CAMPUS
DE EXCELENCIA
INTERNACIONAL

PREDICCIÓN DE LA EVOLUCIÓN DE PACIENTES TRAS DAÑO CEREBRAL CAUSADO POR TRAUMA

Alumno:
Abel de Andrés Gómez

Director:
Antonio LaTorre

ESQUEMA DE TRABAJO

1. INTRODUCCIÓN
2. METODOLOGÍA
    1. ¿Qué es la ciencia de datos
    2. Etapas en el procesamiento de datos
    3. Trabajo relacionado
3. DESARROLLO
    1. Preparación de datos
    2. Pre-procesamiento de datos
    3. Modelado de datos
4. RESULTADOS
    1. Métricas de precisión y Kappa
    2. Matriz de confusión
    3. Curva de ROC
    4. Comparativa de tiempos
5. CONCLUSIONES
6. LÍNEAS FUTURAS

INTRODUCCIÓN

INTRODUCCIÓN

¿Qué son las lesiones traumáticas cerebrales?

- Son lesiones producidas cuando un golpe, impacto, sacudida u otras lesiones en la cabeza causan daños al cerebro.

- Sus **causas principales** son: accidentes vehiculares, caídas, actos de violencia y lesiones deportivas.

Datos interesantes:

- 2 millones de personas sufren LTC al año.
- Aproximadamente 52000 fallecidos al año.
- Mayor índice entre varones de 15 y 24 años.

Consecuencias de un LTC depende de:

- la rapidez del diagnostico y
- el tratamiento adecuado, que pueda aliviar algunas consecuencias.

Es complicado conocer las consecuencias de una LTC en las primeras horas e incluso en los primeros meses. ¡Es importante un estudio de predicción!

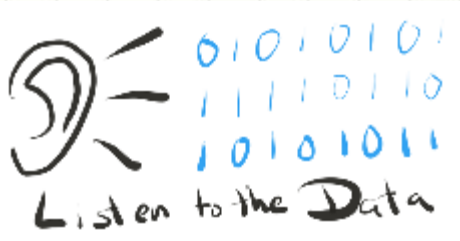
METODOLOGÍA

- ¿Qué es la ciencia de datos?
- Etapas en el procesamiento de datos
- Trabajos relacionados

METODOLOGÍA

¿Qué es la ciencia de datos?

- Campo de la estadística y de las ciencias de la computación.
- Combina los conceptos de estadísticas, minería de datos, análisis de datos y aprendizaje automático.
- Intenta descubrir información y patrones ocultos en grandes volúmenes de datos.
- Los patrones:
  - Muestran relaciones entre las variables
  - Ayudan a interpretar los datos
  - Aportan Información valiosa para la toma de decisiones.



METODOLOGÍA

Etapas en el procesamiento de datos

- Obtención de datos.
- Procesado y limpieza de datos.
- Análisis exploratorio de datos.
- Modelado.
- Producto y visualización de datos.

METODOLOGÍA

Trabajos relacionados

- Articulo de investigación como referencia.
- Muestra variables y número de pacientes a analizar.
- Regresión logística como modelo de predicción.
- Stepwise como técnica de selección de variables.
- Curva ROC y métricas de precisión.
- Redes Bayesianas para la probabilidad de asociación.

DESARROLLO

- Preparación de datos
- Pre-procesamiento de datos
- Modelado de datos

Desarrollo - Preparación de datos

Obtención de los datos

- Los datos provienen de la institución de: “London School of Hygiene and Tropical Medicine”.
- Se encuentran en formato CSV separado por comas.
- Contiene 88 variables, de las cuales solo utilizaremos 31 para la preparación
- Las variables de escáneres y Test de Glasgow se duplican para el hospital de origen y el posible hospital de transferencia.
- Se poseen 10008 pacientes.
- El numero final de variables será de 14.

Desarrollo - Preparación de datos

Clasificación entre: Alive, DEATH/SD, NO-DATA y MD/GR

- Clasificación:
  - Fallecidos o con discapacidades severas.
  - Con discapacidad moderada o buena recuperación.
  - Vivos (pero sin resultados finales).
  - Sin datos.
- 2 variables (GOS5 y GOS8) -> poseen el estado final del paciente.
  - Dichas variables suelen poseer valores de forma alterna.
- Se crea una nueva variable llamada “outcome” que será la que contendrá el resultado final.
- La variable “outcome” poseerá el valor de las variables GOS5 o GOS8.
  - Si estas variables no poseen ningún valor, entonces deberemos tener en cuenta el resto de valores. Por ejemplo, la variable “SYMPTOMS” y la variable “OUTCOME”.
  - La unión de “SYMPTOMS” y “OUTCOME” también nos ha aportado información sobre el estado final del paciente cuando GOS5 y GOS8 se encuentran vacías.

Desarrollo - Preparación de datos

Clasificación entre: Escaneados y no escaneados

- Variables de escáneres duplicadas (Hospital origen y transferencia).
- Duplicidad + Columnas sin valor -> Clasificación.
- **Objetivo**: eliminar filas que no tengan valores en ningún escáner.
- Conservaremos únicamente las que posean valores (escaneados).
- Sobre los escaneados, usamos los escáneres del hospital de transferencia (si los posee).
- **¡Todas las columnas deben estar cumplimentadas!.**

Desarrollo - Preparación de datos

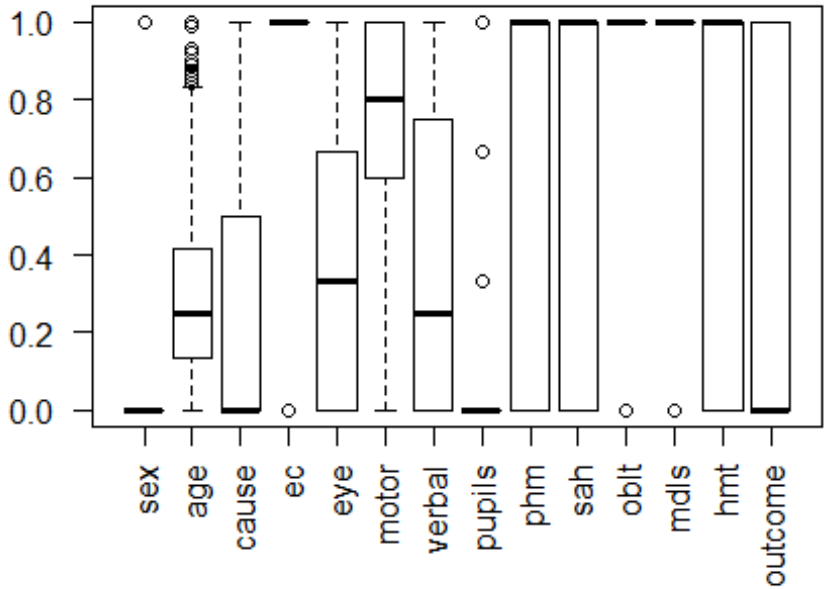
Eliminación y unión de variables

- 2 columnas de reactividad de pupilas: derecha e izquierda.
  - Nueva columna para unir las dos pupilas.
  - Posterior al tratamiento, se eliminan las 2 variables originales.
- 2 columnas de causa de la lesión: Hospital origen y transferencia
  - Generalmente poseen el mismo valor.
  - Si difieren, entonces nos quedamos con la actual (hosp. Transferencia).
  - Nueva columna para unir las dos causas.
  - Posterior al tratamiento, se eliminan las 2 variables originales.
- Se cambia el nombre a todas las variables.
- Se revisa y se eliminan las filas que no posean todos los valores.
- Se revisa y se modifica algunos valores anómalos. p.ej: Escaneado pero sin datos de escáner.
- 6986 pacientes.

Desarrollo - Pre-procesamiento de datos

Búsqueda de outliers

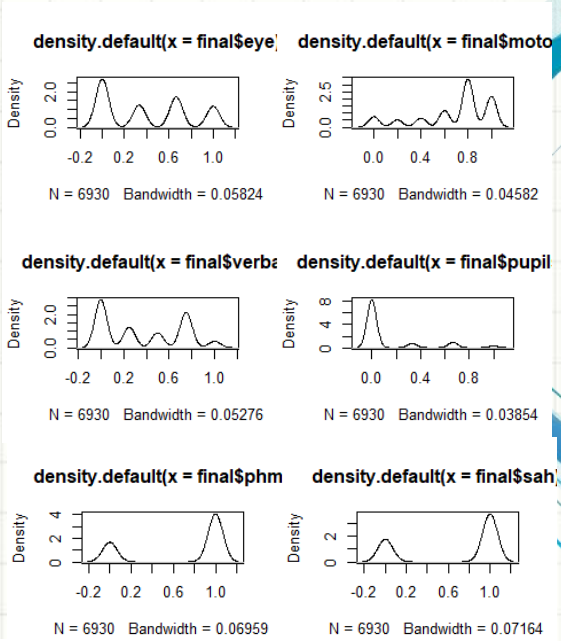
- Estadísticamente son aquellos que se encuentran fuera del rango intercuartil (Q1-Q3).
- Antes de descartarlos, es necesario explicar su presencia.
- En la naturaleza estos datos no son tan anómalos. **Son totalmente posibles.**
- No se descarta ningún dato.

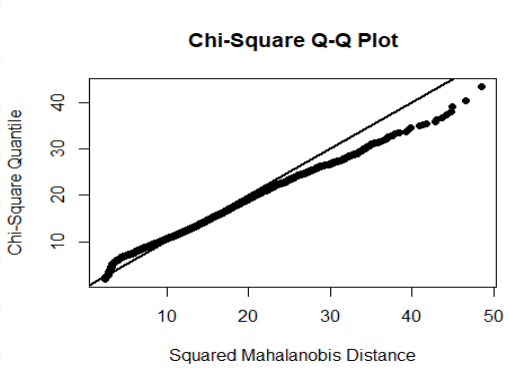


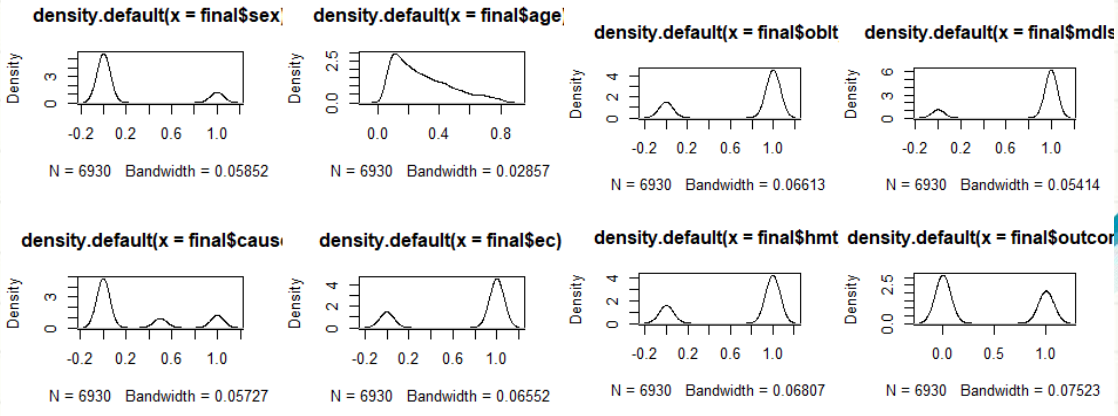
Desarrollo - Pre-procesamiento de datos

Análisis de normalidad

- Se ha realizado el test de Mardia, Henze-Zirkler y Anderson-Darling.
- Todos los test han **rechazado la hipótesis de normalidad.**
- Deberemos tener en cuenta este análisis en el uso de **modelos predictivos no-paramétricos**.





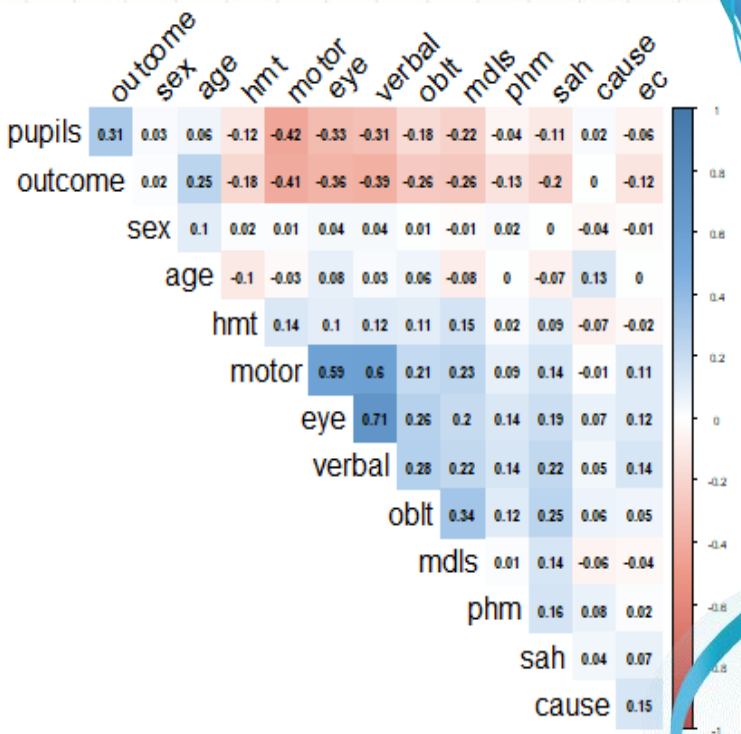


Desarrollo - Pre-procesamiento de datos

Estudio de la correlación

- No existen correlaciones suficientemente fuertes.
- No contienen información redundante.
- “motor”, “verbal” y “eye” son variables con gran relevancia.
- Las variables de “cause” y “sex” apenas tienen correlación con la variable de salida.
- Se deberá estudiar la posible exclusión de estas variables.

	First.Variable	Second.Variable	Correlation
188	motor	outcome	-0.40984112
189	verbal	outcome	-0.38934889
187	eye	outcome	-0.36182805
190	pupils	outcome	0.30999486
194	mdls	outcome	-0.26340926
193	oblt	outcome	-0.25708526
184	age	outcome	0.24528254
192	sah	outcome	-0.20016386
195	hmt	outcome	-0.18246423
191	phm	outcome	-0.13254914
186	ec	outcome	-0.12180868
183	sex	outcome	0.01988148
185	cause	outcome	0.00362152
196	outcome	outcome	0.00000000



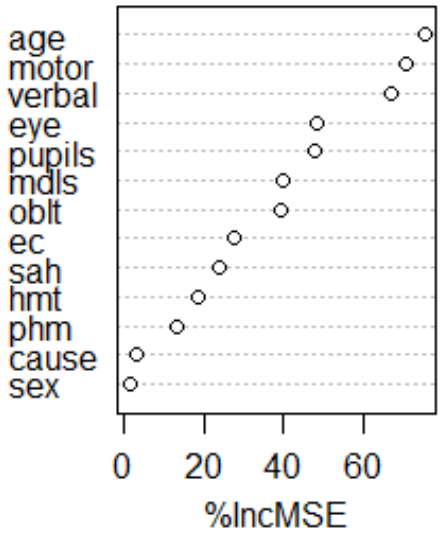
Desarrollo - Pre-procesamiento de datos

Selección de variables:

Random Forest

- Entrenamos los datos

- Obtenemos sus variables mas importantes usando IncMSE



Stepwise Fordward

- Se ha usado el criterio AIC.

- Se ha usado el criterio lambda de Wilks.
  - Utiliza la variable que mejor clasifica a la clase y va incluyendo variables.

```
Initial Model:
outcome ~ sex + age + cause + ec + eye + motor + verbal + pupils +
    phm + sah + oblt + mdls + hmt

Final Model:
outcome ~ age + ec + eye + motor + verbal + pupils + phm + sah +
    oblt + mdls + hmt
```

- Ambos modelos coinciden en que el mejor modelo es aquel que tiene en cuenta todas las variables excepto “sex” y “cause”.

```
Formula containing included variables:

outcome ~ motor + age + verbal + oblt + pupils + mdls + hmt +

    eye + ec + phm + sah
```

Desarrollo - Pre-procesamiento de datos

Análisis PCA

Para elegir el numero de componentes principales podremos utilizar dos criterios:

- Criterio de Kaiser.
  - Conservar aquellos cuya **desviación estándar al cuadrado** alcancen 1.
  - Con este criterio, nos quedaríamos con los primeros 5 componentes.

```
##  [1] 0.49713279 0.22143722 0.20570161 0.18278879 0.17178962 0.15685861
##  [7] 0.15239376 0.13879386 0.12731397 0.08999118 0.05458915 0.04233462
## [13] 0.03867736 0.03481073
```

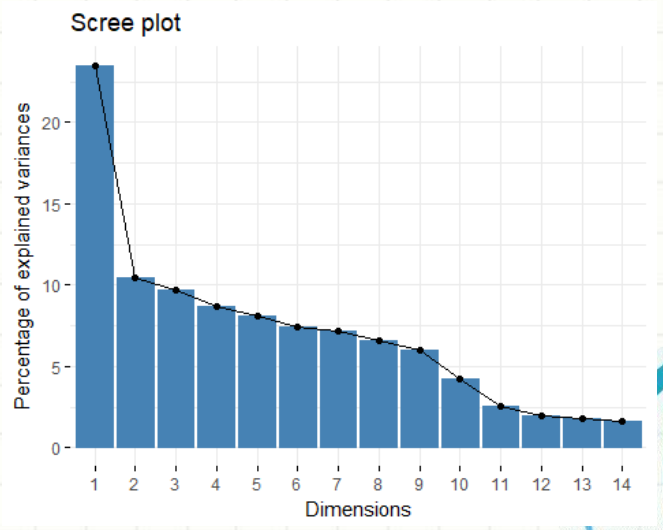
- Explicar al menos un 80% de la varianza.
  - Según este criterio deberíamos seleccionar los 9 primeros componentes principales.

¿Debemos descartar algún componente?

Los criterios nos han indicado que debemos usar 5 y 9 componentes.

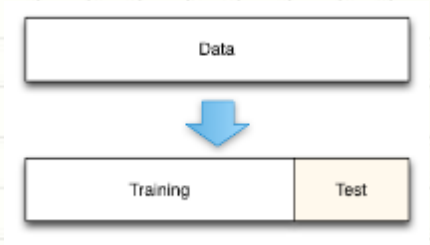
Teniendo en cuenta que tenemos 14 variables:

- La reducción no es significativa.
- Se perdería interpretabilidad.



Desarrollo - Modelado de datos

- En esta fase, se han dividido mediante muestreo los datos en un conjunto de entrenamiento -70%- y pruebas -30%-.
- Las técnicas utilizadas son las siguientes:
  - Arboles de decisión (Random Forest)
  - Regresión logística
  - Redes neuronales
  - Análisis bayesiano (Naïves Bayes)
  - Gradient Boosting (AdaBoost)
  - Gradient Boosting Trees (GBM y XGBoost)
  - Combinación de modelos.



- Utilizaremos modelos de clasificación, debido a que la variable a predecir es dicotómica.

- Proceso:
  - Construimos el modelo con los datos de entrenamiento.
  - Usamos el modelo construido y los datos de pruebas para obtener las predicciones.
  - Visualizamos las predicciones en forma de matriz de confusión.

- Se ha tenido en cuenta la eliminación de las variables de “sex” y “cause” en la construcción de modelos.

RESULTADOS

- Métricas de precisión y Kappa
- Matriz de confusión
- Curva de ROC
- Comparativa de tiempos

Resultados - Métricas de precisión y Kappa

Precisión

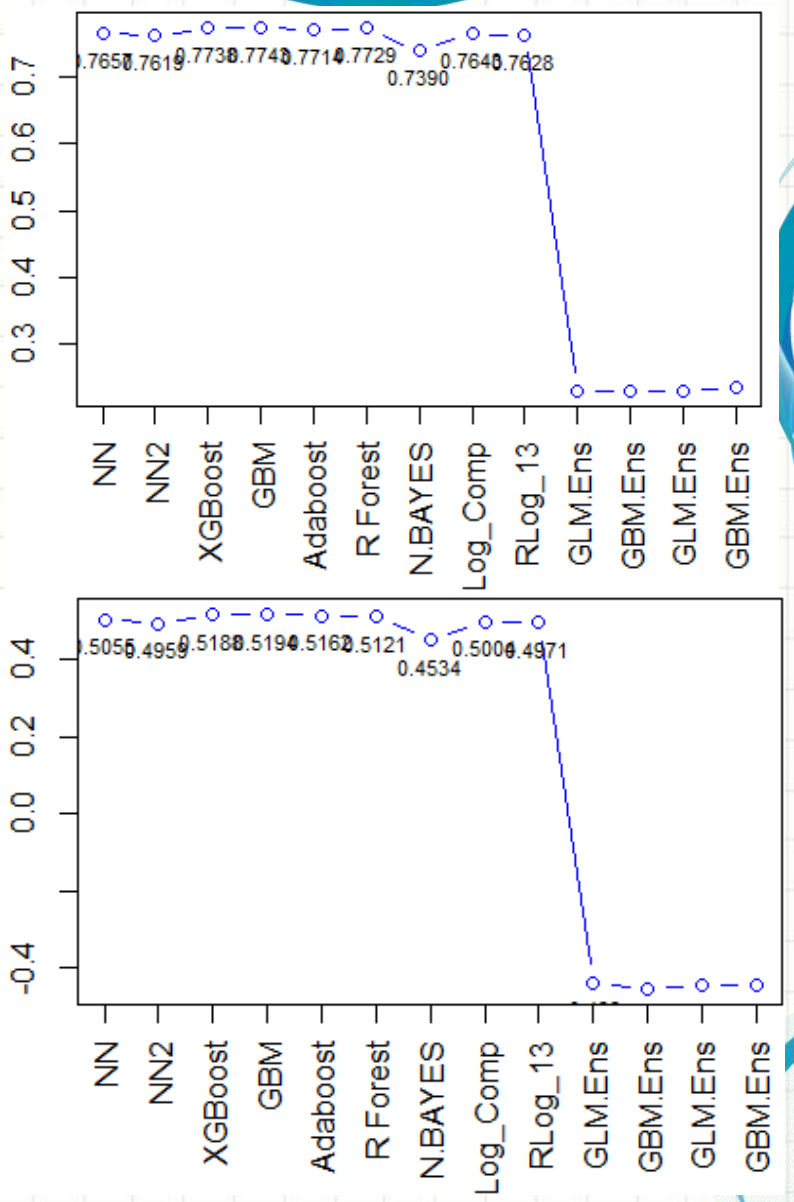
Porcentaje de pacientes correctamente clasificados.

Teniendo en cuenta nuestro conjunto de datos:

- GBM ha sido el modelo que mejor precisión nos ha aportado
- La combinación de modelos ha sido la que peor precisión nos ha proporcionado.

Kappa

- Podemos observar que los resultados son similares a la precisión.



Resultados – Matriz de confusión

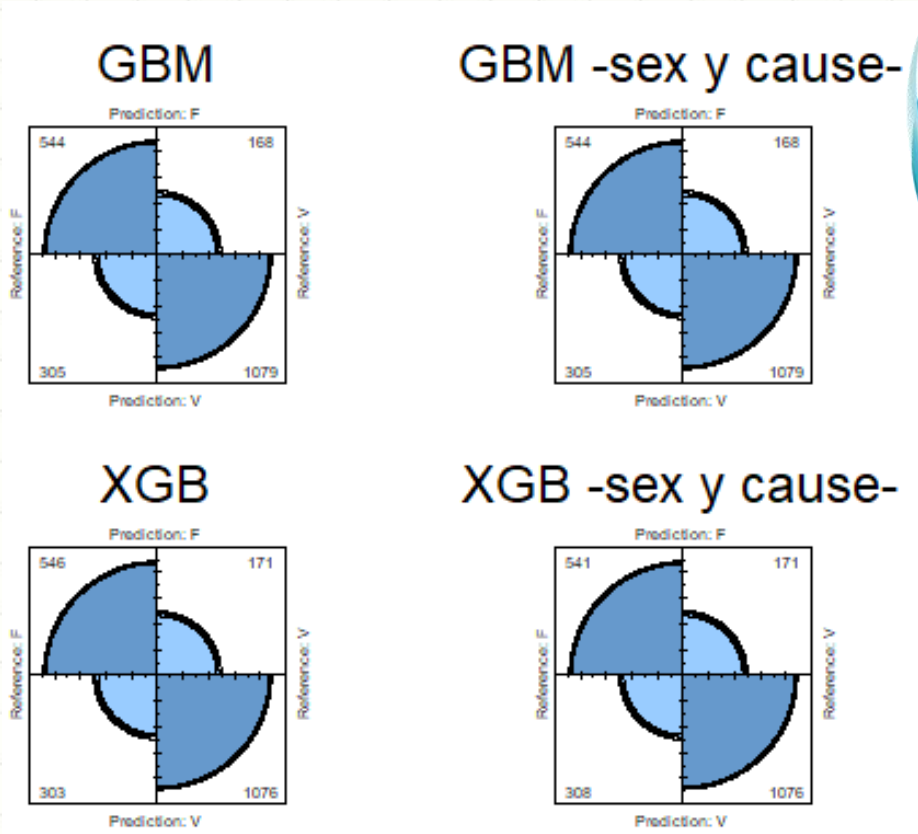
Tabla que se usa para medir el rendimiento de un modelo de clasificación.

Los términos mas básicos son:

- Verdaderos Positivos.
- Verdaderos Negativos.
- Falsos Positivos
- Falsos Negativos

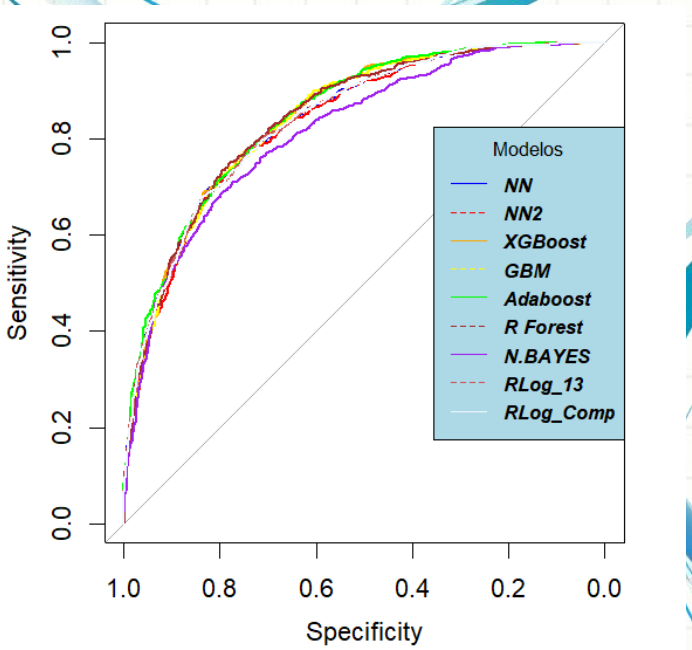
EN GBM:

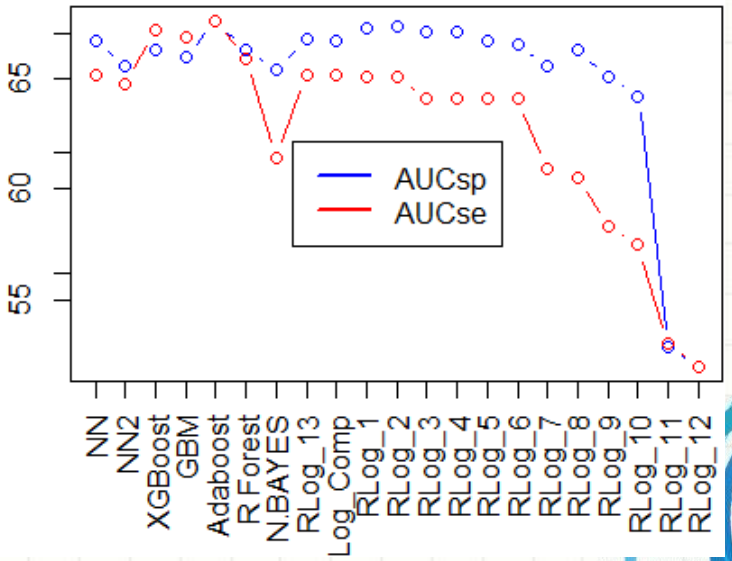
- VP= 544
- VN=1079
- FP=168
- FN=305



Resultados – Curva de ROC

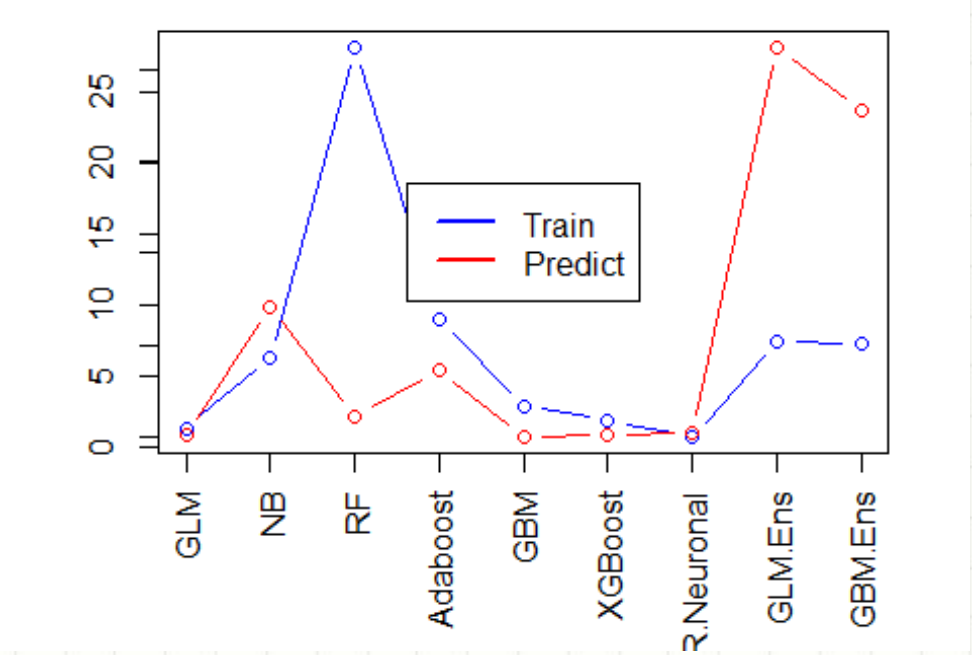
- ROC muestra la capacidad discriminativa (distinguir entre fallecidos y vivos) de un modelo.
- Calculamos AUC
  - Probabilidad de clasificar correctamente un par de individuos vivo y fallecido, seleccionados al azar de la población.
- Cuanto mas se acerque el área a 1, mayor capacidad de discriminación del modelo.
- Medidas
  - Sensibilidad-> proporción de verdaderos positivos sobre el total que fallecen.
  - Especificidad-> proporción de verdaderos negativos sobre el total que viven.
- Calculo del área parcial para el rango 90%-100% de 'se' y 'sp'. Rango especifico para determinada situación clínica. Es un valor estandarizado.





Resultados – Comparativa de tiempos

- Además de las métricas predictivas y la capacidad discriminatoria, hemos utilizado una comparación entre tiempos.
- Se han medido **los tiempos de entrenamiento** (sin tener en cuenta la búsqueda de híper-parámetros).
- Se han medido **los tiempos de predicción** usando el conjunto de pruebas y el modelo construido.



CONCLUSIONES

Conclusiones

- Determinaciones en análisis exploratorio concuerdan con resultados en fase de modelado posteriores.
- Tener en cuenta varios modelos y no quedarse con uno solo. (Precisión, tiempo, pruebas de diagnostico).
- Necesidad de estudio detallado de la combinación de modelos. Hemos obtenido peores resultados que utilizando modelos independientes.
- Se han utilizado un mayor numero de modelos y se han obtenido para las configuraciones seleccionadas, mejores resultados.

LÍNEAS FUTURAS

Líneas futuras

- Es evidente que se puede realizar nuevos trabajos con el objetivo de obtener **nuevos resultados**.
- Utilizar **nuevas variables**, realizar un análisis exploratorio y establecer **nuevos modelos**.
- Utilizar otras configuraciones de **híper-parámetros** al construir los modelos.
- Realizar nuevos estudios centrados en la **combinación de modelos**. Seleccionar otros modelos para obtener mejores resultados.
  - Modelos que estén menos correlacionados entre sí.

MUCHAS GRACIAS POR SU ATENCIÓN

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

ABEL DE ANDRÉS GÓMEZ