**SELECCIÓN DE VARIABLES Y ANÁLISIS PREDICTIVO**

Elaborado por:

DIANA CATALINA VELÁSQUEZ GAVIRIA

CAMILA MEJÍA QUINTERO

MAURICIO CUSCAGUA

Materia: Métodos estadísticos avanzados

Profesora: ANDRÉS RAMIREZ HASAN

Maestría en ciencia de datos y analítica

Universidad Eafit

Medellín

2019

## Introducción

En el presente trabajo se analizarán tres bases de datos, las cuales corresponden a variables dependientes con soportes continuo, binario y conteo. El objetivo sobre cada una de ellas será seleccionar las variables regresoras y predecir la variable respuesta. Para medir el desempeño de los modelos, la capacidad predictiva general será el error cuadrático medio para las variables continuas y conteo, y el accuracy (verdaderos positivos + verdaderos negativos dividido el taman ̃o muestral) para la variable binaria. La capacidad predictiva específica, será la correcta clasificación de los valores inferiores y superiores a -1 para la variable continua. Para la variable binaria, será el ́area bajo la curva ROC. En el caso de la variable de conteo, ser ́a la correcta clasificación de los valores iguales a 0, y mayores a 0.

Para llevar a cabo el análisis se utilizará la metodología CRISP-DM ( Cross Industry Standard Process for Data Mining).

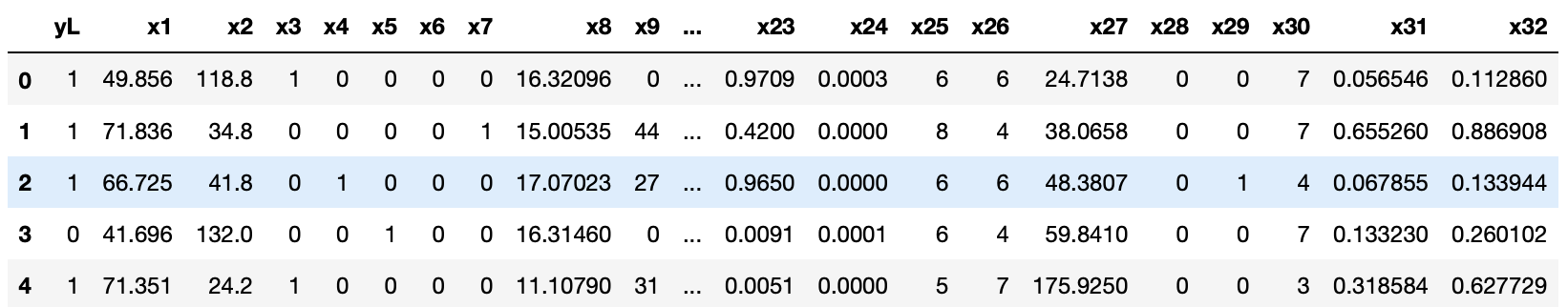
## Fase I. Business Understanding. Definición de necesidades del cliente (comprensión del negocio)

Para este ejercicio no se tiene detalle de las bases de datos suministradas, sin embargo, el objetivo será obtener el mejor modelo predictivo, con una previa selección de las variables regresoras que mejor expliquen la variable respuesta.

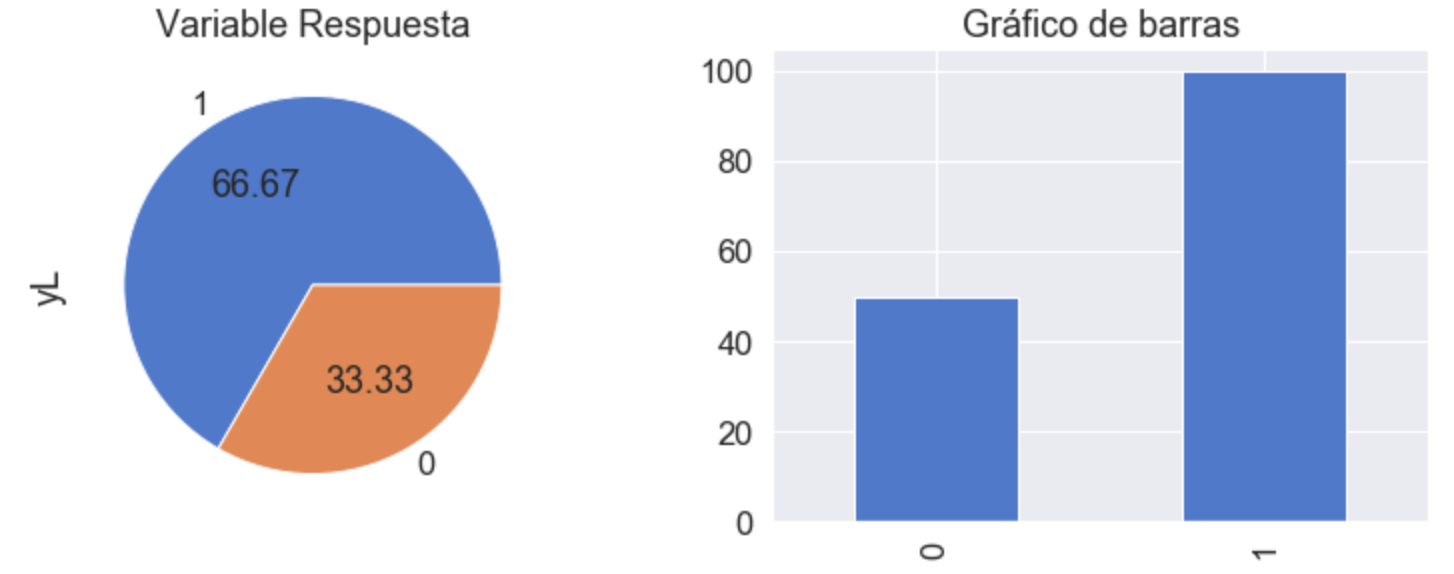
## Fase II. Data Understanding. Estudio y comprensión de los datos

## Para llevar a cabo el análisis sobre la base objeto de estudio (en este caso la binaria) se realizó primero que todo un entendimiento general de los datos. Para esto se realizó una visualización inicial y un resumen general de las variables de la base:

Como se mencionó al inicio se tiene un total de 150 registros y 32 variables.



La variable de interés (variable respuesta) es “yL” la cual toma valores de 0 y 1 y está conformada así:



El 33.33% de los datos pertenecen al nivel 0, lo cual corresponde a 50 observaciones y el 66.67% permanecen al nivel 1 y equivalente a 100 observaciones.

**Análisis de variables numéricas:**

Posteriormente se realizó un análisis de cada una de las variables de la base con respecto a la variable de interés (yL). Para esto se dividió la base de datos en dos subset. Uno para variables numéricas (17 variables) y otro para variables dicótomas (15 variables).

Para el conjunto de variables numéricas se analizó el promedio de cada variable para cada una de las categorías de la variable respuesta así:

Imagen que contiene captura de pantalla

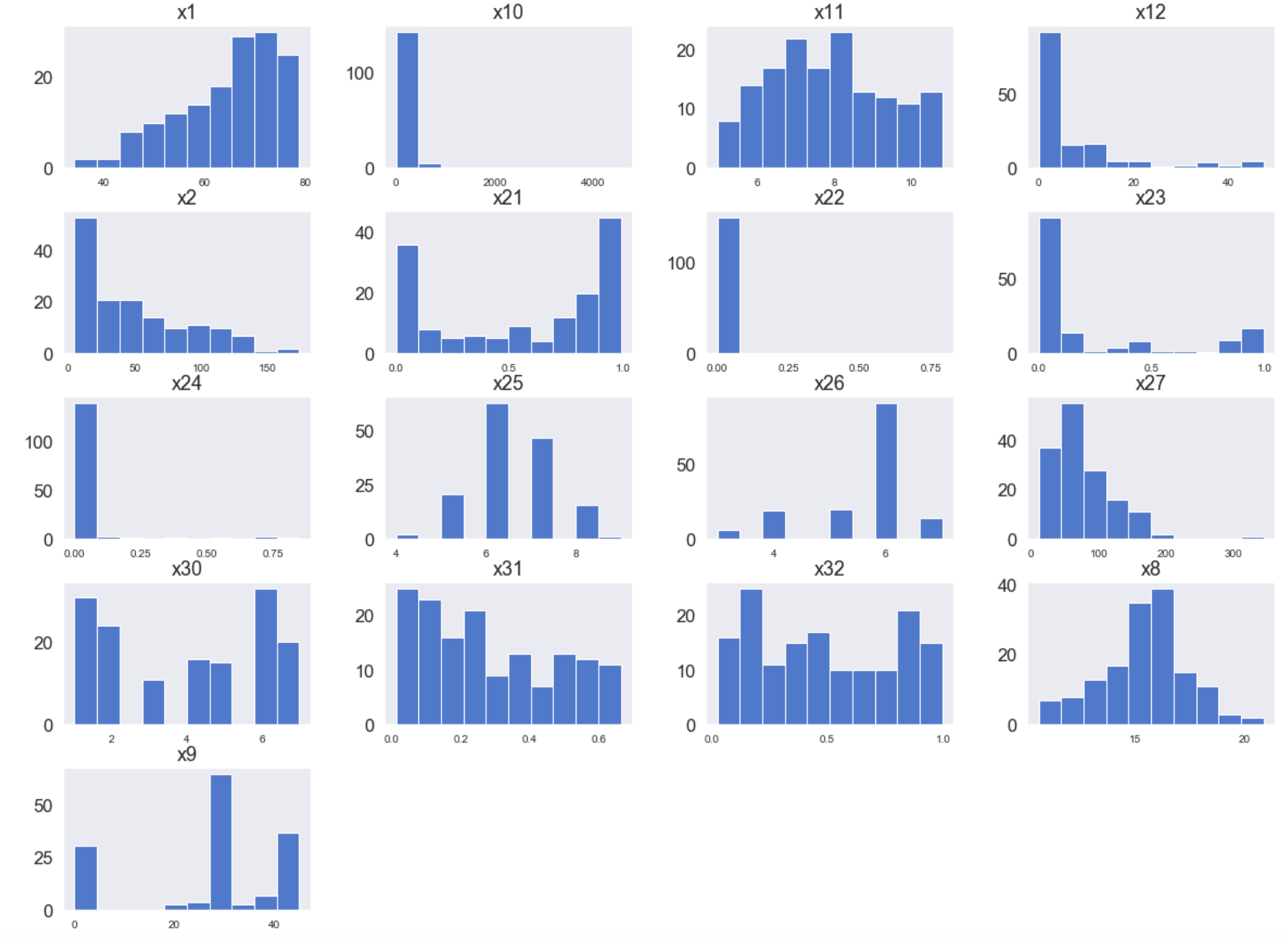
Descripción generada automáticamente

Con este análisis se buscó entender como variaba el promedio tanto para ambos niveles de la variable respuesta e identicar las variables donde se presentaba la mayor diferencia a simple vista, los cuales se encuentran resaltados en círculo naranja.

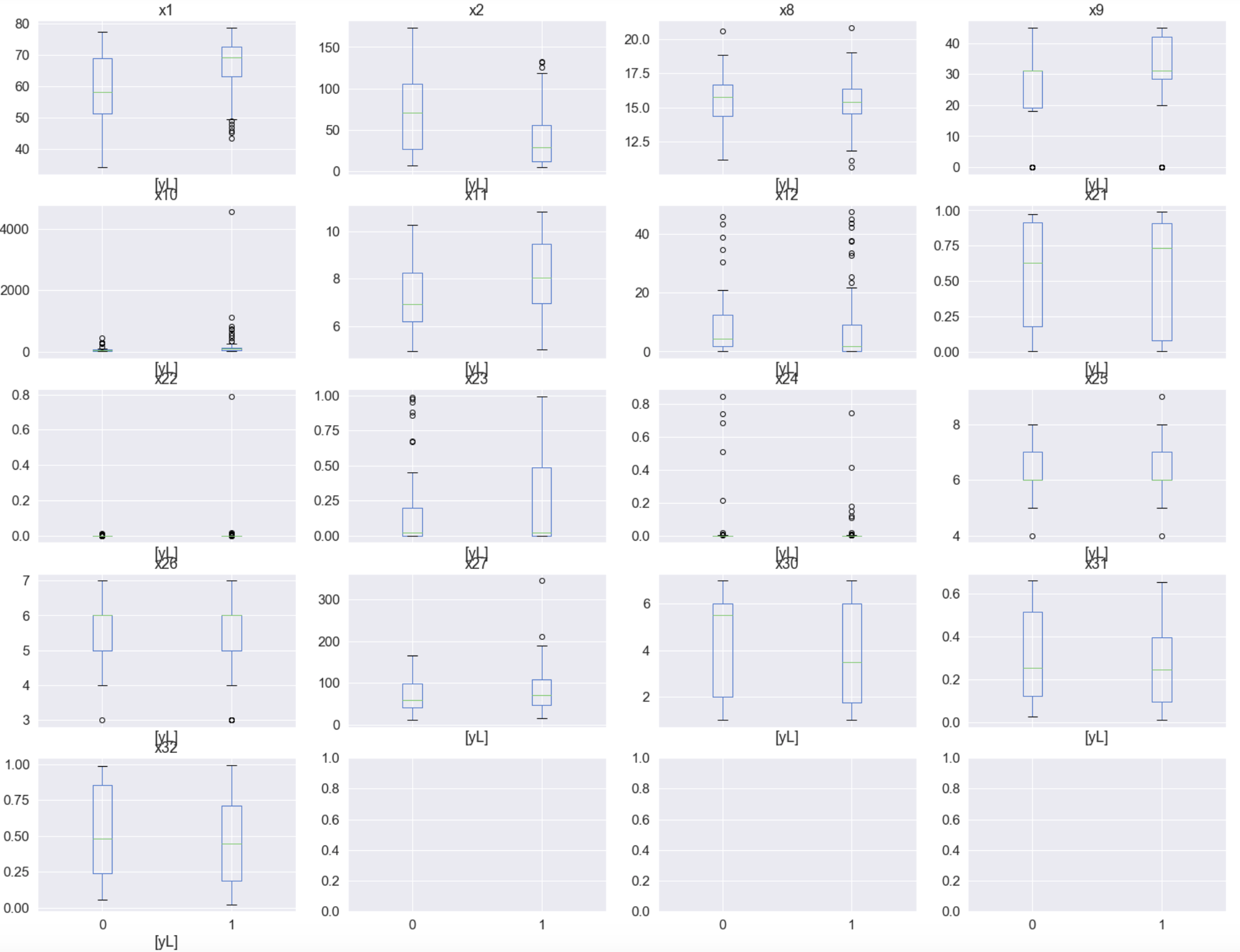
Para complementar el entendimiento se realizaron histogramas de cada una de las variables numéricas y boxplot comparativos que permitieran ver la distribución y variabilidad de los datos tanto de manera univariada como con respecto a cada categoría de la variable respuestra. Esto permite dar claridad del rango de valores en los que se mueven cada una de las variables. (ver notebook adjunto).

En este análisis se pudo detectar que hay tres variables con muy pocos niveles, estás son: x25, x26, x30. Las demás regresoras variaban en un rango amplio de datos como se puede ver en los gráficos siguientes:

**Histogramas variables numéricas**



**Box plot comparativos variables numéricas vs ‘yL’**



En los boxplot se pueden visualizar de mejor manera aquellas variables donde pareciera haber diferencia entre ambas categorías de ‘yL’, lo cual da una idea inicial de que variables pueden ser de interés para explicar la variable respuesta.

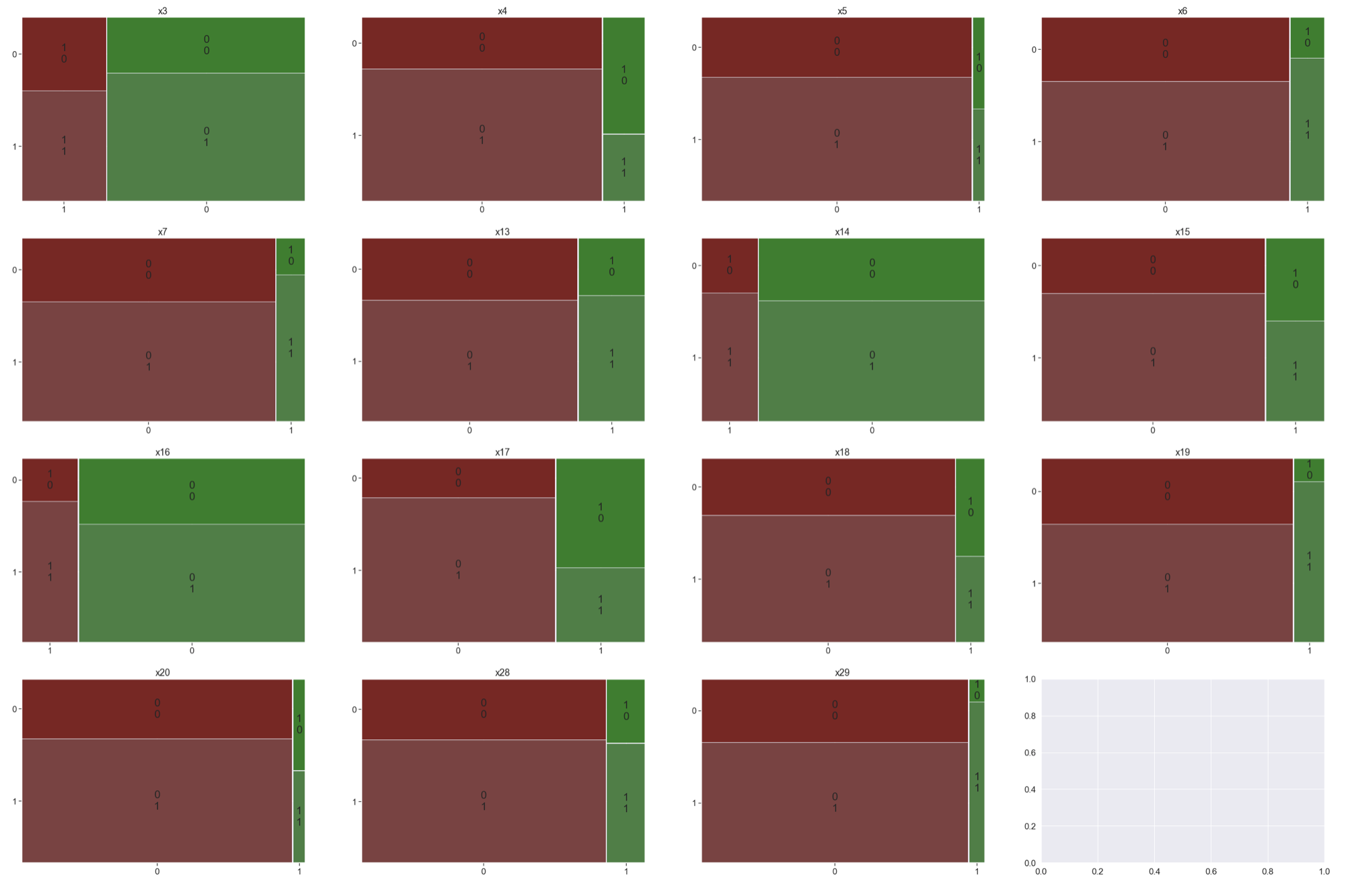
También para este tipo de variables se realizó una matriz de correlación y se encontró que las variables con mayor coeficiente de correlación son las resaltadas en azul oscuro para las correlaciones negativas y naranja fuerte para las correlaciones positivas. (ver notebook con tabla completa). Por ejemplo las variables de correlación mayor a 0.7 fueron: (x1 y x2),(x1 y x11) , (x31 y x32), (x21 y x23).



**Análisis de variables categóricas**

De la misma manera se realizó un análisis sobre las variables categóricas con respecto a la variable “yL” para entender como variaban las proporciones de cada categoría entre los niveles de la variable respuesta. Para llevar a cabo este análisis se construyeron mosaics plot y tablas de contingencia. Y de la misma manera se identificaron aquellas variables donde dichas proporciones parecía diferentes entre ambos niveles de la variable respuesta.

**Mosaics Plots variables categóricas**

Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba Señal de pulgar hacia arriba 

Se resaltan con una manito amarilla aquellas variables donde parece tenerse diferencia en los niveles. Las tablas de contingencia pueden verse en el notebook adjunto.

## Fase III. Data Preparation. Análisis de los datos y selección de características.

Luego de realizar el entendimiento de los datos y tratar de detectar patrones, correlaciones, etc. Se procedió a preparar los datos para la selección de caraterísticas y el proceso de modelado. Lo primero que se hizo fue una estandarización de las varibles con el fin de evitar problemas de escalas. Posteriormente se realizó una partición de la base de datos en dos subconjuntos: Train y test (80% y 20% respectivamente).

**Partición de los datos:**

Para llevar a cabo la partición y garantizar una real aleatoriedad se realizó primero una mezcla de los datos con la función shufle de python y luego sobre esta base de datos mezclada se realizó la partición aleatoria de la misma en las proporciones definidas. De esta manera se obtuvo un conjunto de train de 120 registros, y uno para test de 30 regristros..

**Selección de variables (feature selection)**

Antes de realizar todo el proceso de modelación, se dedicó un buen tiempo a la selección de variables. Para esto se utilizaron diversas técnicas, tales como: filter methods (varianza constante, cuasi-constante, características duplicadas, correlación, mutual information, fisher score, univariate AUC score), wrapper methods (step fordward, step backward, subset) y embedding methods (lasso regularisation, trees, random forest importance, gradient. Boosted machines, se usó una técnica de bayes (spike and slab) y algoritmos evolutivos, especificamente una técnica de differential evolution con un naive bayes como función de optimización ).

Con los primeros métodos se encontró que no se contaba con carácterísticas duplicadas, ni con variables con varianza constante o casi contante, por lo tanto con estos métodos básicos no se eliminó ninguna característica. Adicional a todos estos métodos, se utilizó un criterio propio con base en lo observado en el análisis descriptivo para la preselección de algunas características.

Al utilizar los wrapper methods, se realizó validación cruzada con kfold = 5 para tomar las variables donde se maximizaba el resultado del AUC en el dataset de validación. De la misma manera se realizaron varias particiones aleatorias de train y test para hacer más robusto el proceso de selección y todo este proceso se corrió sobre el dataset de entranamiento, lo cual es lo recomendado para evitar el sobreajuste.

Estos fueron los métodos utilizados:

* **Descriptivo/fisher para varibles dicótomas**
* **Mutual Information**
* **AUC univariate score**
* **Step forward**: Se iteró para diferente número de características y se tomó el de mayor AUC. Además se realizó validación cruzada con k=5
* **Step backward:** Se iteró para diferente número de características y se tomó el de mayor AUC. Además se realizó validación cruzada con k=5
* **Subset** cv=2: se realizó para un total de hasta 12 combinaciones de características con validación cruzada con k = 2. (Esto por el costo computacional que presenta).
* **Laso Regularisation** (c=1): se aplicó la regularización lasso con varios criteriores de regularización. (c = 0.5, c = 1 y c=1.5) mientras más alto el valor del hiperparámetro de regularización, menos estricta es la penalidad. Se tomaron las características de los tres modelos.
* **Elasticnet regularisation**
* **Random Forest importance**
* **Recursive feature selection using random forests importance RFE:** se remueve una característica en cada iteración. La menos importante.
* **Recursive feature selection using random forests importance RFECV:** Es la misma técnica anterior pero con validación cruzada. Se tomó k=5.
* **Gradient Boosted trees importance**
* **Spike and slab**

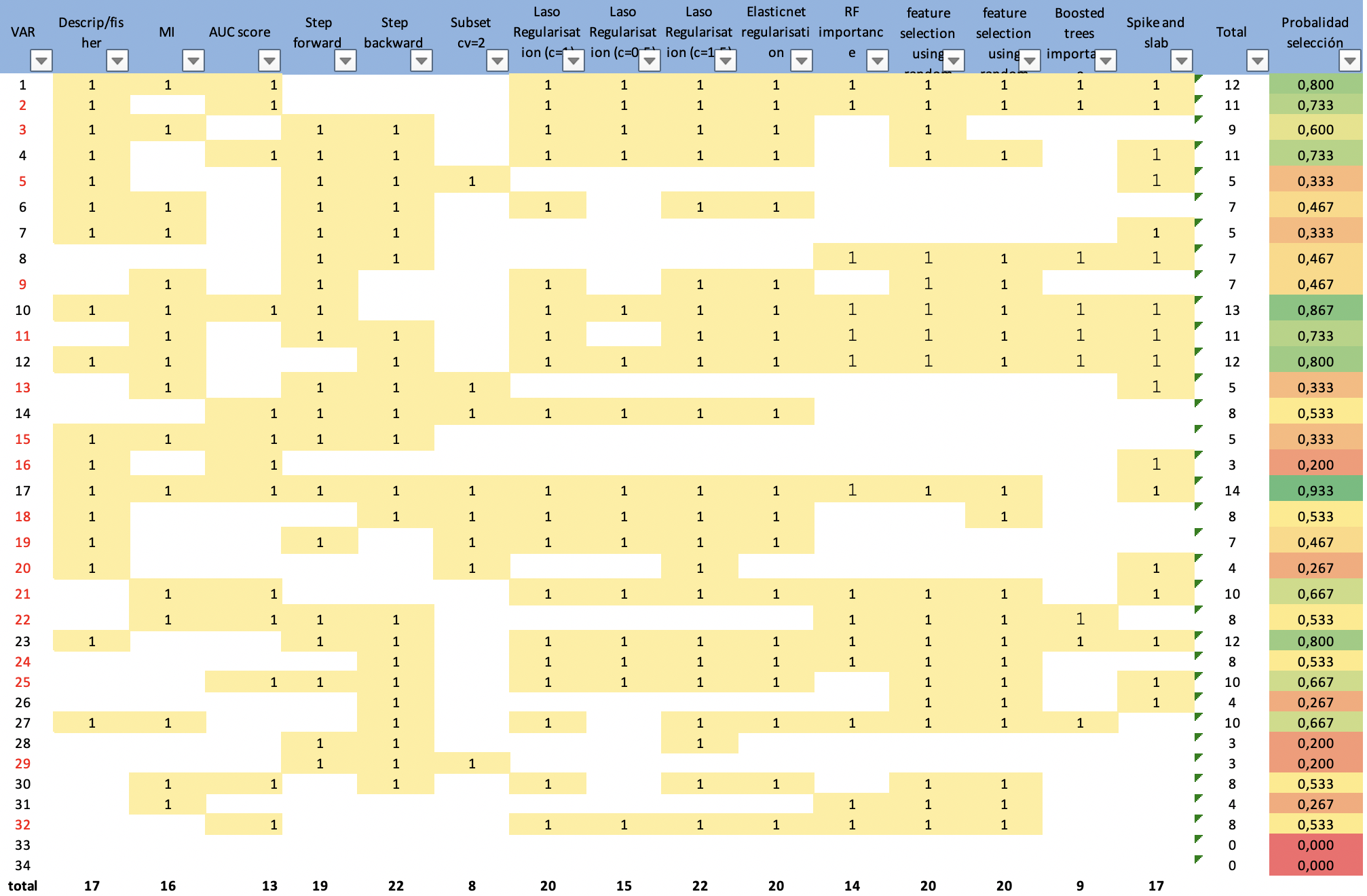
Algunos de los resultados del AUC de validación en algunos métodos son los siguientes:

Imagen que contiene texto

Descripción generada automáticamente

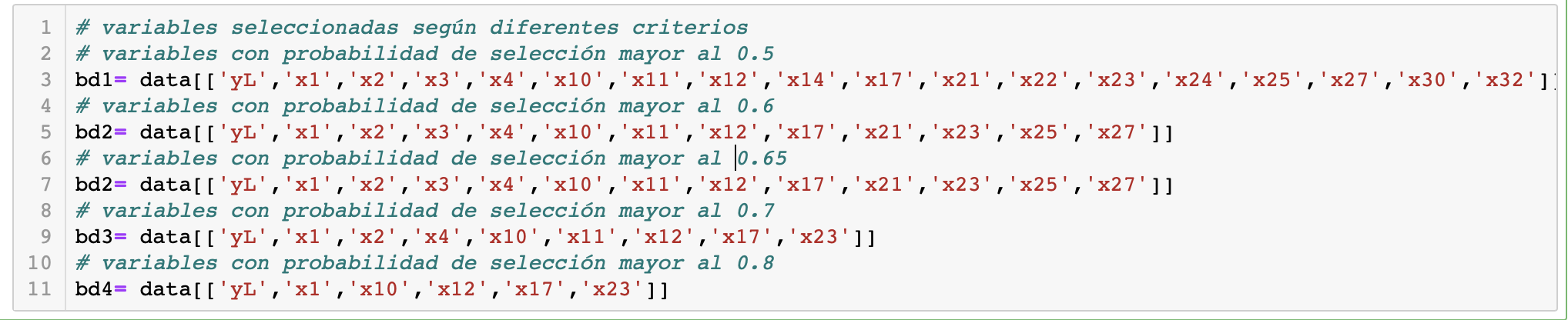
De acuerdo a los resultados de la tabla, se seleccionaron la cantidad de variables donde se lograba el mayor resultado de AUC. (resaltados en el recuadro naranja).

Finalmente luego de aplicar todos los métodos con sus múltiples corridas se creó una tabla con el listado de variables de la base original y la cantidad de métodos usados de la siguiente manera:



Cada columna representa un método de selección. La primera coumna contiene todas las variables y se asigna un valor de 1 en cada método si la variable fue seleccionada por el mismo. Al final se crearon dos medidas, una representa el número total de veces que la variable fue seleccionada por los diferentes métodos y la última columna representa la probabilidad de selección de cada variable (# de veces que fue seleccionada/métodos usados)

Con base en la tabla obtenida se utilizaron varios criterios para la selección de características, el primero fue tomar las variables que tuvieron una probabilidad de selección mayor a 0.50, el segundo fue tomar las variables con probabilidad de selección mayor a 0.5, 0.6, luego mayor a 0.65, a 0.70 y a 0.8 y este fue el resutado de la selección:



Con base en lo anterior se construyeron cinco subconjuntos de bases adicionales, cada una con las variables seleccionadas por cada criterio de selección. La bd1 tiene un total de 17 variables, la bd2 de 12 variables, la bd3 de 11, la bd4 de 8 y la bd5 de 5 variables.

## Fase IV. Modeling. Modelado

Una vez terminados los procesos de preparación y exploración de los datos y la selección de carácterísticas se realizó el proceso de modelado. Se utilizaron 6 modelos de aprendizaje sobre los 5 subset mencionados anteriormente: Regresión logística, Gradient boosting, Decission Tree, Random Forest y suport vector machines y k nearest neighbors. Para evaluar el rendimiento de los modelos se analizó el ROC AUC en el subconjunto de test y también se analizó el accuracy.

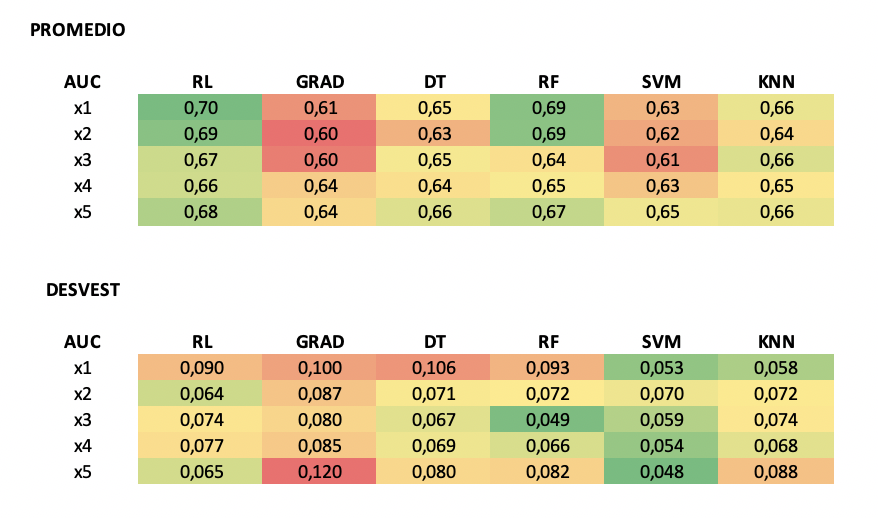
Adicionalmente para evitar sobreajuste se realizaron varias particiones aleatorias sobre la base de datos original (como una especie de validación cruzada) y al final se promedió el resultado de los AUC sobre las 10 corridas realizadas (ver tabla adjunta en excel con todo el detalle). Adicionalmente durante el proceso de modelado se tuvieron en cuenta algunos aspectos tales como:

Al usar la regresión logística, teniendo en cuenta la diferencia presentada entre los AUC de entrenamiento y los de validación, se aplicó una regularización variando un parámetro en el código en diferentes valores y se seleccionó el criterio que maximizaba el resultado el AUC. Este varía para cada subconjunto de variables.

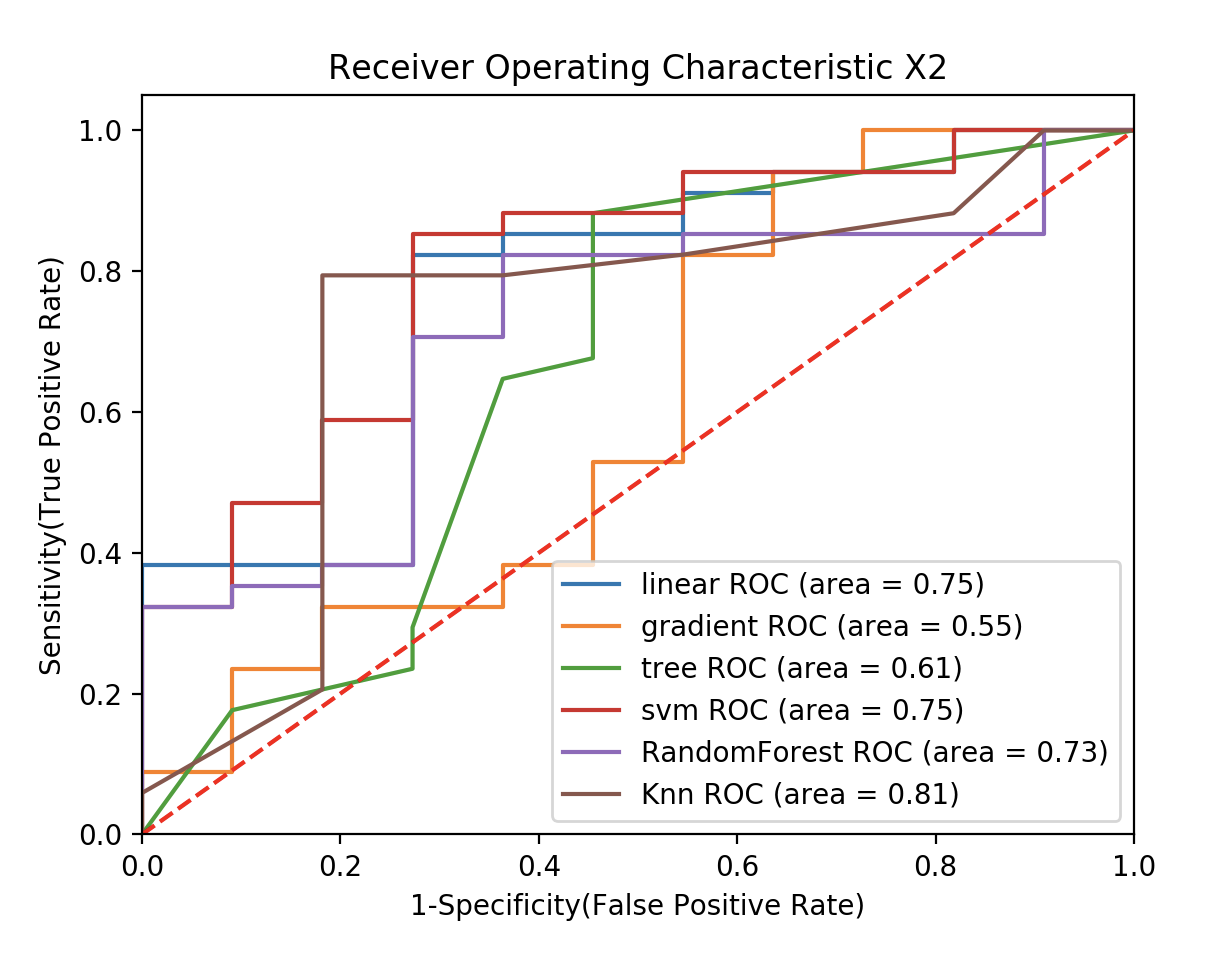
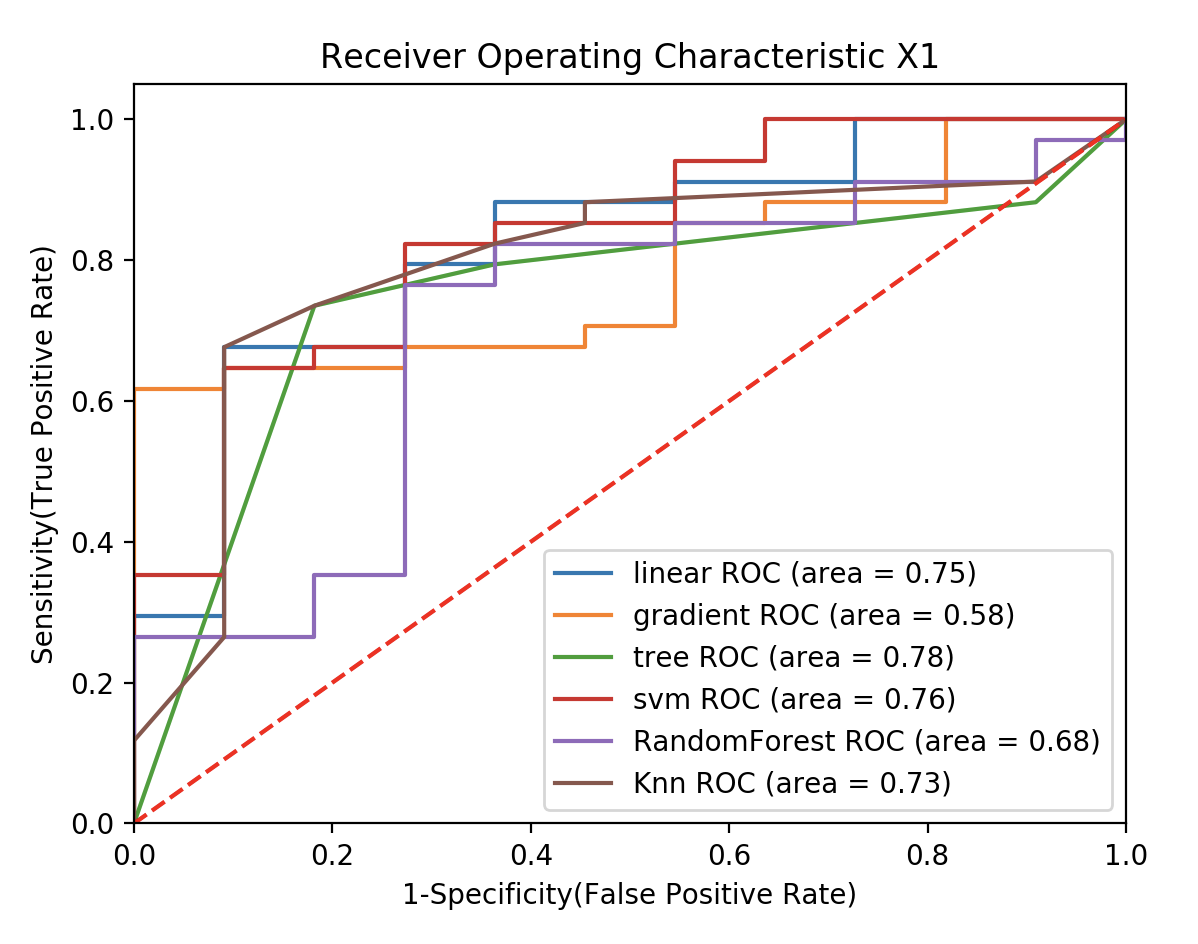
En las máquinas de soporte vectorial se realizó una variación a los kernel, se utilizaron los siguientes: Lineal, sigmoide, polinomial y radial. Teniendo en cuenta que es un modelo costoso computacionalmente y que los resultados con los diferentes kernel no fueron los mejores se decidió utilizar el kernel lineal que arrojó los mejores resultados.

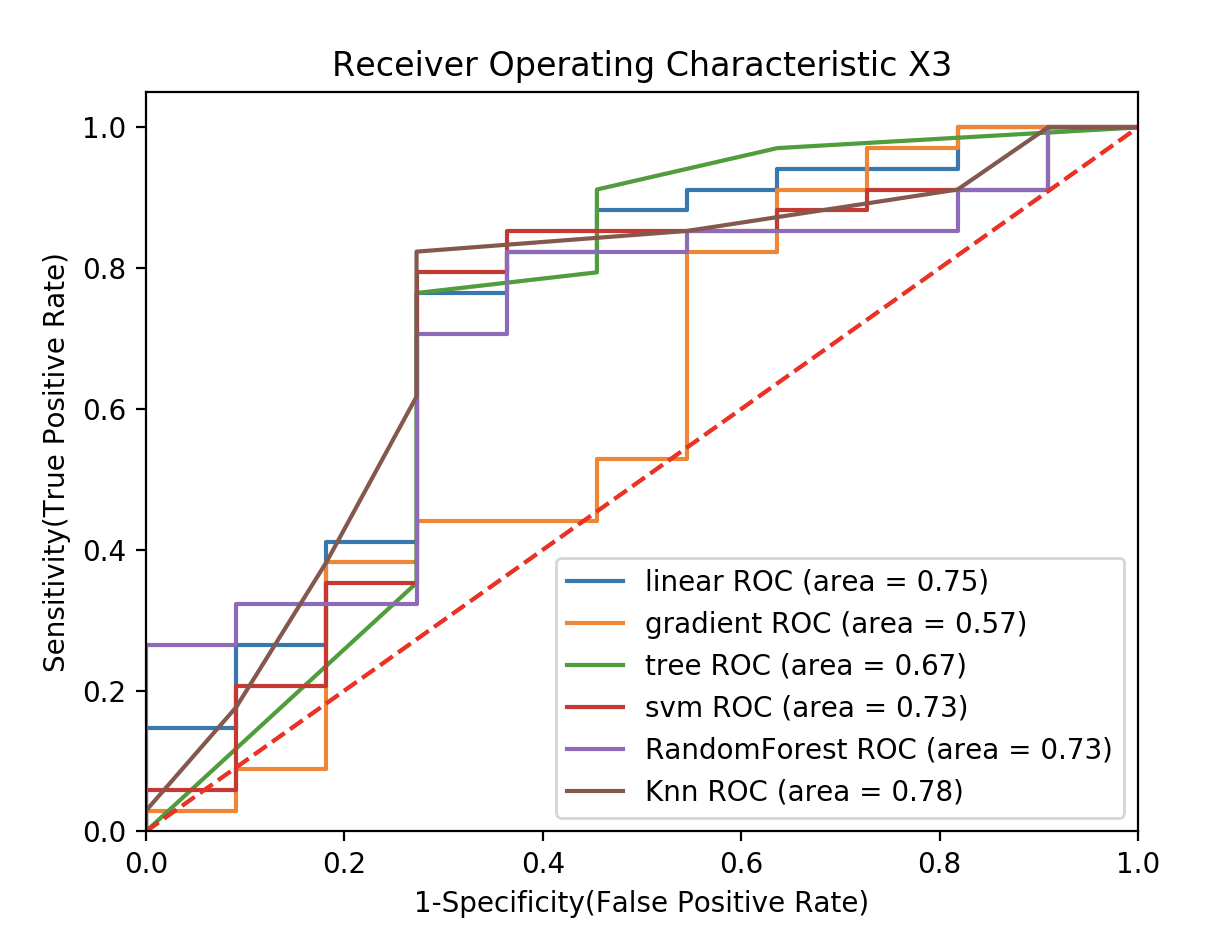
## Fase V. Evaluation. Evaluación (obtención de resultados)

Los mejores resultados fueron los arrojados por la regresión logística en las base de datos bd1 y bd2, seguido del random forest, ambos con variabilidad en los resultados en todas las corridas muy similar. La siguiente tabla muestra los resultados obtenidos en cada corrida y el promedio y desviación estadándar del total.



Por último para comprender mejor el resultado de cada uno de los modelos en los 5 datasets se realizó un comparativo de las curvas ROC como se observa a continuación.



Imagen que contiene texto, mapa

Descripción generada automáticamente

Imagen que contiene texto, mapa

Descripción generada automáticamente

## Fase VI. Deployment. Despliegue (puesta en producción)

Esta sería la última fase del proceso, y es la puesta en producción del modelo. Teniendo en cuenta todo el análisis anterior el modelo recomendado para la puesta en producción sería el obtenido con la regresión logística con la bd1 ya que presenta los mejores resultados de AUC. También puede utilizarse el random forest y tomar el mejor resultado entre ambos, ya que dependiendo de la selección del conjunto de train y test los resultados pueden variar.

## Conclusiones

Luego de realizar todo el análisis, correr diferentes modelos, con los diferentes conjuntos de variables seleccionadas, se observa que el modelo recomendado para predecir la variable ‘yL’ es la regresión logística con la bd1, ya que se obtiene un nivel del AUC promedio el 0.7 y el mejor valor de la métrica Recall con respecto a los demás modelos.

Durante todo el proceso de modelado se pudo observar la sensibilidad de los resultados al variar la partición de la muestra (train, test), por lo tanto para evitar el sobreajuste y escoger el mejor modelo se realizaron diferentes particiones (10 en total) y se compararon los resultados del AUC. Al final se tomó la decisión con el valor promedio del AUC en todas las corridas realizadas.