# Proyecto Final

# Machine Learning

# Yorch Sepúlveda

**Problema**

Los datos proporcionados tienen un origen no revelado, pero corresponden a una aplicación real en el campo de la ingeniería. El problema consiste en encontrar un modelo de regresión para predecir de manera efectiva la variable dependiente (Y).

**Hipótesis**

La NO eliminación de datos atípicos que se observan al graficar el promedio de las columnas, no afecta a la reducción de dimensiones y elección de atributos.

**Datos**

Los datos para este estudio se encuentran públicamente disponibles de la siguiente URL: <https://goo.gl/64CUV7>. Estos datos cuentan con 481 columnas de números reales y 42 instancias.

**Preprocesamiento de los datos**

Primeramente, se verifica que no existan datos Na dentro del set de datos. Luego se procedió a graficar los datos en brutos ara analizar como se encuentran distribuidos los datos. Para ello se obtuvieron set de datos de los mínimos, los promedios y los máximos de cada columna X. Tras esto se procede a graficar cada uno de ellos y se compara si hay alguna anomalía. Al realizar lo anterior se ve que existe datos que están muy extremos que perfectamente se pensaría que son atípicos, pero no se eliminará ya que se ocupara la técnica de normalización y esta se encargara de detectar los datos atípicos.

Al normalizar los datos se percata que existen columnas completas con Na las cuales se procede a eliminar. En total fueron nueve columnas las eliminadas. En cuanto a datos, de un total de 20160 datos (que se encuentran en las columnas X) se eliminaron 378 datos lo cual equivale a 1,875% de los datos.

**Normalización en etapa de post-procesamiento**

Tal como se dijo en el enunciado, se aplica nuevamente normalización de los datos para ver si los rangos de las variables cambiaron. Al analizar esta normalización, los rangos no variaron tras la eliminación de los datos atípicos.

**Visualización**

Para la etapa de visualización se utiliza Principal Component Analysis (PCA), en donde al procesar los datos normalizados en un grafico 3D se aprecia una anomalía en donde existen algunos de los datos que se encuentran demasiado cerca y que en un grafico 2D es muy poco clara su visualización.

**Reducción de la dimensionalidad**

Una de las alternativas para mejorar la predicción es reducir el número de dimensiones, es por el que se utiliza el algoritmo Boruta y según lo que recomiende este algoritmo se visualizaran los datos realizando un análisis de correlación de los datos.

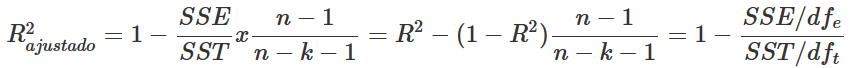
**Regresión**

Dato que la variable dependiente Y se determina a partir de un conjunto de variables independientes llamadas predictores (X1, X2, X3, …) es que utiliza la regresión lineal múltiple.

**Medidas de desempeño**

La medida de desempeño que se utiliza es R2 (coeficiente de determinación). Sin embargo, cuanto mas predictores se encuentren en el modelo, mayor es el valor de R2 ya que, por poco que sea, cada predictor va a explicar una parte de la variabilidad observada en Y.

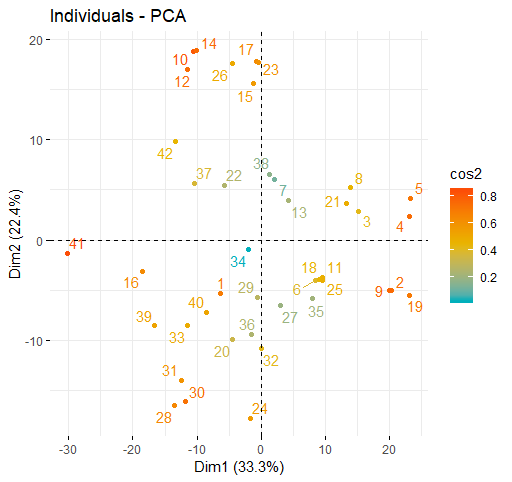
Además, se utiliza R2ajustado el cual penaliza al valor de R2 por cada predictor que se introduce al modelo. El valor de la penalización depende del número de predictores utilizados y del tamaño de la muestra, es decir, del número de grados de libertad. Cuanto mayor es el tamaño de la muestra, más predictores se pueden incorporar en el modelo. Esta medida de desempeño ayuda a encontrar el mejor modelo que permita explicar la variabilidad de Y con el menor número de predictores.

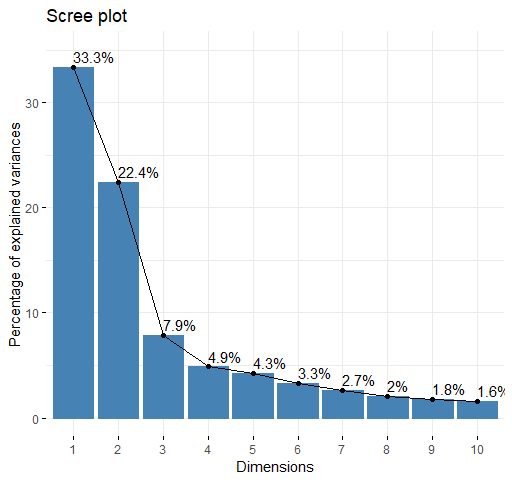


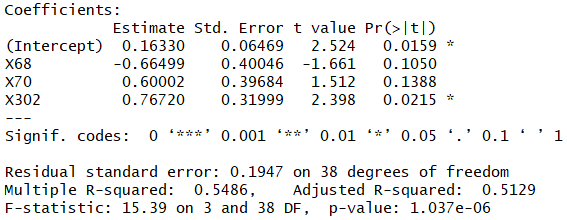
Siendo SSE la variabilidad explicada por el modelo (Sum of Squares Explained), SST la variabilidad total de Y (Sum of Squares Total), n el tamaño de la muestra y k el número de predictores introducidos en el modelo.

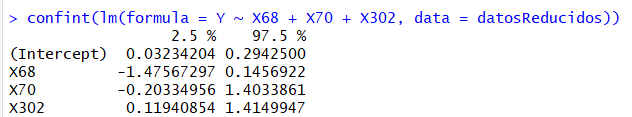
**Validación**

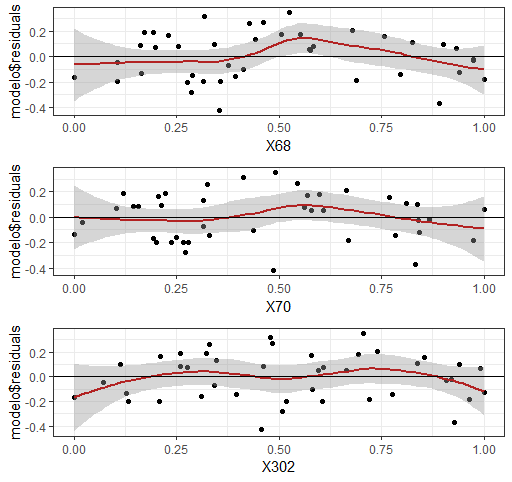
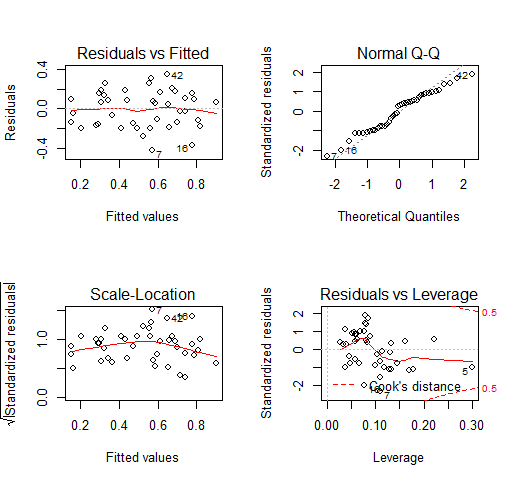
Se utiliza knn.cv para la validación de los datos, y el clasificador de datos es KNN

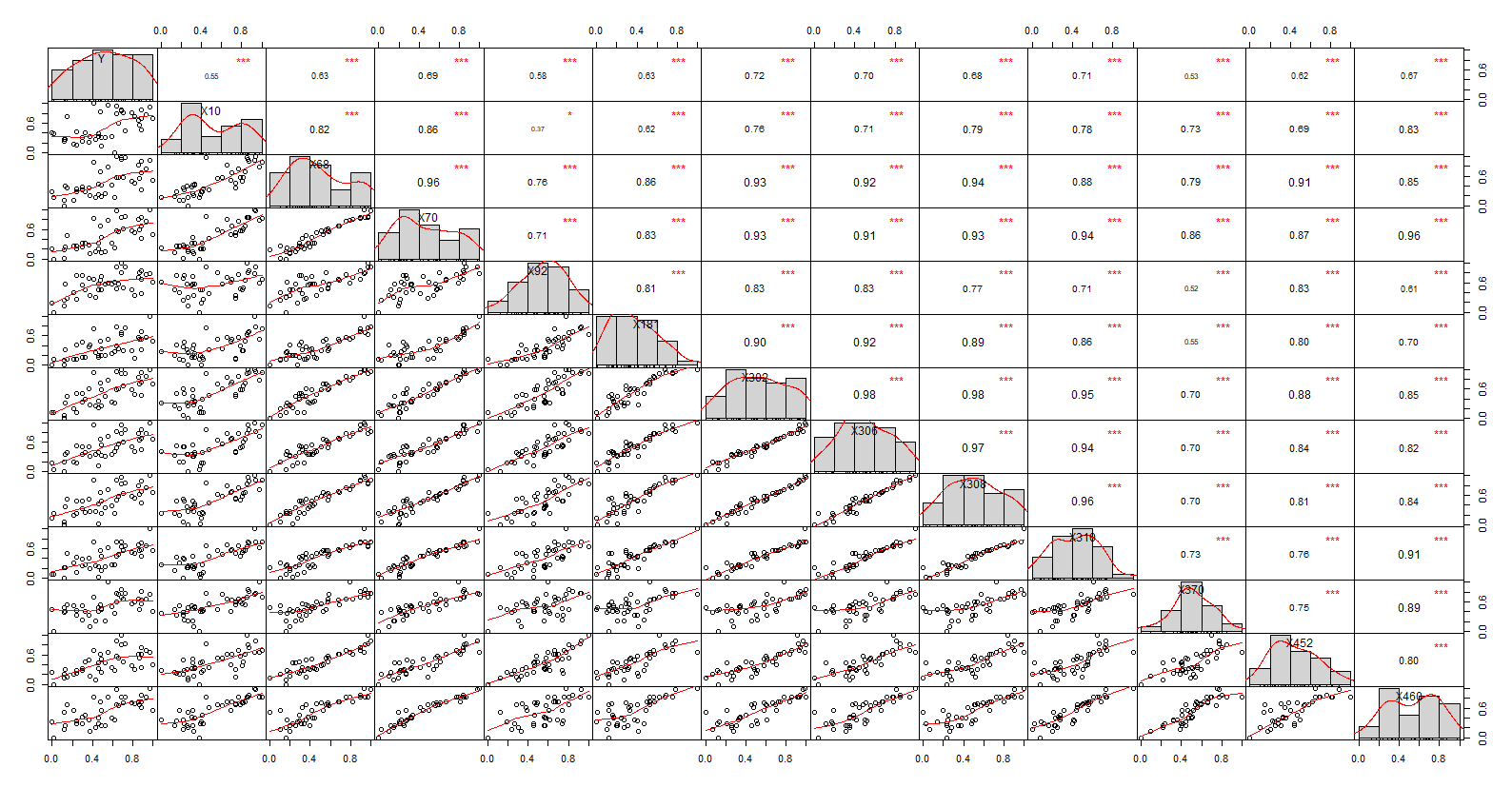
**Resultados**

****





****



**Conclusión**

Para concluir, en lo anterior se realiza un análisis de los datos los cuales son de origen desconocido, pero al realizar los procedimientos mencionados con anterioridad con lm la cual es una función de modelo lineal, como el análisis de regresión lineal, se obtuvo el modelo con todas las variables introducidas como predictores la cual tiene un R2 considerable (0.5486) y un R2ajustado similar (0.5129) , que es capaz de explicar el 50,86% de la variabilidad observada en Y. El p-value del modelo es significativo (1.037e-10) por lo que se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales de regresión es distinto de 0. Muchos de ellos no son significativos, lo que es un indicativo de que podrían no contribuir al modelo.

En cuanto a la hipótesis, no queda bien en claro de que pueda ser aceptada. Si bien al pasar por los procesos mencionados con anterioridad donde los datos atípicos fueron otros (los que se encargaron de encontrar los datos atípicos son la normalización de los datos y a la reducción de dimensiones que realiza a través del algoritmo boruta) y no los que previamente se revisaron al graficar el promedio de las columnas (X’s), el modelo no logra ser consistente ya que solo puede explicar el 50,86% de la variabilidad observada en Y.