**Conjunto Binario**

Luego de realizar el análisis descriptivo para el entendimiento de los datos y la detectección de patrones y correlaciones, 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).

**Selección de variables**

Para realizar el proceso de selección de variables 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**

Para los métodos donde de variaron los hiperparámetros de regularización se seleccionaron la cantidad de variables donde se lograba el mayor resultado de AUC.

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:

Pegar imagen matriz.png

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:

**# variables con probabilidad de selección mayor al 55% (18 variables):**

x1, x2, x3, x4, x10, x11, x12, x14, x17, x18, x21, x22, x23, x24, x25, x27, x30, x32

**# variables con probabilidad de selección mayor al 60% (12 variables):**

x1, x2, x3, x4, x10, x11, x12, x17, x21, x23, x25, x27

**# variables con probabilidad de selección mayor al 65% (11 variables):**

x1, x2, x4, x10, x11, x12, x17, x21, x23, x25, x 27

**# variables con probabilidad de selección mayor al 70% (8 variables):**

x1, x2, x4, x10, x11, x12, x17, x23

**# variables con probabilidad de selección mayor al 80% (5 variables):**

x1, x10, x12, x17, x23

Con base en lo anterior se construyeron cinco subconjuntos de bases adicionales, cada una con las variables seleccionadas por cada criterio de selección.

**Selección del modelo**

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 7 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 y Gaussian Process Classifier. Para evaluar el rendimiento de los modelos se analizó el ROC AUC en el subconjunto de test y también se analizó el accuracy. Cada modelos en cada conjunto de variables seleccionadas se corrió variando los hiperparámetros y con validación cruzada con kfold = 5. De cada modelo se escogió el mejor para cada subset tanto teniendo en cuenta la métrica de accuracy como la de auc.

Este proceso se corrió muchas veces (realizaron varias particiones aleatorias sobre la base de datos original) y se detectó que el mejor modelo variaba significativamente, pero el conjunto 5 (de menor número de variables varias veces fue el mejor). Luego se corrieron con estas 5 características sobre el test cada uno de estos modelos y se seleccionó el que mejor estaba generalizando. En este caso los mayores valores de auc y de accuracy se obtuvieron con el árbol de decision, a pesar de que muchas veces en la validacion inicial se lograba un mejor ajuste con vecinos cercanos.

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.

## Conclusiones

Los mejores resultados sobre el conjunto test fueron los arrojados por el arbol de decisión en la base con 5 carácterísticas, seguido del random forest.

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 y el conjunto de variables más pequeño (5 variables), ya que lograba resultados muy similares al conjunto con mayor número de variables