros 2 métodos mencionados clasifican mal 93 de los 345 datos y el método lvq3 clasificó mal 96 de los 345, esto nos indica que los métodos lvq1 y lvq2.1 mejoraron en poco la clasificación. En la primer clasificación que se realizó con olvq1 teniamos 94 individuos mal clasificados. Ya vimos que se mejoró, ahora observemos como fue su clasificación por cada clase, lo que podemos ver es que lvq2.1 clasifica 16 individuos mal mientras que lvq1 clasifica mal 19 individuos pero si recordamos olvq1 clasificó mal sólo 15 individuos lo que nos estaría indicando que para la clase "0" sigue siendo mejor el método de olvq1. Mientras que para la clase "1", lvq1 clasifica mal a 80 individuos, lvq2.1 a 77 y olvq1 clasificó mal a 79 individuos, lo que nos lleva a pensar que para esta clase las variables no logran explicar la diferencia, si es que existe, entre ambos grupos, y no nos da facilidad de decidir si un nuevo individuo pertenece a la clase "1" o a la clase "0" y así saber si se le otorga o no se le otorga el crédito y la cantidad de crédito solicitada.

Con lo anterior podemos apreciar que la clasificación es mala pero podría mejorar si se aumentan o se disminuyen la cantidad de prototipos o si quitamos alguna variable. Presentamos el cuadro 5.5 en el que se muestran las clasificaciones de inicialización con diferente número de prototipos y con las mismas variables que se utilizaron para hacer el análisis anterior:

Iniciación			
# prototipos	***Errores "0"***	***Errores "1"***	% ***Error total***
1	0	154	44.63
5	14	106	34.78
25	21	72	26.95
50	26	71	28.11

Cuadro 5.5: Clasificación variando el número de prototipos

Como podemos ver en el cuadro 5.5 a mayor número de prototipos mayor error en la clase "0" pero menor en la clase "'1", pero no es algo uniforme, ya que si ahora vemos los porcentajes totales notamos que poner muchos prototipos no refleja una mejor clasificación si no que logra obtener el mismo porcentaje de error que se obtiene al utlizar 25 prototipos. Entonces podemos observar que al menos para esta base de datos no es necesario colocar muchos prototipos. Pero como se puede ver aunque se coloquen más o se disminuyan la clasificación sigue siendo mala.

Ahora consideraremos quitar alguna variable o colocar la variable 13 que fue la que se quitó en el primer análisis o hacer combinaciones entre las variables. En el primer análi-

sis observamos que utilizando 25 prototipos se conseguía un porcentaje mejor de error, por lo que observamos las clasificaciones que se obtenía mediante esto pero los resultados no fueron favorables. Si colocamos la variable 13 para conseguir mayor información obteniamos que de los 191 individuos pertenecientes a la clase "1" solo clasificaba bién a 151 y a los 40 restantes decidía otorgarles el crédito cuando tenía que negárselos, en la clase "1" de los 154 individuos solo a 66 clasificaba bién y a los restantes 88 decidia negárselos cuando en realidad tenia que otorgarle el crédito. Esta clasificación provoca un error total del 37 %, mayor error al obtenido con las primeras variables, después se decidió quitar otra variable, es decir de las 6 variales solo quedarnos con 3, estamos quitando más información y es posible que nos arroje un mayor porcentaje de error, pero veamos, de los 191 individuos de la clase "0" clasificó bien a 115 y erroneamente a los restantes 76, y de la clase "1" clasifica bién a 97 y mal a 57, provocando un error total del 39 %. Una vez más mayor al que se tenía. Así mismo realizamos combinaciones con las variables pero los resultados no mejoraron por lo que se decidió que la "mejor" clasificación a la que se había llegado era con las primeras 4 variables y con 25 prototipos.

Se observó que con 25 prototipos se obtenía un error total inferior al que obtuvimos con 15 prototipos, entonces veamos si aplicando los métodos logramos mejorar un poquito más las clasificación. A continuación mostraremos una tabla con las clasificaciones que realizaron los diferentes metodos:

	olvq1		lvq1		lvq2.1		lvq3	
Decisión	***s***	***n***	***s***	***n***	***s***	***n***	***s***	***n***
n	23	168	20	171	24	167	22	169
s	82	72	81	73	86	68	84	70

Cuadro 5.6: Clasicación con 25 prototipos y LVQ

En el cuadro 5.6 se observa que se logra un menor porcentaje de error total ya que de 27 % se llega casi al 26 % aunque observemos que el error en la clase "0" aumentó pero disminuyó en la clase "1" y de esta tabla la "mejor" clasificación obtenida es a la que llega el metodo lvq3. Pero, es mejor quedarnos con la primera clasificación o con esta última? Si lo vemos desde la perspectiva de una entidad financiera, lo recomendable sería quedarnos con la primera, ya que como lo hemos estado mencionando a lo largo de este trabajo, es más costoso otorgar créditos a personas que no son para nada rentables y que además no regresaran el crédito ya que esto ocasionarían pérdidas financieras superiores en comparación a no otorgar créditos a personas que si son rentables, por lo tanto la "mejor" clasificación sería la que se obtuvo mediante lvq2.1, 15 prototipos y 4 variables.

5.3. KNN y LVQ

En esta sección compararemos ambos métodos para intentar decidir si alguno es mejor que el otro. Como pudimos observar no se logró llegar a una buena clasificación, por la estructura de la información de la base de datos y por que algo muy importante que notamos es que las variables medidas no pueden hacer diferencias entre un grupo y otro para poder llegar a una buena clasificación.

Observemos las diferencias o similitudes, si es que existen, entre ambos métodos mostrando las tablas que se obtuvierón cuando se hicieron los análisis respectivos para cada método. Observemos si podemos señalar a alguno de ellos como "mejor" método.

Estos datos fueron obtenidos mediante KNN:

# Prototipos	**Errores 1**	**Errores 0**	% Error 1	% Error 0	% Total
1	63	41	40.9	21.46	30.14
5	68	30	44.15	15.70	28.40
10	**70**	**21**	**45.45**	**10.99**	**26.37**
15	76	20	49.35	10.47	27.82
25	76	21	49.35	10.99	28.11
45	81	18	52.59	9.42	28.69
225	132	1	85.71	0.52	38.55

Cuadro 5.7: Clasificación mediante KNN

Estos fueron los datos obtenidos por LVQ:

Inicialización					
# prototipos	***Errores 0***	***Errores 1***	% ***Error 0***	% ***Error 1***	% ***Total***
1	0	154	0	100	44.63
5	14	106	7.32	68.83	34.78
15	**15**	**79**	**7.85**	**51.59**	**27.24**
25	21	72	10.99	46.75	26.95
50	26	71	13.61	46.10	28.11

Cuadro 5.8: Clasificación con el conjunto de Inicialización

Los resutados presentados en el cuadro 5.7 y en el cuadro 5.8 fueron arrojados bajo la misma muestra de entrenamiento, así mismo mediante la misma muestra de prueba, sin embargo en lvq se pudo mejorar utilizando el metodo lvq2.1 por 15 prototipos para construir los vectores "codebook" y esta clasificación se muestra a continuación:

	olvq1		lvq1		lvq2.1		lvq3	
Decisión	***s***	***n***	***s***	***n***	***s***	***n***	***s***	***n***
n	15	176	19	172	**16**	**175**	25	166
s	75	79	80	74	**77**	**77**	71	83

Cuadro 5.9: Clasificación con 15 prototipos y todas las rutinas de LVQ

En la clasificación obtenida mediante el método de KNN que se presenta en el cuadro 5.9 se ve claramente que en la clase "0" a mayor número de K-prototipos menor es el error que se comete en esta clase, mientras que en el cuadro 5.9 se observa que en el método de LVQ pasa lo contrario ya que a mayor número de prototipos mayor es el error que se comete. Sin embargo, para ambos métodos el porcentaje de error en la clase "0" es menor en comparación con la clase "1", pero se obtiene un menor porcentaje de error mediante el método de lvq2.1 con 15 prototipos ya que se logra un porcentaje de error del 8 % mientras que en KNN se logra obtener un porcentaje del 11 %, 3 % superior con respecto al obtenido por lvq2.1. Para la clase"0" indica que es mejor LVQ ya que el porcentaje obtenido es inferior al de KNN y nos estaría indicando que con una menor probabilidad le otorgaremos crédito a personas que no son rentables y que a consecuencia se obtendran menores pérdidas.

Por otro lado para la clase "1" mediante KNN se obtiene un error del 45 % mientras que mediante lvq2.1 se obtuvo un error del 50 %, superior en 5 % al obtenido por el método de KNN. Para esta clase el método KNN es "mejor" que LVQ, aunque el error cometido en general para esta clase en considerablemente grande y esto también nos trae pérdidas en una entidad financiera dado que no se esta concediendo el crédito a un porcentaje considerable las pérdidas se vuelven mayores. Pero en general el error cometido es casi el mismo, razón por la que no podemos afirmar que un método sea mejor que el otro.

Conclusiones

En este trabajo se analizaron cuatro métodos de clasificación dos de ellos de tipo paramétrico: *análisis discriminante* y *regresión logística*, y otros dos más de tipo no paramétrico: "*K-nearest neighbour*" y "*learning vector quantization*". Asímismo se hizo la respectiva aplicación a la base de datos australiana de *"Credit Scoring"*.

- De los métodos paramétricos no resulta claro si el análisis discriminante es mejor que la regresión logística, debido a que los errores de clasificación no disminuían conjuntamente para las dos clases.
- En el caso de los métodos no paramétricos la conclusión es muy similar ya que los errores cometidos en las clasificación en KNN era mejor para la clase "1", mientras que LVQ era mejor para la clase "0". El error total, es casi el mismo, por lo tanto inclinarse por un método en especial es difícil.
- Entre los métodos paramétricos y no paramétricos podríamos inclinarnos por los paramétricos ya que el error de clasificación fue menor para éstos y más en el caso de regresión logística, además de que son métodos de mejor comprensión en cuanto a que son más familiares en diversas áreas.
- Llegamos a la conclusión que las variables expuestas por la base de datos no medían su objetivo, ya que no lograban mostrar ninguna diferencia entre los dos grupos. Esto no permite tomar una decisión adecuada basada exclusivamente en el análisis estadístico de los datos para un nuevo solicitante de crédito. Con mucha facilidad podríamos equivocarnos.
- Hay que tener en cuenta que la base de datos es real además de que es una base un poco antigua y la muestra es una muestra truncada, uno esperaría que en la actualidad esta entidad financiera halla modificado las variable medidas para un mejor estudio y tener una mejor toma de decisión. Es claro que estos métodos no actúan de forma automática; sin embargo es de mucha ayuda lo que indique el método estadístico para poder tomar una decisión.

- Una gran desventaja fue que no conociamos qué era especificamente lo que medían cada una de las variables, desventaja que nos conducia a no saber cuál era el contenido de las variables que se implementaban a los modelos.

- Es importante hacer notar que estas cuatro técnicas pueden ser manejadas en **R - project**, sin embargo, en el caso de los métodos no paramétricos es necesario procesar estas técnicas hasta su "mejor" convergencia, ya que uno de los objetivos computacionales de los métodos "automatizados" es que trabajan con el conjunto completo de referencia e intentan reducir el número de calculos, de tal manera que no realizan aquellos calculos que consideran innecesarios, mientras qué los métodos paramétricos, proporcionan un modelo para la toma de una decisión.

Bibliografía

Agresti Alan (2007) *"An introduction to categorical data analysis"*. Second edition. Universidad de Florida. Wiley–Interscience.

Dillon William R., Goldstein Matthew (1984) *"Multivariate Analysis"* `Methods and applications`. John Wiley & Sons.

Friedberg Stephen, Insel Arnold y Spence Lawrence (1982) *"Algebra lineal"*, Illinois State University. Primera Edición. Publicaciones Cultural, S.A.

Hastie Trevor, Tibshirani y Friedman Jerome (2008) *"The Elements of statistical learning"* `Data mining, inference and prediction`. Second Edition. Springer.

Mardia K.V., Kent J.T., Bibby J. M. (1995) *"Multivariate Analysis"*. Academic Press Limited.

Ripley B. D. (1996) *"Pattern Recognition and neural network"*. Cambridge.

Rencher Alvin C.(1934) *"Methods of multivariate analysis"*.John Wiley & Sons

Thomas Lyn C., Edelman David B. y Crook Jonathan N. (2002) *"Credit Scoring and its Applications"*. *Siam*, Monographs of Mathematical Modeling and Applications.