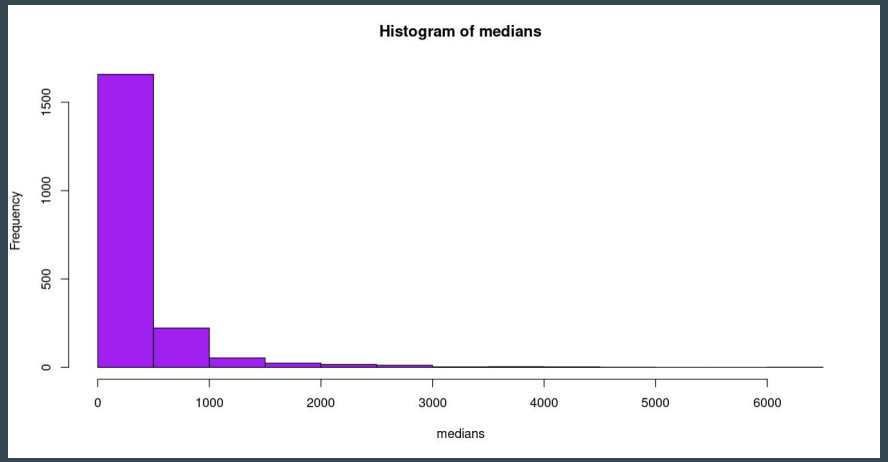
Trabajo Práctico Final
Aprendizaje Estadístico

Clasificación de muestras de tejido en ‘sanos’ o ‘con tumor’ a partir de expresiones de genes

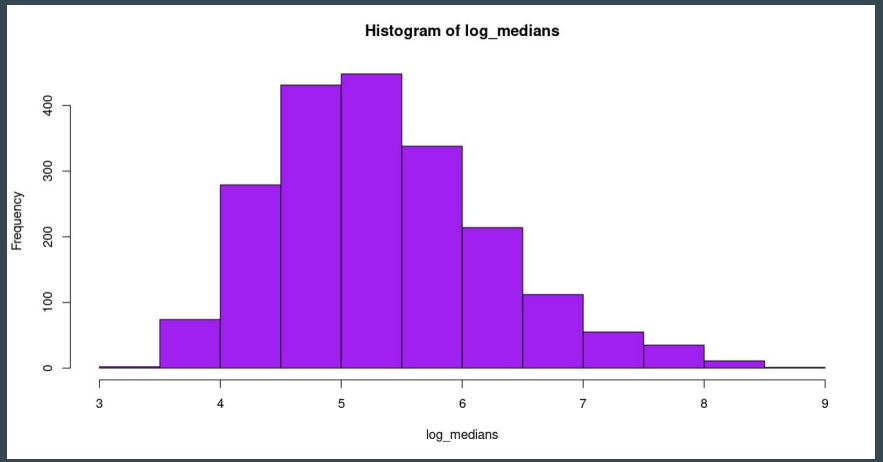
---

Facultad de Ingeniería - Universidad de Buenos Aires
Profesora: Jemina García
Alumnos: Federico Elías - Ignacio Brusati
Cuatrimestre: 1C2022

Normalización de los datos



Medianas de los datos originales



Medianas luego de la normalización

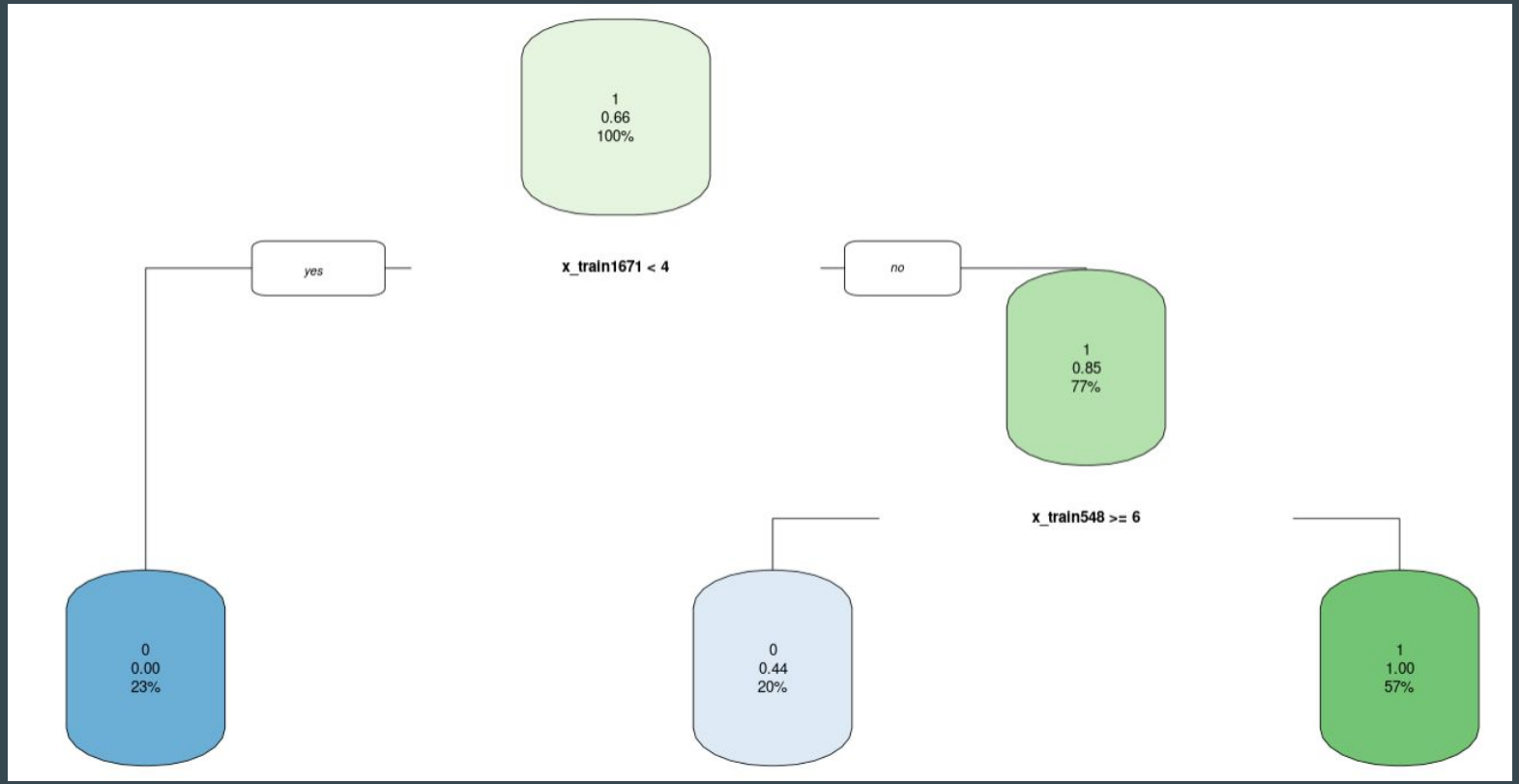
Árbol de Clasificación

- Este método va segmentando el espacio de covariables a partir de ciertas reglas hasta llegar a varias regiones pequeñas (llamadas nodos terminales u 'hojas')

- Cuando llega una nueva observación, se la ubica en una de esas hojas dependiendo de las reglas de decisión que se fueron tomando para dividir el espacio de covariables

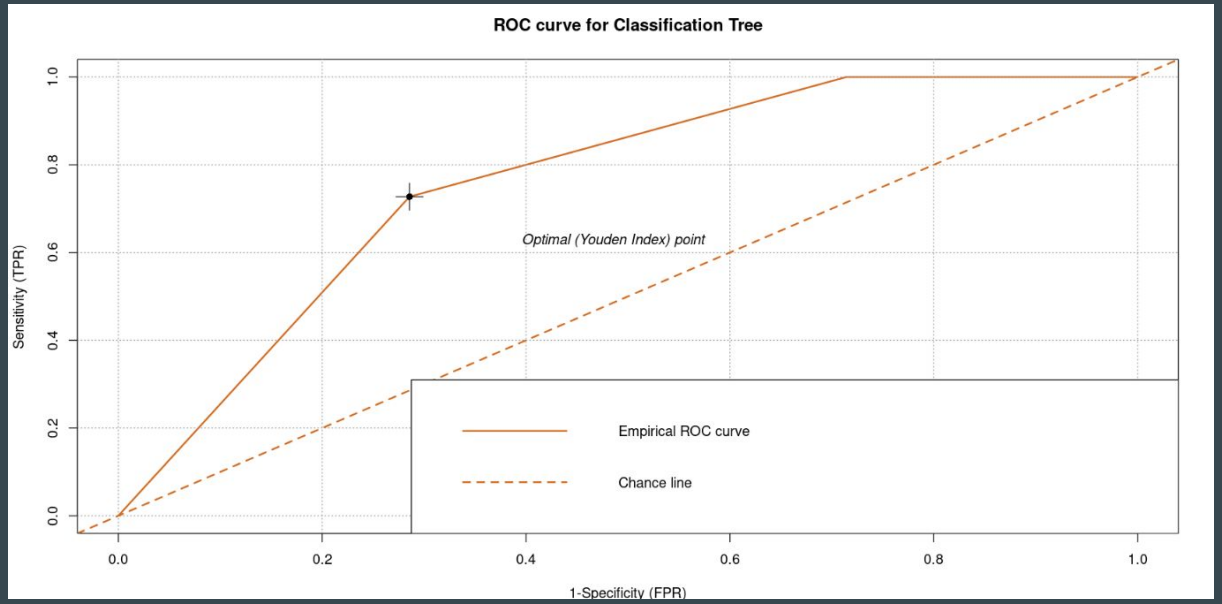
- Para dar una predicción a esa nueva observación se utilizan las observaciones del set de entrenamiento que se tenían en la hoja en la que cayó la nueva observación usando la regla de la mayoría

Árbol de Clasificación



Error para el modelo de Árbol de Clasificación

La probabilidad a partir de la cual se clasifica como verdadero es 1.0000000



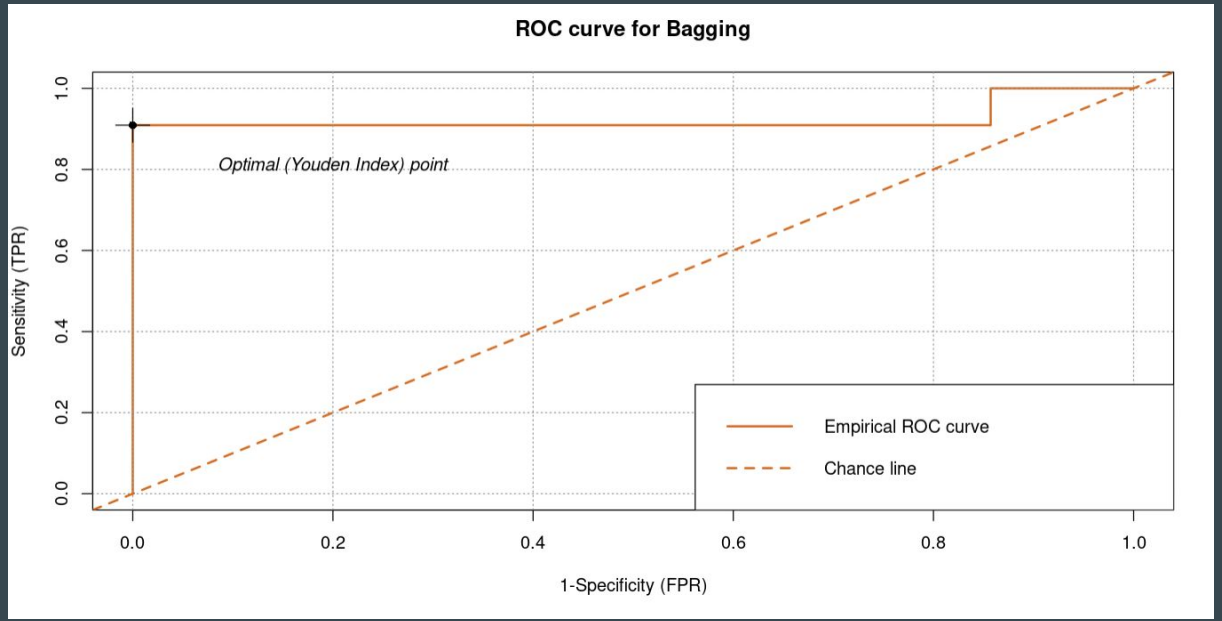
El área bajo la curva ROC es 0.7597403 y el error de clasificación es 0.4176955

Bagging

- Este método propone utilizar diferentes muestras y construir un árbol para cada una de ellas
- Para la clasificación de una nueva observación se utiliza la regla de la mayoría usando las clasificaciones de esos árboles
- Las nuevas muestras se consiguen mediante bootstrap
- Se realizó Bagging con 15 árboles

Error para el modelo de Bagging

La probabilidad a partir de la cual se clasifica como verdadero es 0.6699144

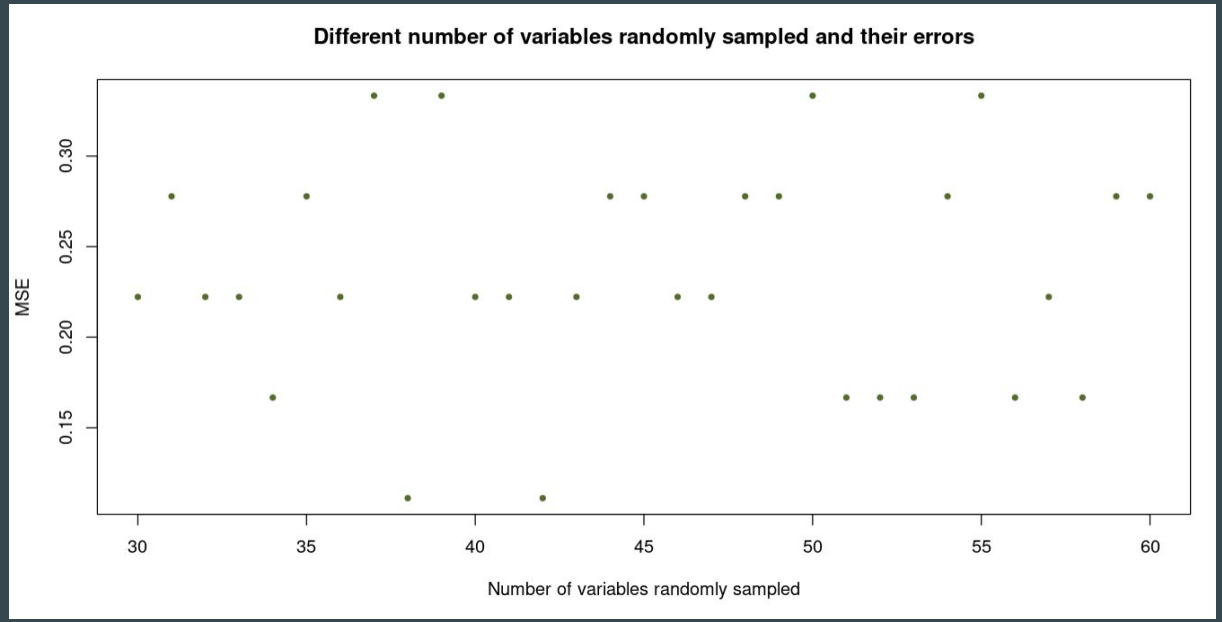


El área bajo la curva ROC es 0.9220779 y el error de clasificación es 0.139266

Random Forest

- Como en Bagging, se construyen árboles bootstrapeando muestras de entrenamiento
- La diferencia con Bagging es que este método propone que cada vez que se realiza una partición se elija al azar m de las p covariables disponibles
- Se logra independizar a los estimadores buscando una mayor reducción en la varianza del estimador final

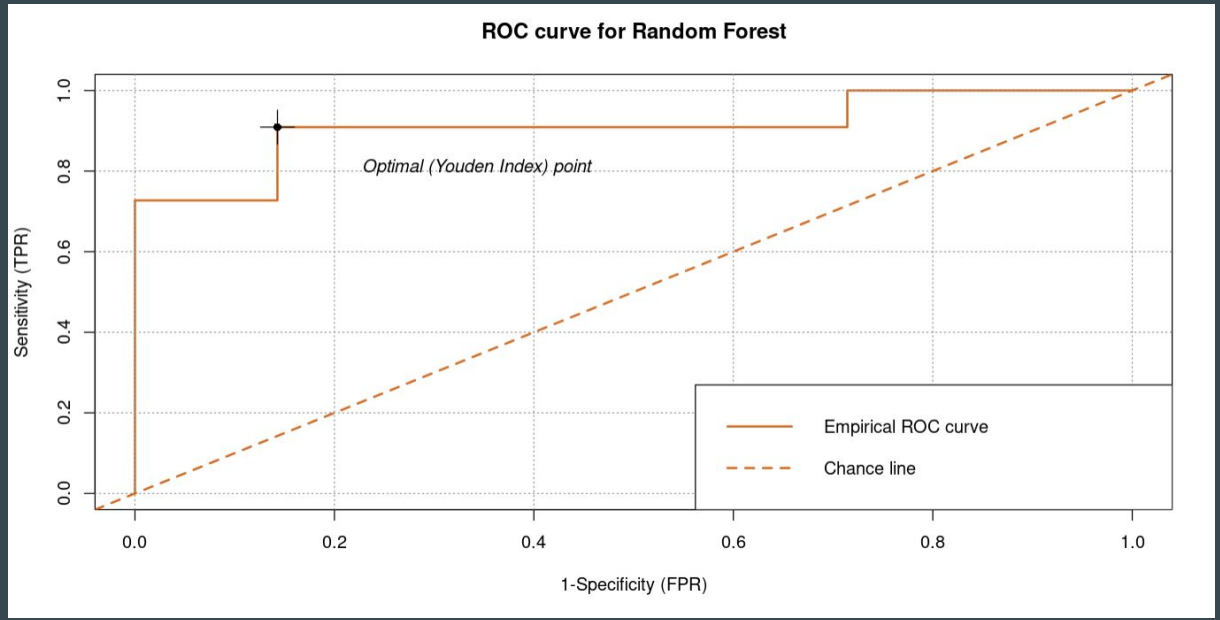
Búsqueda del parámetro de cantidad de variables separadas



Se probaron todos los valores enteros entre 30 y 60 y el óptimo fue 38

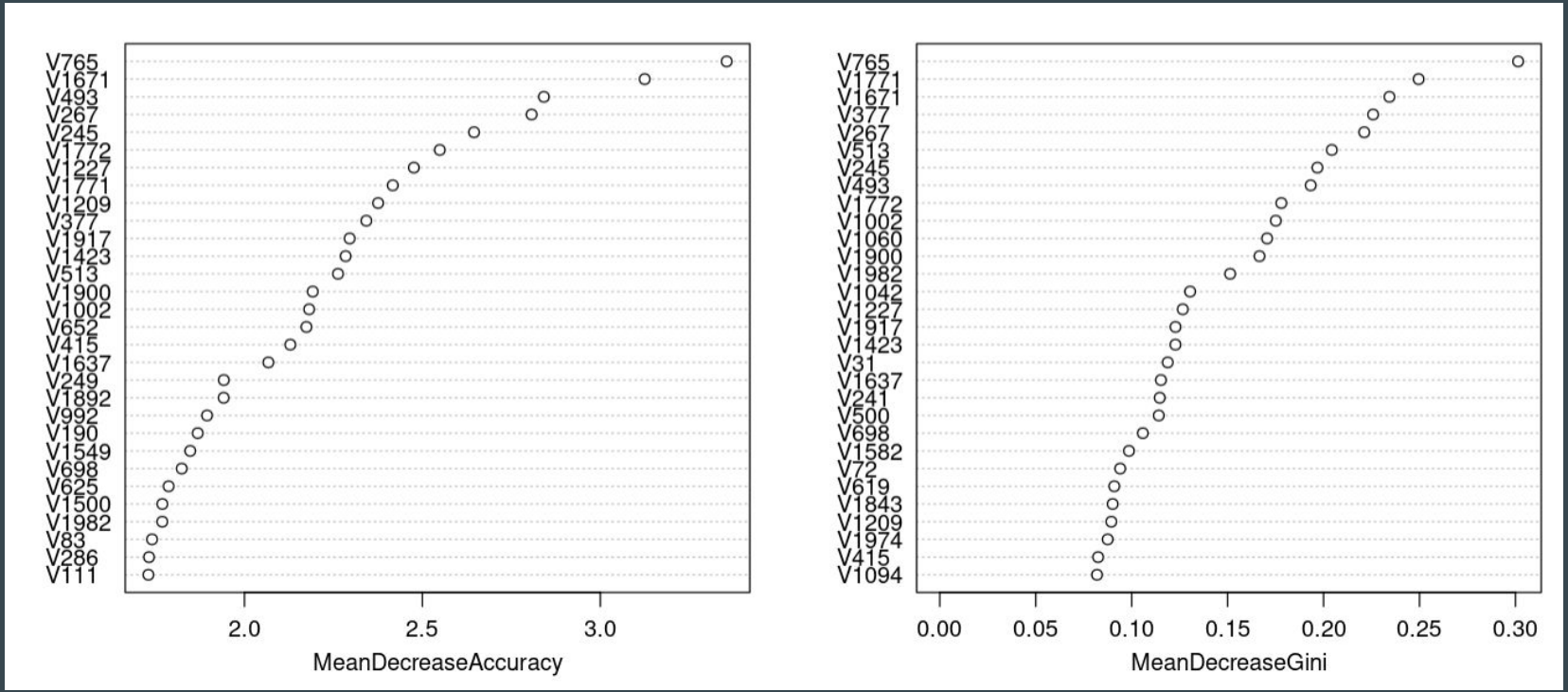
Error para el modelo de Random Forest

La probabilidad a partir de la cual se clasifica como verdadero es 0.5920000



El área bajo la curva ROC es 0.9090909 y el error de clasificación es 0.1111111

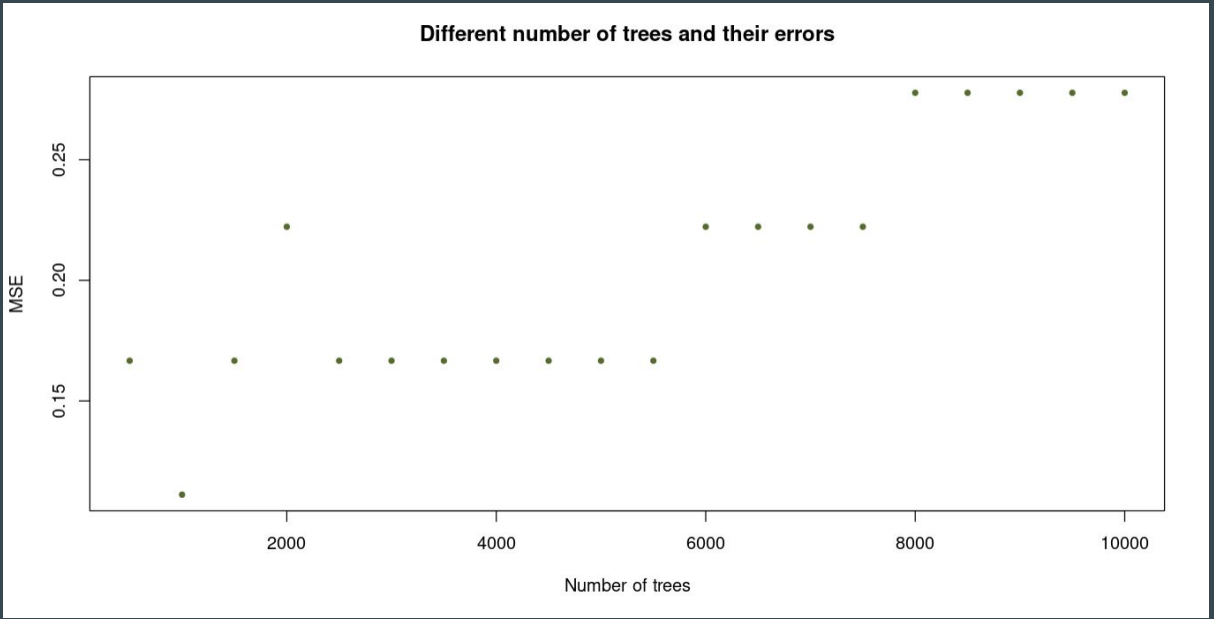
Selección de variables con Random Forest



Boosting

- Funciona de una forma similar a Bagging pero los árboles se crean de forma secuencial
- Se predice a los residuos, no a las respuestas
- En cada paso se actualiza la estimación final sumando una proporción de la estimación realizada en ese paso

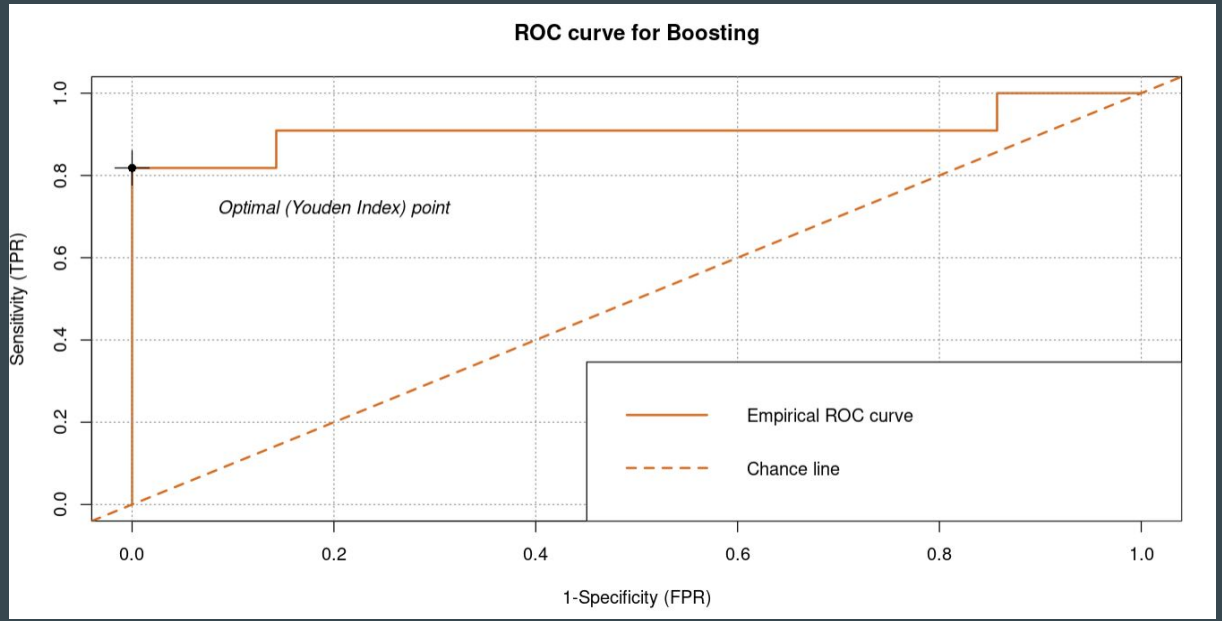
Búsqueda del parámetro de cantidad de árboles



Se probaron un rango de valores entre 500 y 5000 con un salto de 500 y el óptimo fue 1000

Error para el modelo de Boosting

La probabilidad a partir de la cual se clasifica como verdadero es 0.8514867

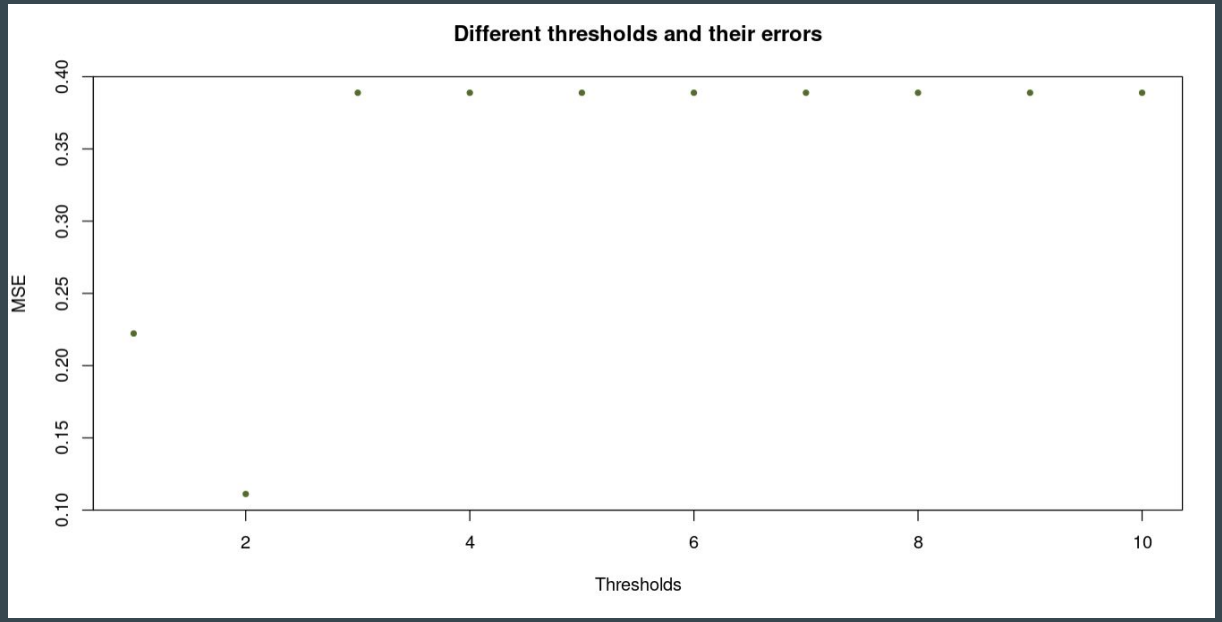


El área bajo la curva ROC es 0.909090 y el error de clasificación es 0.1111111

Nearest Shrunken Centroids

- Calcula un centroide para cada gen para cada clase
- Reduce los centroides en una cantidad llamada *threshold*
- Descarta los centroides de los genes que den 0 para todas las clases
- Al llegar una nueva observación, se toman los valores de esos genes y se los comparan con cada uno de estos centroides de clase: la clase cuyo centroide está más cerca es la clase predicha para esa nueva observación
- Para los siguientes modelos, se usaron los genes encontrados por este método

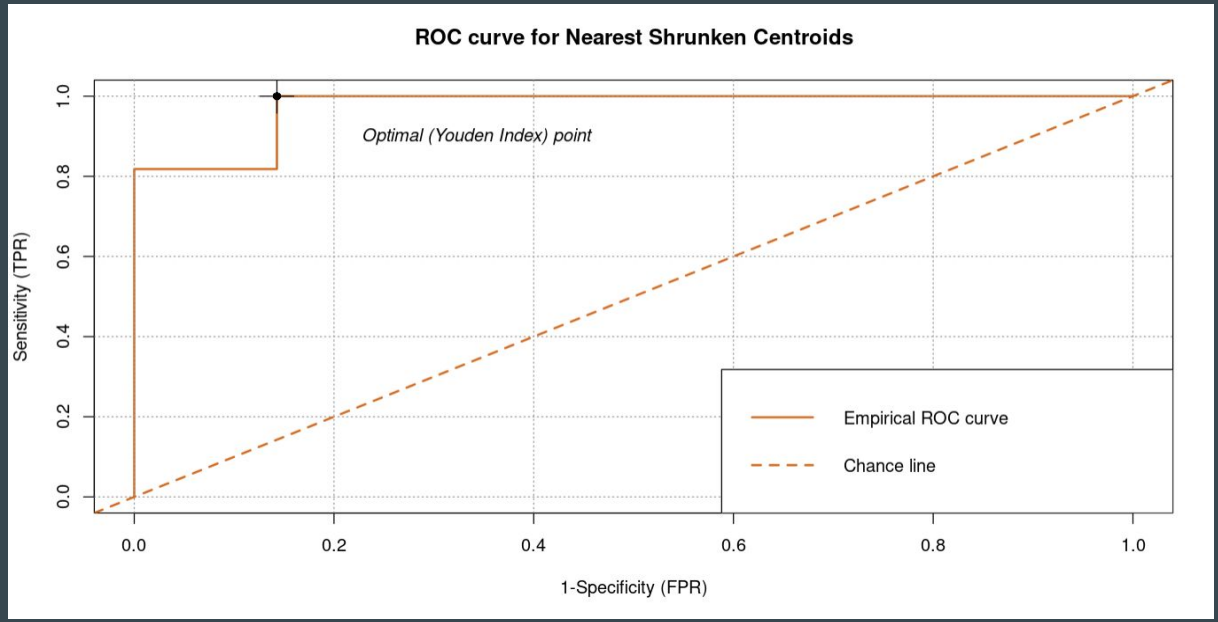
Búsqueda del parámetro de threshold



Se probaron un rango de valores entre 1 y 10 con un salto de 1 y el óptimo fue 2

Error para el modelo de Nearest Shrunken Centroids

La probabilidad a partir de la cual se clasifica como verdadero es 0.6044035



El área bajo la curva ROC es 0.974026 y el error de clasificación es 0.1111111

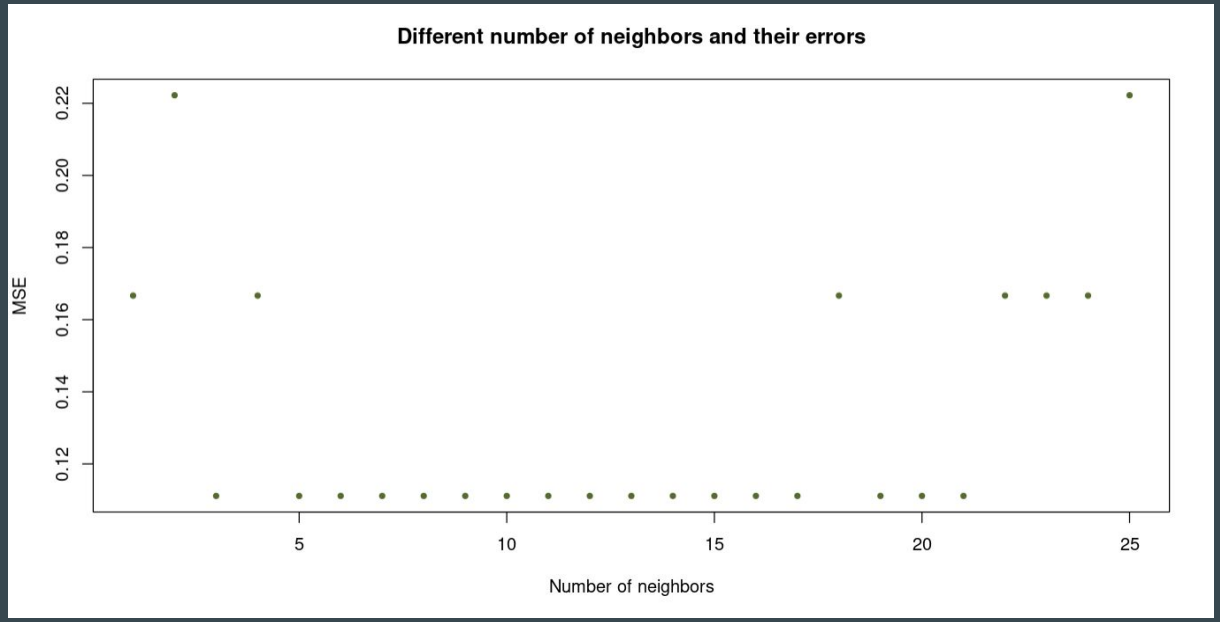
K-Nearest Neighbors (KNN)

- Sirve para estimar la distribución de Y dado X y después clasificar una observación en la clase con mayor probabilidad estimada

- A un nuevo punto se le asigna una categoría por voto de la mayoría entre los k vecinos más cercanos

- Se utilizó la distancia euclídea como métrica de distancia, luego se usó cross validation para obtener el k óptimo

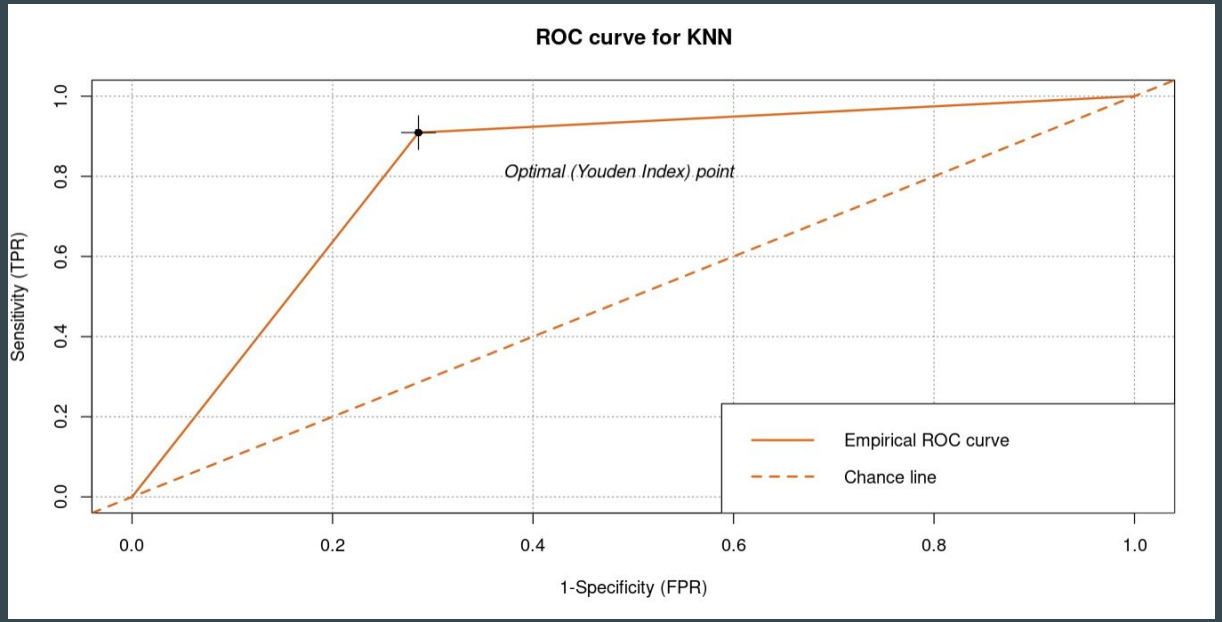
Búsqueda del parámetro de cantidad de vecinos



Se utilizó un rango de valores de k entre 1 y 25 y el óptimo resultó ser un valor de k=3 vecinos

Error para el modelo de KNN

La probabilidad a partir de la cual se clasifica como verdadero es 1.0000000



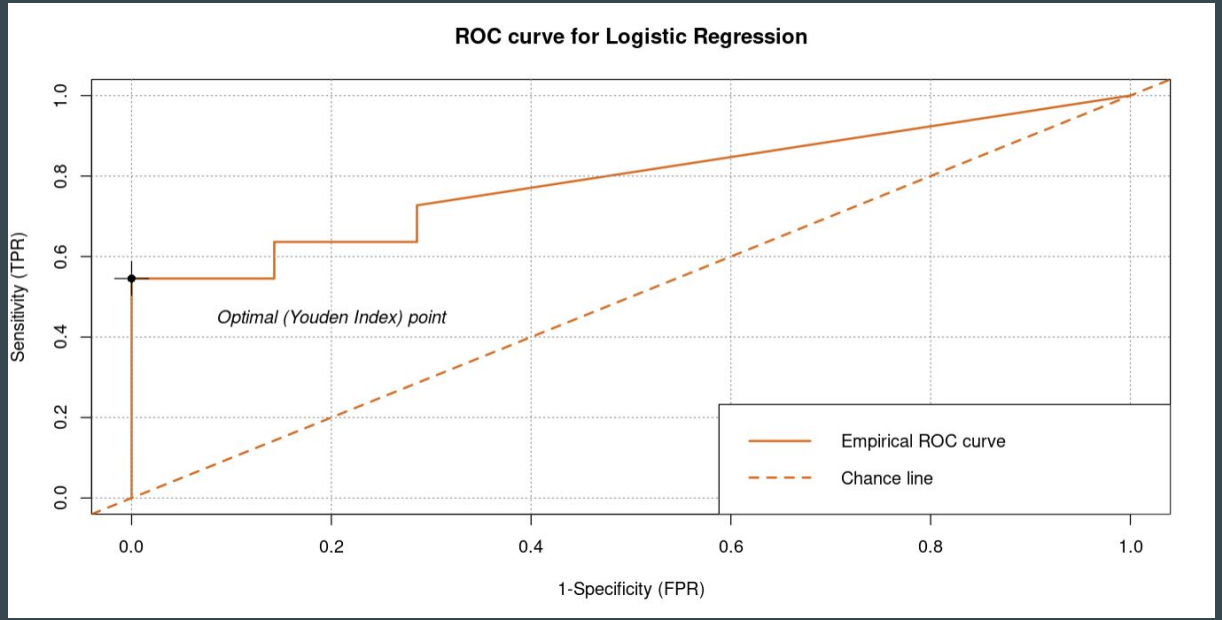
El área bajo la curva ROC es 0.8116883 y el error de clasificación es 0.1111111

Regresión Logística

- Modela la probabilidad de que una nueva observación Y pertenezca a una clase en particular dada una serie de covariables X1, X2, ..., Xp
- Se estima una esperanza condicional usando la función logística
- Se estiman los parámetros β mediante el método de máxima verosimilitud
- El sistema al que se llega con el método de máxima verosimilitud debe ser resuelto mediante métodos numéricos

Error para Regresión Logística

La probabilidad a partir de la cual se clasifica como verdadero es casi 0

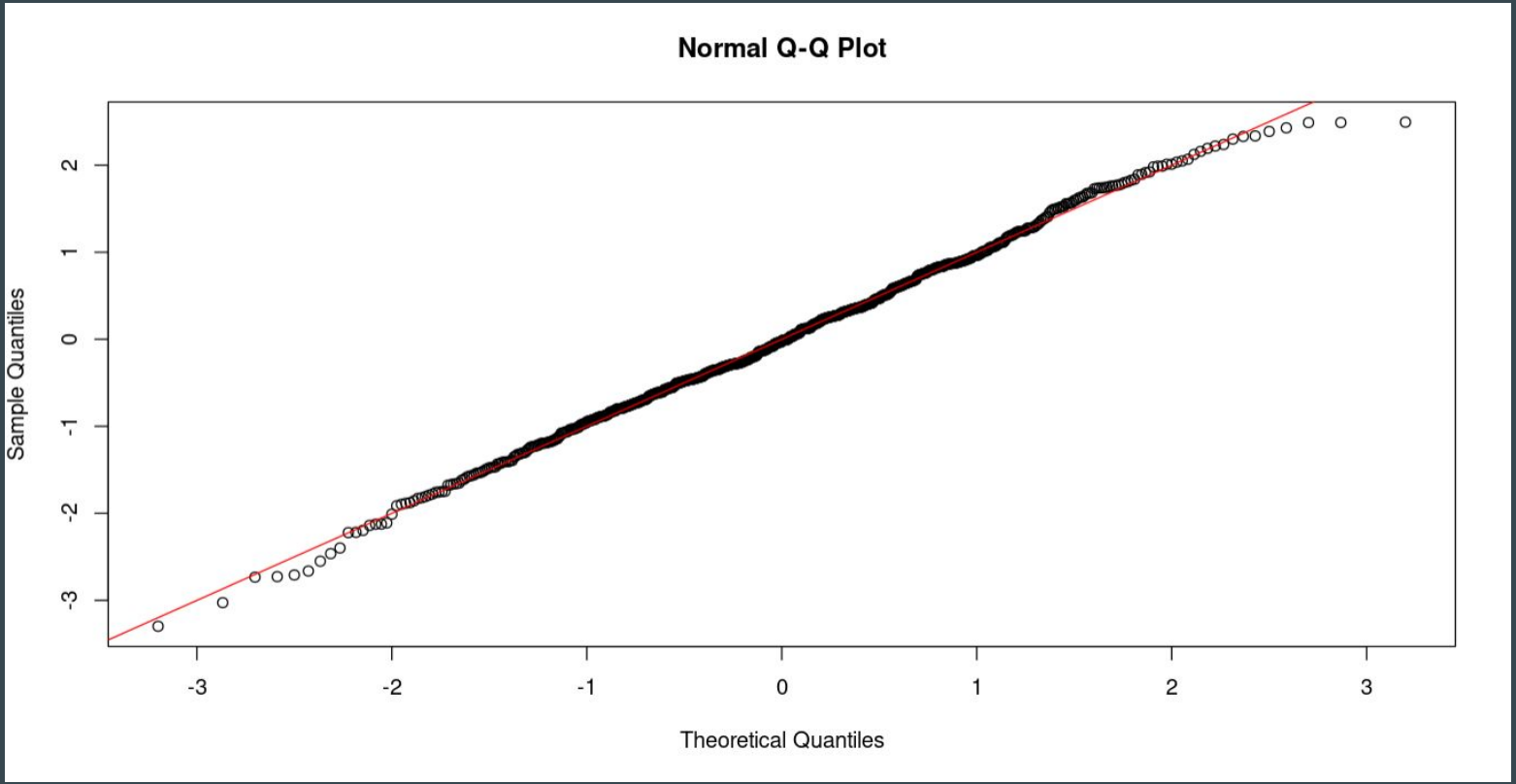


El área bajo la curva ROC es 0.7857143 y el error de clasificación es 0.3333333

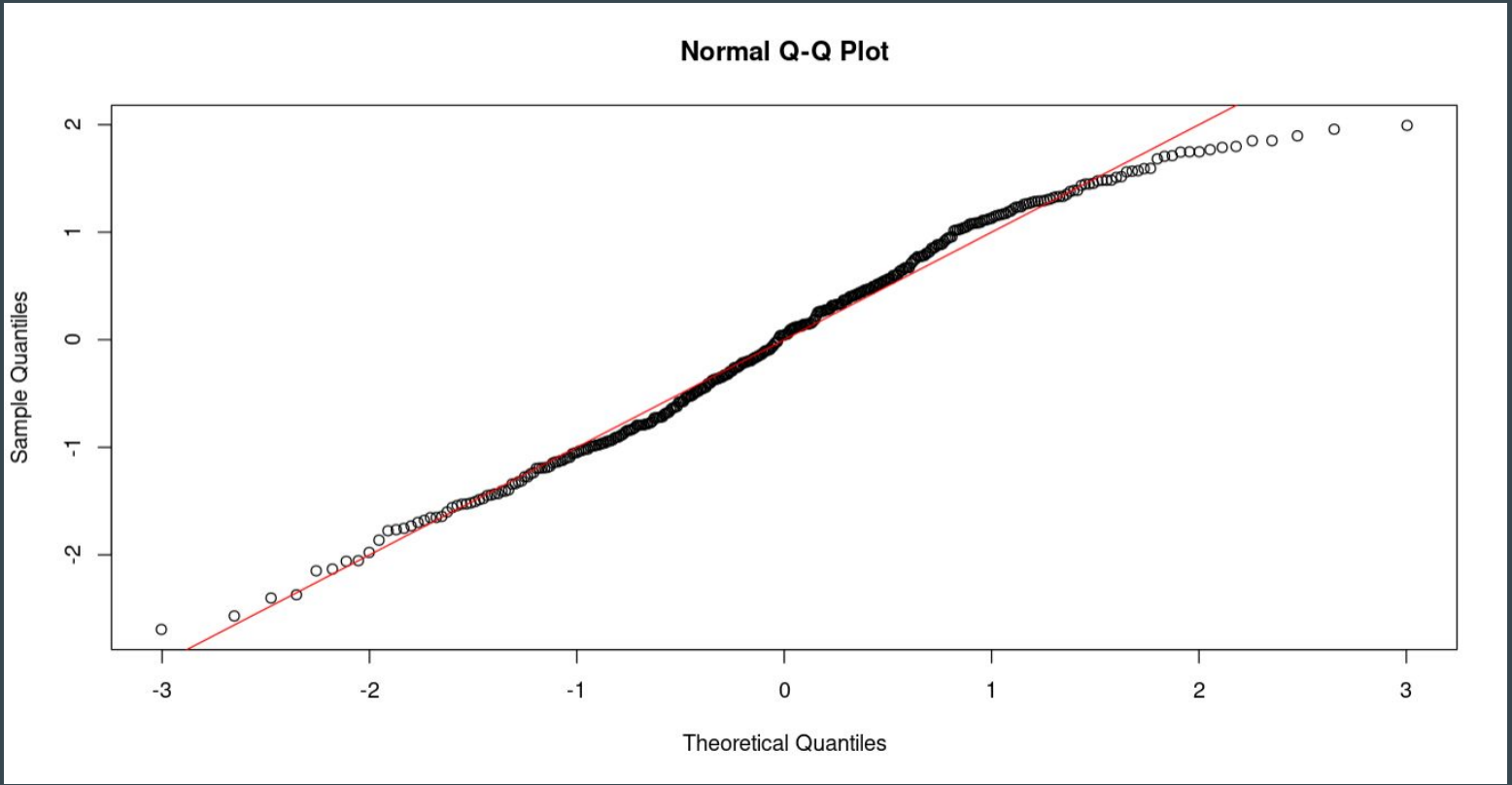
Análisis Lineal Discriminante (LDA)

- Sirve para clasificar los genes dentro de dos o más grupos de poblaciones
- Las poblaciones en este caso son “enfermos” y “sanos”
- Este método asume que las clases tienen la misma varianza
- También asume que las clases tienen distribución normal multivariada
- Primero se verifica normalidad para ambos grupos de pacientes

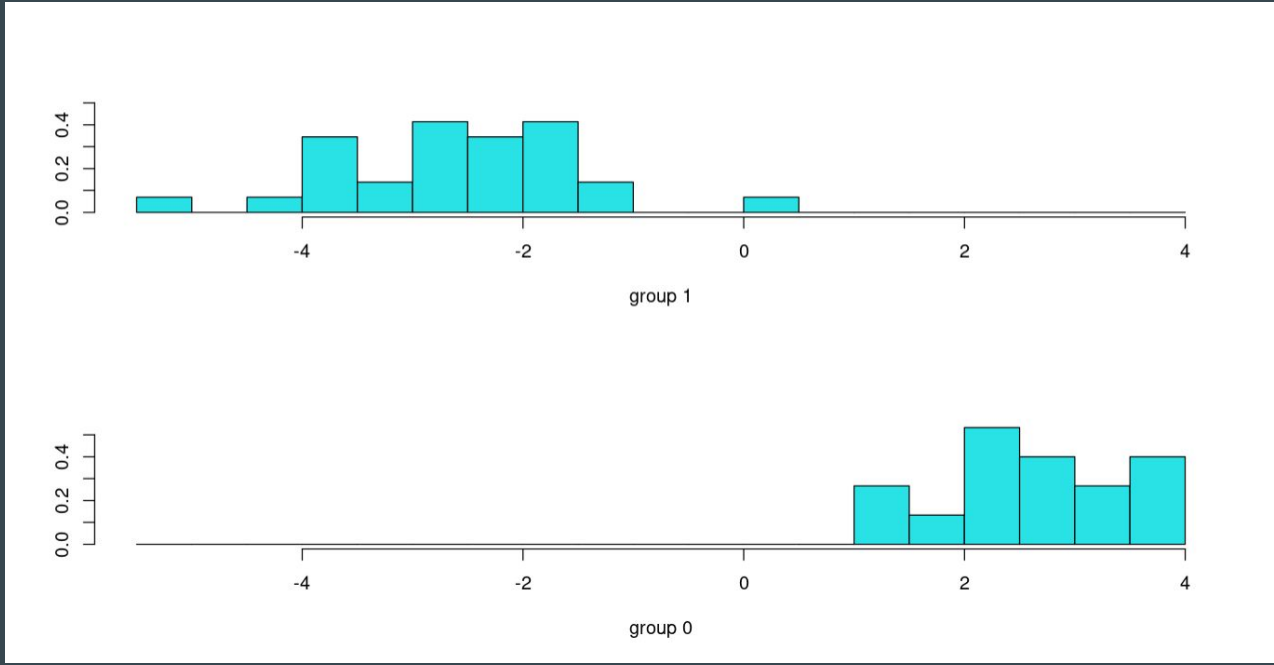
Normalidad para los enfermos



Normalidad para los sanos



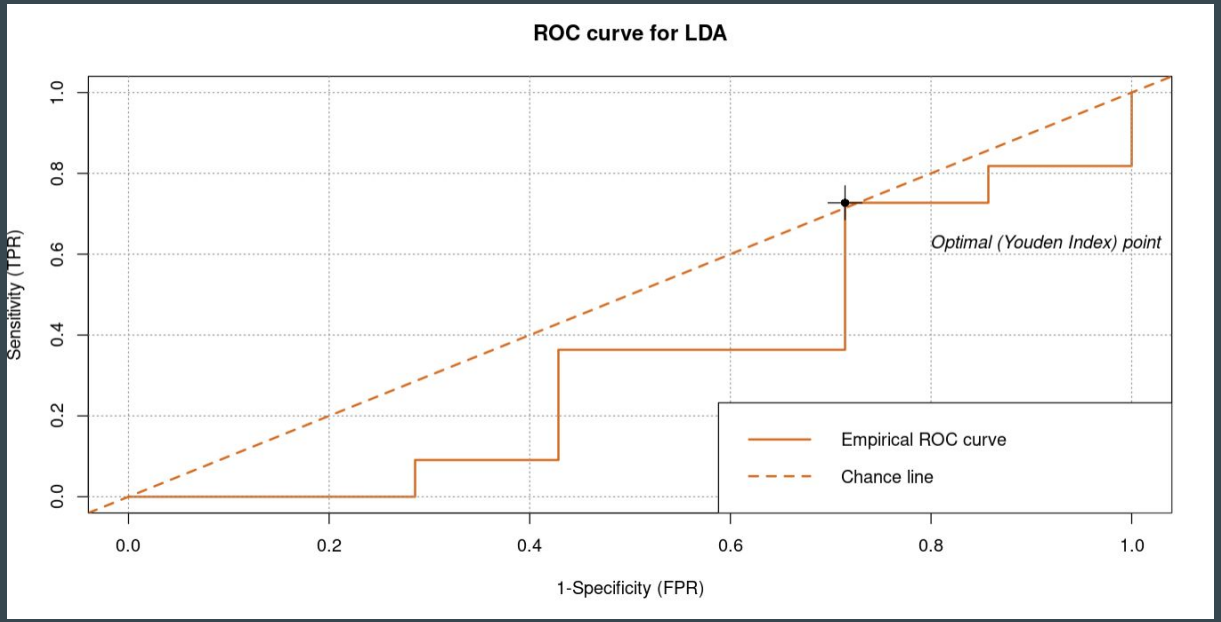
Distribuciones de enfermos y sanos



Las distribuciones tienen diferentes medias pero no tienen la misma varianza

Error para el modelo LDA

La probabilidad a partir de la cual se clasifica como verdadero es casi 0



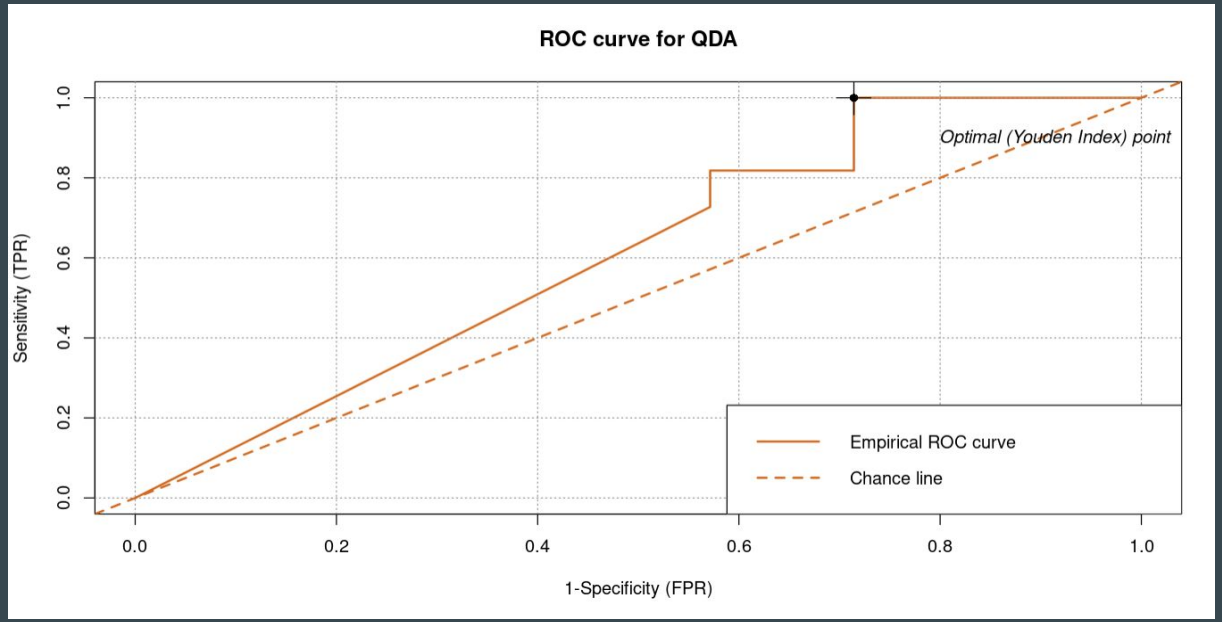
El área bajo la curva ROC es 0.3376623 y el error de clasificación es 0.4444444

Análisis Cuadrático Discriminante (QDA)

- Similar al LDA, pero con el supuesto de que las matrices de covarianza son diferentes
- De los 25 genes obtenidos por el NSC, usamos 12

Error para el modelo QDA

La probabilidad a partir de la cual se clasifica como verdadero es 1.0000000



El área bajo la curva ROC es 0.6103896 y el error de clasificación es 0.7222222

¡Muchas gracias!