Deep Learning Proyecto

Parámetros críticos que afectan el índice de Eficiencia Global de Equipamiento (OEE - Overall Equipment Effectiveness)

Annia Díaz Trivel

Objetivo

El objetivo de este proyecto es poder identificar los parámetros críticos que afectan el índice de OEE (Overall Equipment Effectiveness, Eficiencia Global de Equipamiento)

Ind. OEE = disponibilidad x desempeño x calidad

disponibilidad% = (tiempo produciendo / tiempo programado para producir) x 100

desempeño% (cantidad de producción real / cantidad de producción teórica) x 100

calidad% = (cantidad de productos buenos / cantidad total producida) x 100

Disponemos de información de 3 equipos distintos desde julio del 2016. Detallado por hora, día, mes, año, equipo, modelo a producir entre otros parámetros. Realicé una limpieza de los datos con dropna y me quedaron 21417 registros
Tenemos información de 50 parámetros la idea es poder identificar cuáles son los más relevantes para obtener un Índice de OEE óptimo (ValOEE)

```
'OEEDIA' : Día
'OEEMES': Mes
'OEEANIO': Año
'OEEHORARIO': Horario
'OEETURNO': Turno
'OEEEQUIPO':Equipo
'OEETREALMI': Tiempo Real en Minutos
'MODELOCOD':Modelo
'OEEPRD':Producto
'MODELOANT': Modelo Anterior
'OEEPRDANT': Producto Anterior
'OEECNTOP': Cantidad de Operarios
'OEEDESAC': Desacoplado
'OEEOBJ': Objetivo
'OEEOBJAC': Objetivo Acumulado
'OEEPRREAL': Producción Real
'PORCPROD': Porcentaje de Producción
'OEEPRAC': Producción Real Acumulada
'OEEFALL1': Código Falla 1
'OEEMFA1': Minutos Falla 1
'OEEFALL2': Código Falla 2
'OEEMFA2': Minutos Falla 2
'OEEFALL3': Código Falla 2
'OEEMF3': Minutos Falla 3
'OEEFALL4': Código Falla 4
'OEEMF4': Minutos Falla 4
'OEEMIND': Minutos Causa Indeterminada
'OEETCP': Tiempo Ciclo Segundos por Pieza
'OEEPPU': Pérdida Por Performance en Unidad
'OEEPPS': Pérdida Por Performance en Segundos
'OEETMP': Total Minutos Perdidos
'OEEPPPH': Productividad en Piezas por Persona / Hora
```

```
'OEEESTU': Código Estuche
'OEEPREST': Proveedor Estuche
'OEEDDES':Días en depósito
'OEESOBRE': Código Sobre
'OEEPSOB': Proveedor Sobre
'OEEDDS': Días en depósito
'OEELITER': Código Literatura
'OEEPLIT': Proveedor Literatura
'OEEDDL': Días en depósito Literatura
'OEEETIH': Código Etiqueta Holográfica
'OEEPREH': Proveedor Etiqueta Holográfica
'OEEDDEH': Días en depósito Etiqueta Holográfica
'OEEETH2': Código Etiqueta Holográfica 2
'OEEPEH2': Proveedor Etiqueta Holográfica 2
'OEEDEH2': Días en depósito Etiqueta Holográfica 2
'OEEPRES': Presentación
'OEECNTPRID': Cantidad Producida
'OEEORDEN': Orden
'VALOEE': Si Valor de índice de OEE supera el 80 %
```

Split Data

```
#Preparo test y train
new_train_size = int(len(df) * 0.67)
new_test_size = len(df) - new_train_size

new_x_train, new_xtest, new_y_train, new_ytest =
train_test_split(X_new, y_new, test_size=new_test_size,
random_state=42)
```

Separo los datos en Train (14349) y Test (7068)

Técnicas de Feature Selection

Feature Selection es el proceso de reducir el número de variables de entrada al desarrollar un modelo predictivo. Agregar variables redundantes reduce la capacidad de generalización del modelo y también puede reducir la precisión.
Es deseable reducir el número de variables de entrada para reducir el costo computacional del modelado y, en algunos casos, para mejorar el rendimiento del modelo.

Podemos resumir las técnicas de Feature Selection de la siguiente forma:

- **Feature Selection:** Seleccionando un subconjunto de parámetros de entrada del conjunto de datos completo.
    - **Unsupervised:** Ejemplo eliminando variables redundantes.
        - Basados en Correlación
    - **Supervised:** Eliminando variables irrelevantes
        - **Filter (Filtrado):** Selecciona las características en forma independiente del clasificador, usando un criterio de “relevancia”.
            - Statistical Methods
            - Feature Importance Methods
        - **Wrapper (Encapsulado):** Selecciona los subconjuntos de características en función del desempeño de un clasificador. Costoso computacionalmente. Necesita estrategia de búsqueda para explorar en forma eficiente el espacio de subconjuntos.
            - RFE (Recursive Feature Elimination)
        - **Intrinsic or embedded (Intrínseco):** Realizan la selección en el proceso de aprendizaje devuelve un subconjunto de características y el clasificador entrenado. n entrenamientos, evalúo costo de agregar o quitar característica pero no reentreno.
            - **Decision Trees**
- **Dimensionality Reduction:**

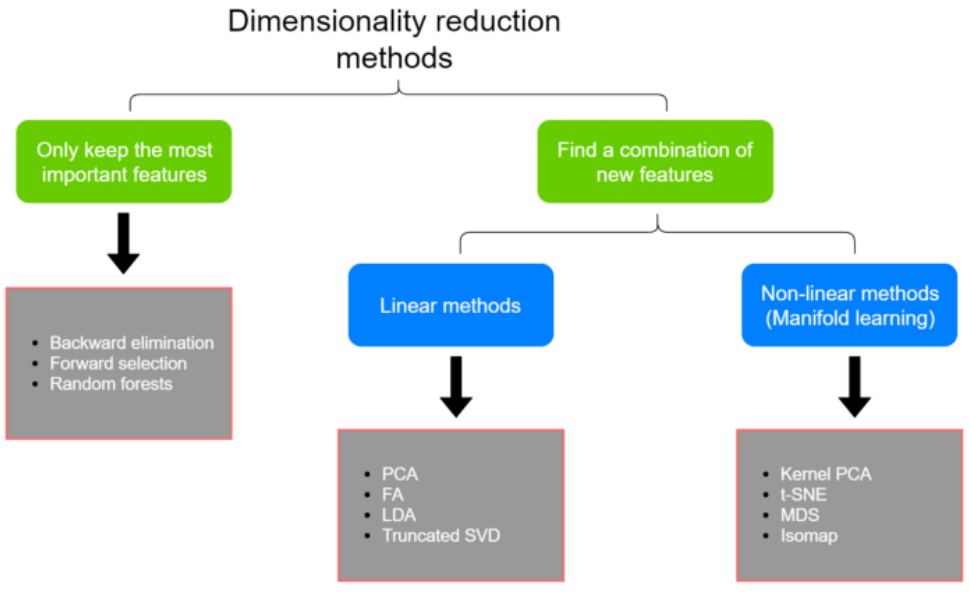
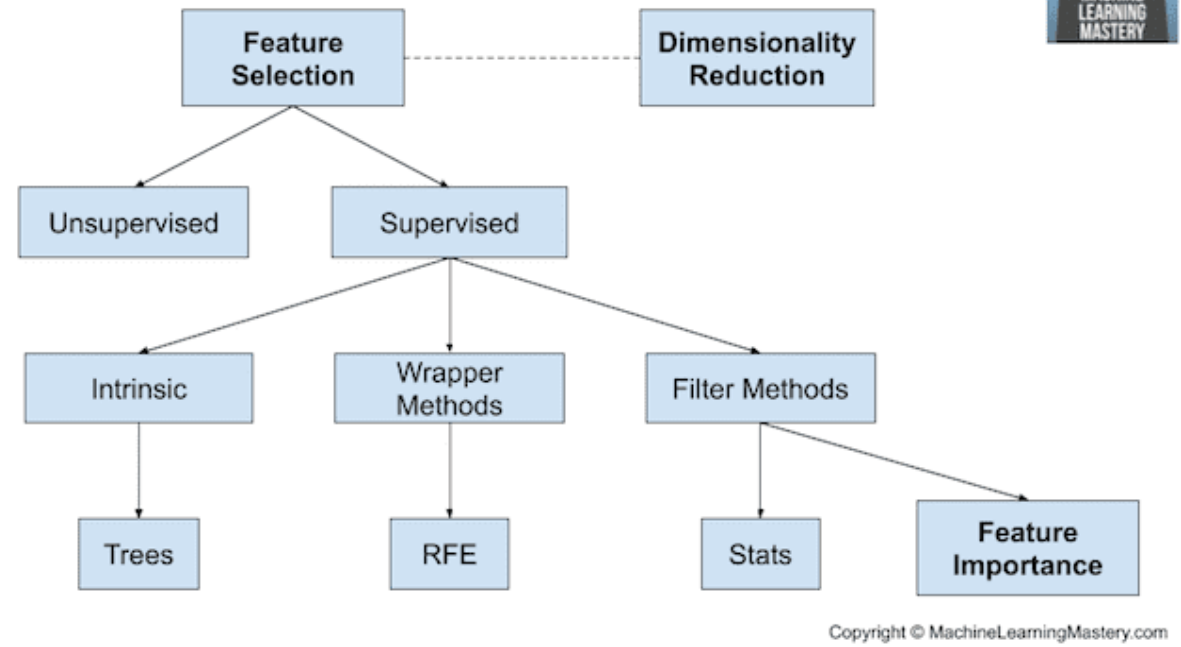
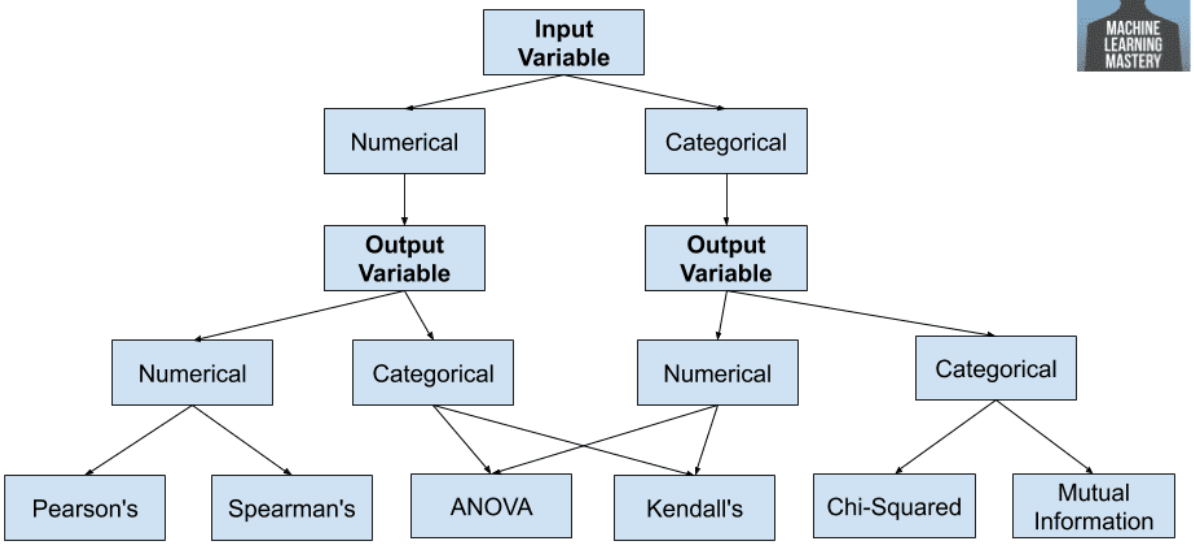


Image copyright: Rukshan Pramoditha

Overview of Feature Selection Techniques



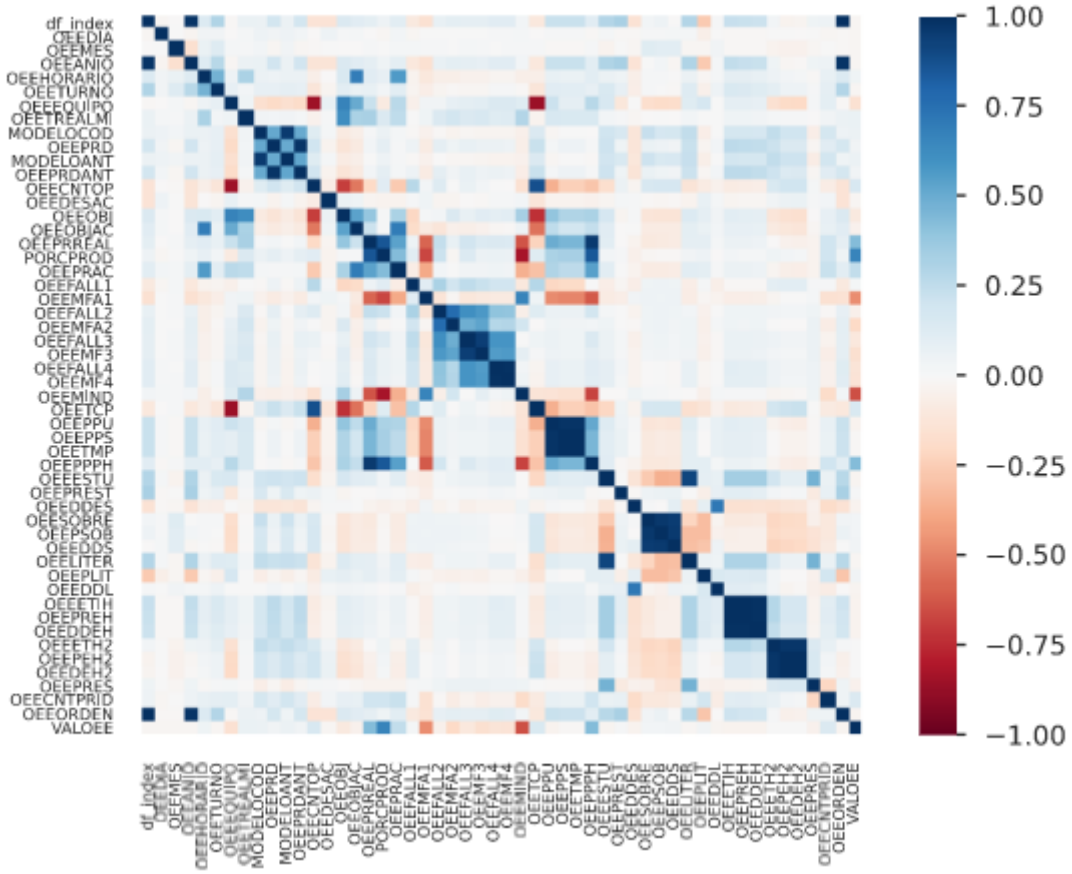
How to Choose a Feature Selection Method



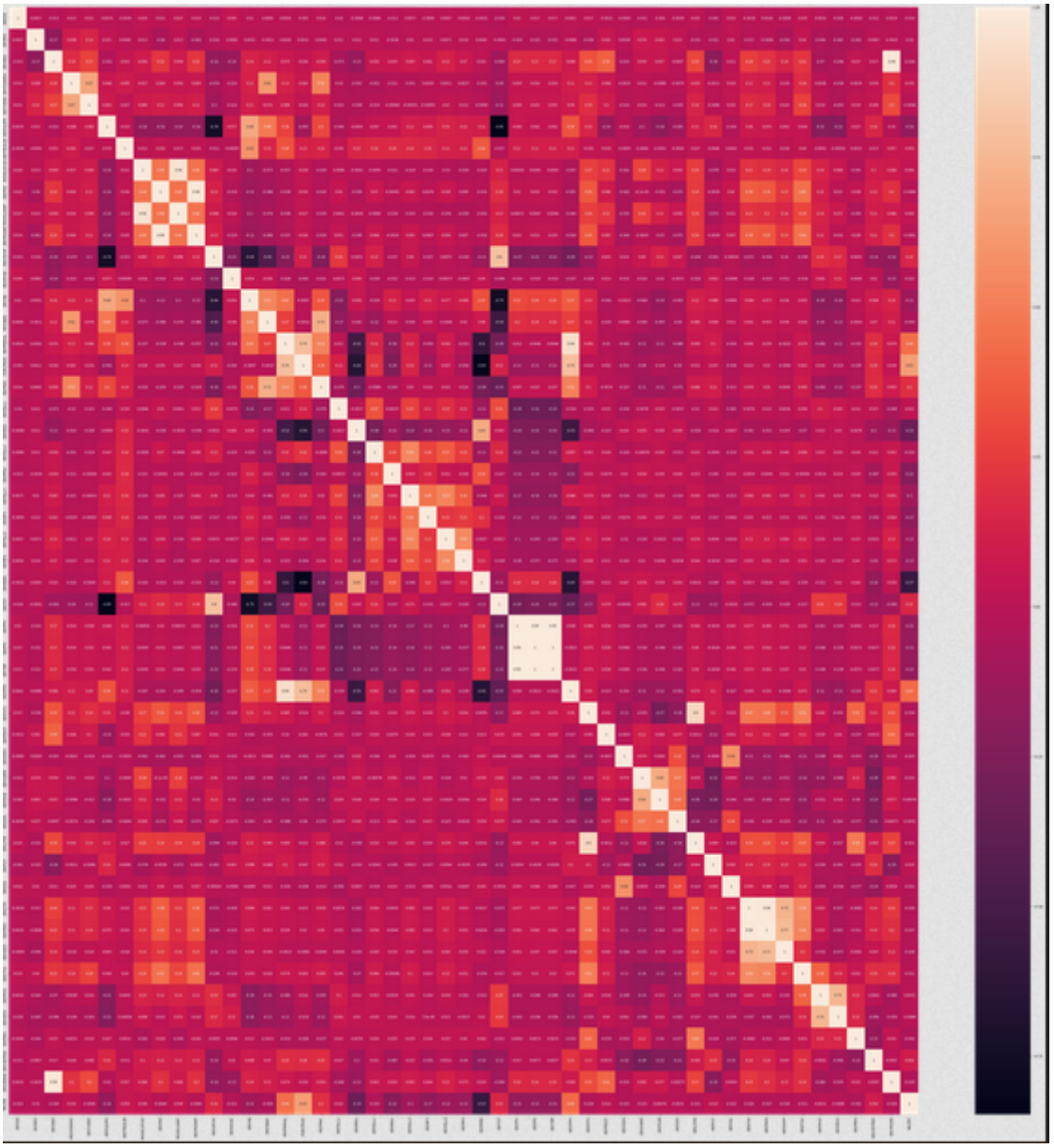
Estas son las técnicas de Feature Selection que utilicé:

Pandas Profiling y Matriz de Correlación

Utilicé el módulo Pandas Profiling de Python para hacer un rápido análisis exploratorio de los datos. El análisis obtenido se encuentra en el informe interactivo (IndiceOEE.html)
Al ejecutar Pandas Profiling obtuve la siguiente matriz de correlación



Matriz de Correlación de Pandas Data Frame



La matriz de correlación de Pandas DataFrame coincide con la correlación indicada de Pandas Profiling en que los 6 más correlacionadas son:

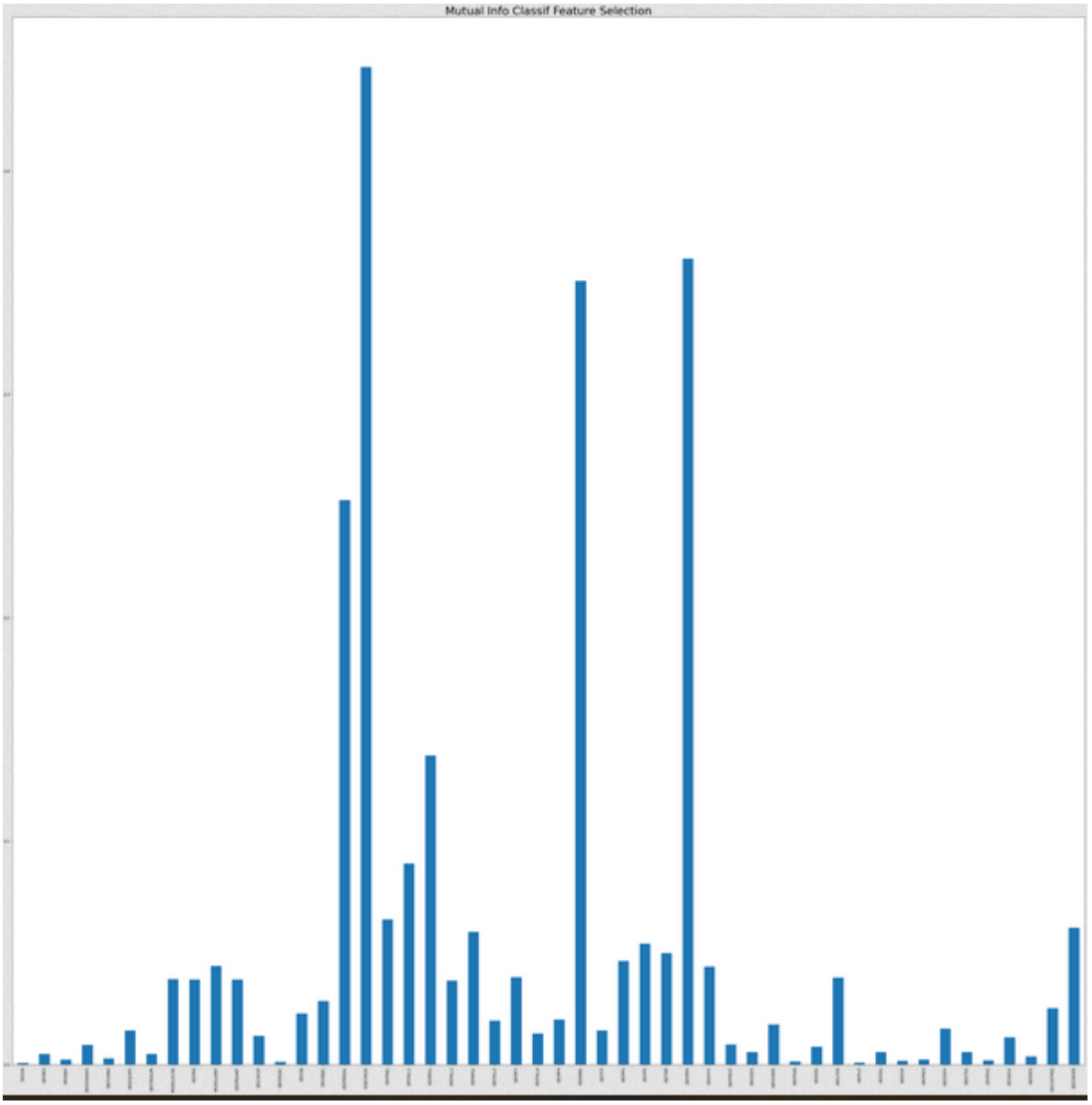
```
PORCPROD      0.634260 Porcentaje de Producción
OEEPRREAL     0.442007 Producción Real
OEEPPPH       0.427464 Productividad en Piezas por Persona / Hora
OEEPRAC       0.197518 Producción Real Acumulada
OEECNTOP      0.155355 Cantidad de Operarios
OEETCP        0.154570 Tiempo Ciclo Segundos por Pieza
```

Mutual Info Classif

Calcula la reducción de entropía a partir de la transformación de un conjunto de datos. Se puede utilizar para la selección de funciones mediante la evaluación de la ganancia de información de cada variable en el contexto de la variable de destino

Los parámetros más importantes según Mutual Info Classif Feature Selection fueron:

PORCPROD	0.445617
OEEPPPH	0.361114
OEEMIND	0.351444
OEEPRREAL	0.253246
OEEMFA1	0.146362
OEEFALL1	0.091558
OEEORDEN	0.065926
OEEPRAC	0.064727
OEEMFA2	0.059071
OEETMP	0.052889
OEEPPS	0.051766
OEEESTU	0.047249
OEEPPU	0.045187
MODELOCOD	0.042780
OEELITER	0.041840
MODELOANT	0.041125



SelectKBest with Chi-square

Usé K = 20

```
SelectKBest with Chi-square Test Feature_Selection
Original feature number:  50
Reduced feature number:  20

Index:  [ 5 12 14 15 16 17 19 21 22 23 24 25 26 27 28 29 30 31 41 42]

Important Features:  Index(['OEEEQUIPO', 'OEEDESAC','OEEOBJAC',
'OEEPRREAL','PORCPROD', 'OEEPRAC','OEEMFA1','OEEMFA2','OEEFALL3',
'OEEMF3', 'OEEFALL4', 'OEEMF4','OEEMIND','OEETCP','OEEPPU','OEEPPS',
'OEETMP','OEEPPPH','OEEETIH','OEEPREH'], dtype='object')
```

Los parámetros más importantes según SelectKBest y ChiSquare fueron:

```
'OEEEQUIPO': Equipo
```

'OEEDESAC': Desacoplado
'OEEOBJAC': Objetivo Acumulado
'OEEPRREAL': Producción Real
'PORCPROD': Porcentaje de Producción
'OEEPRAC': Produccción Real Acumulada
'OEEMFA1': Minutos Falla 1
'OEEMFA2': Minutos Falla 2
'OEEFALL3': Falla 3
'OEEMF3': Minutos Falla 3
'OEEFALL4': Falla 4
'OEEMF4': Minutos Falla 4
'OEEMIND': Minutos Causa Indeterminada
'OEETCP': Tiempo Ciclo Segundos por Pieza
'OEEPPU': Pérdida Por Performance en Unidad
'OEEPPS': Pérdida Por Performance en Segundos
'OEETMP': Total Minutos Perdidos
'OEEPPPH':Productividad en Piezas por Persona / Hora
'OEEETIH':Código de Etiqueta Holográfica
'OEEPREH':Proveedor Etiqueta Holográfica

Feature selection is performed using Pearson's Correlation Coefficient via the f_regression() function

Usé K = 20

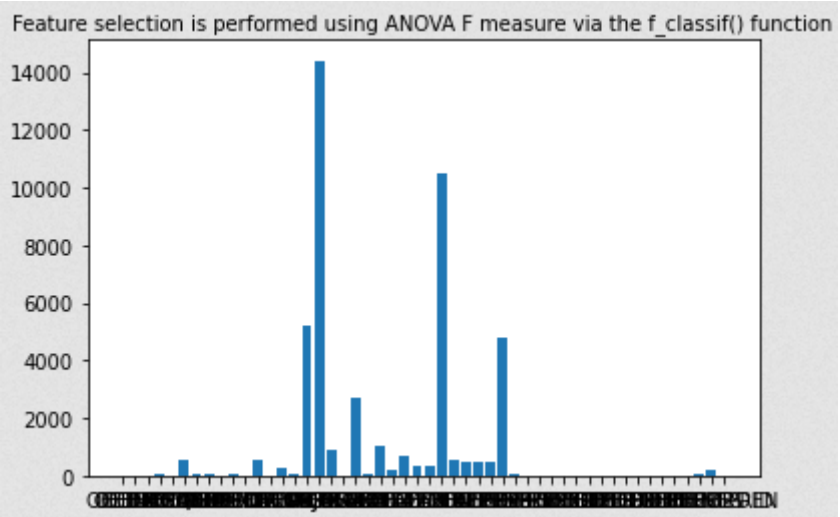
```
Feature selection is performed using Pearson's Correlation Coefficient
via the f_regression() function
Original feature number:  50
Reduced feature number:  20

Index:  [ 5  6 11 13 15 16 17 19 21 22 23 24 25 26 27 28 29 30 31 48]

Important Features:  Index(['OEEEQUIPO','OEETREALMI','OEECNTOP',
'OEEOBJ','OEEPRREAL','PORCPROD','OEEPRAC','OEEMFA1','OEEMFA2',
'OEEFALL3','OEEMF3','OEEFALL4','OEEMF4','OEEMIND', 'OEETCP','OEEPPU',
'OEEPPS','OEETMP','OEEPPPH', 'OEECNTPRID'],dtype='object')
```

Los parámetros más importantes según SelectKBest y f_regression fueron:

'OEEEQUIPO',
'OEETREALMI',
'OEECNTOP',
'OEEOBJ',
'OEEPRREAL',
'PORCPROD',
'OEEPRAC',
'OEEMFA1',
'OEEMFA2',
'OEEFALL3',
'OEEMF3',
'OEEFALL4',

```
'OEEMF4',
'OEEMIND',
'OEETCP',
'OEEPPU',
'OEEPPS',
'OEETMP',
'OEEPPPH',
'OEECNTPRID'
```

Feature selection is performed using ANOVA F measure via the f_classif() function

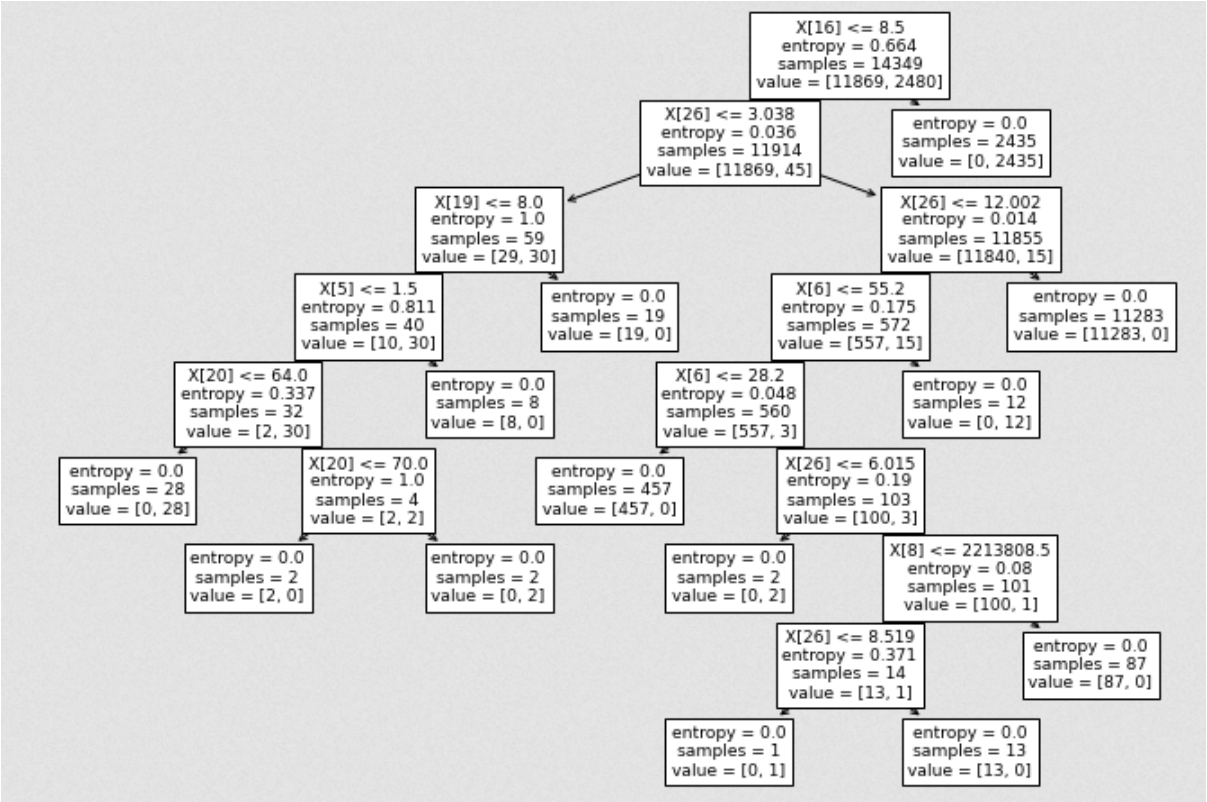


Best Scores:

PORCPROD	14413.130004
OEEMIND	10463.015843
OEEPRREAL	5199.718554
OEEPPPH	4787.933463
OEEMFA1	2666.909433
OEEMFA2	1043.090207
OEEPRAC	869.387560
OEEMF3	675.441094
OEEEQUIPO	545.704485
OEECNTOP	529.635330
OEETCP	524.170174
OEEPPS	494.887923
OEETMP	494.887033
OEEPPU	480.992321
OEEMF4	347.308345
OEEFALL4	329.973148
OEEOBJ	254.646615
OEEFALL3	223.266304
OEECNTPRID	182.497905

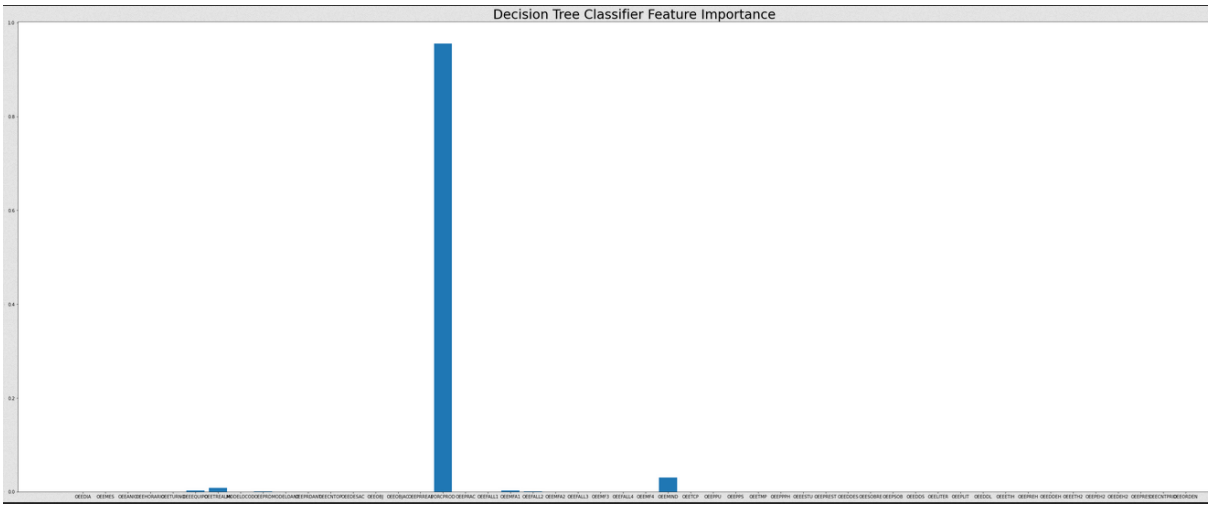
Decision Tree Classifier

Ejecuté con un profundidad 10, obtuve este árbol de Decisión



Detectó los siguientes parámetros como importantes:

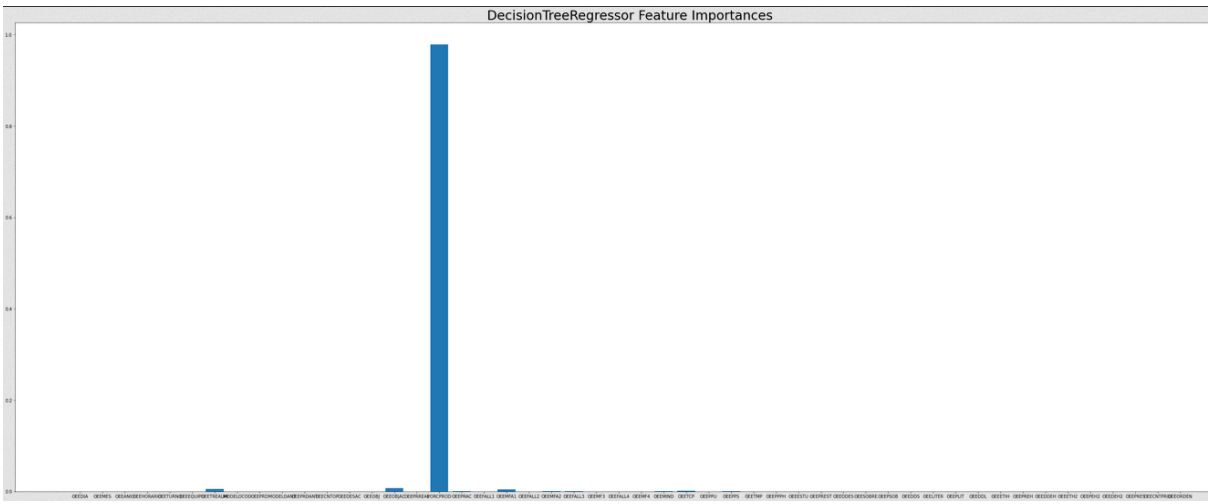
```
PORCPROD      0.955195 Porcentaje de Producción
OEEMIND       0.029855 Minutos Causa Indeterminada
OEETREALMI    0.008456 Tiempo Real Minutos
OEEMFA1       0.002785 Minutos Falla 1
OEEEQUIPO     0.002273 Equipo
OEEFALL2      0.001133 Código Falla 2
OEEPRD        0.000304 Producto
```



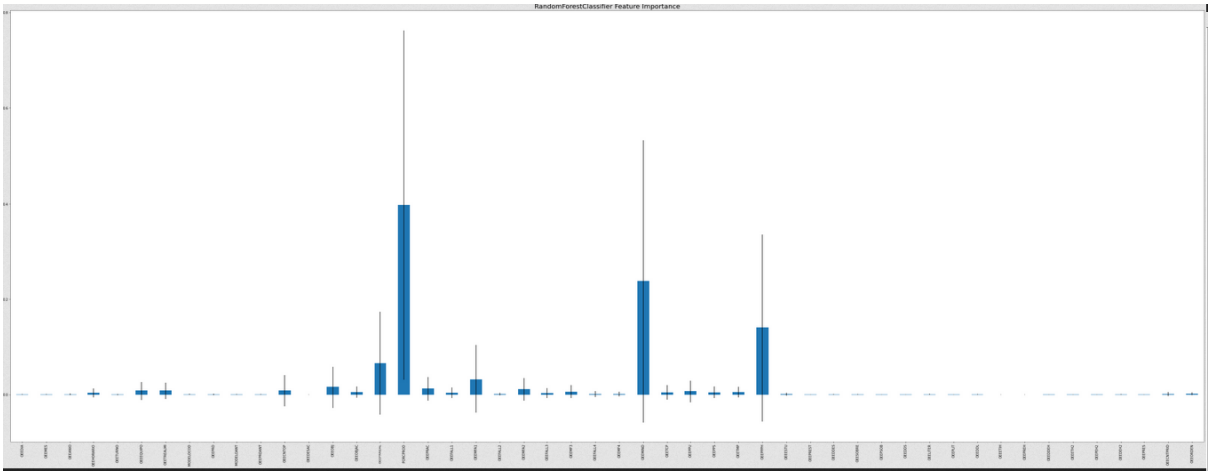
Decision Tree Regressor

Detectó los siguientes parámetros como importantes:

```
PORCPROD      0.978146 Porcentaje de Producción
OEEOBJAC      0.008864 Objetivo Acumulado
OEETREALMI    0.005666 Tiempo Real Minutos
OEEMFA1       0.004523 Minutos Falla 1
OEEMIND       0.000697 Minutos Causa Indeterminada
OEEPRREAL     0.000474 Producción Real
OEEMFA2       0.000469 Minutos Falla 2
OEEMF3        0.000436 Minutos Falla 3
OEEPPU        0.000366 Pérdida Por Performance en Unidad
OEEESTU       0.000239 Código de Estuche
OEEORDEN      0.000119 Orden
```



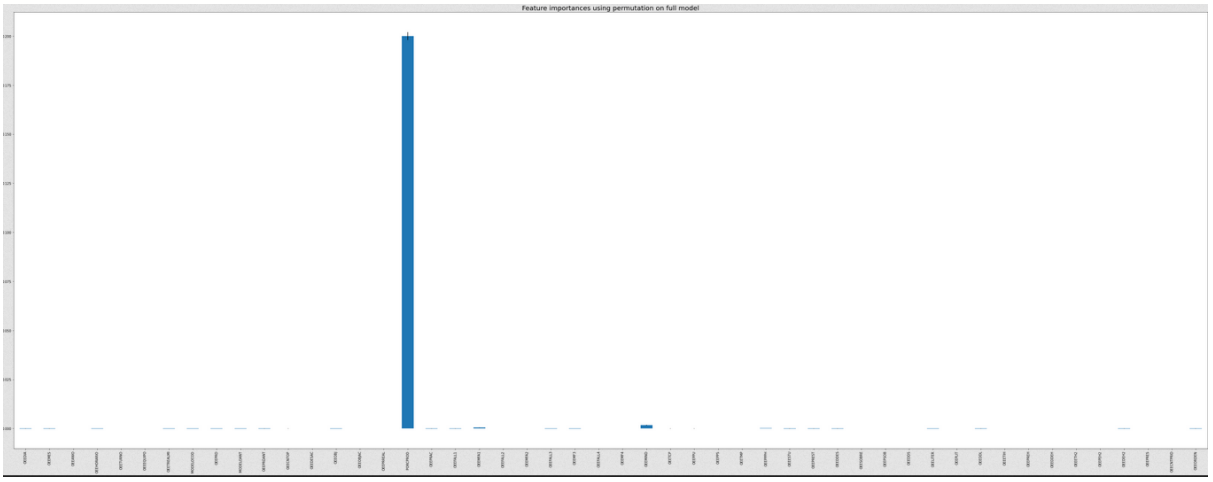
RandomForestClassifier



Detectó los siguientes parámetros como importantes:

PORCPROD	0.397916
OEEMIND	0.237778
OEEPPPH	0.139889
OEEPRREAL	0.066725
OEEMFA1	0.033436
OEEOBJ	0.015825
OEEPRAC	0.012943
OEEMFA2	0.011525

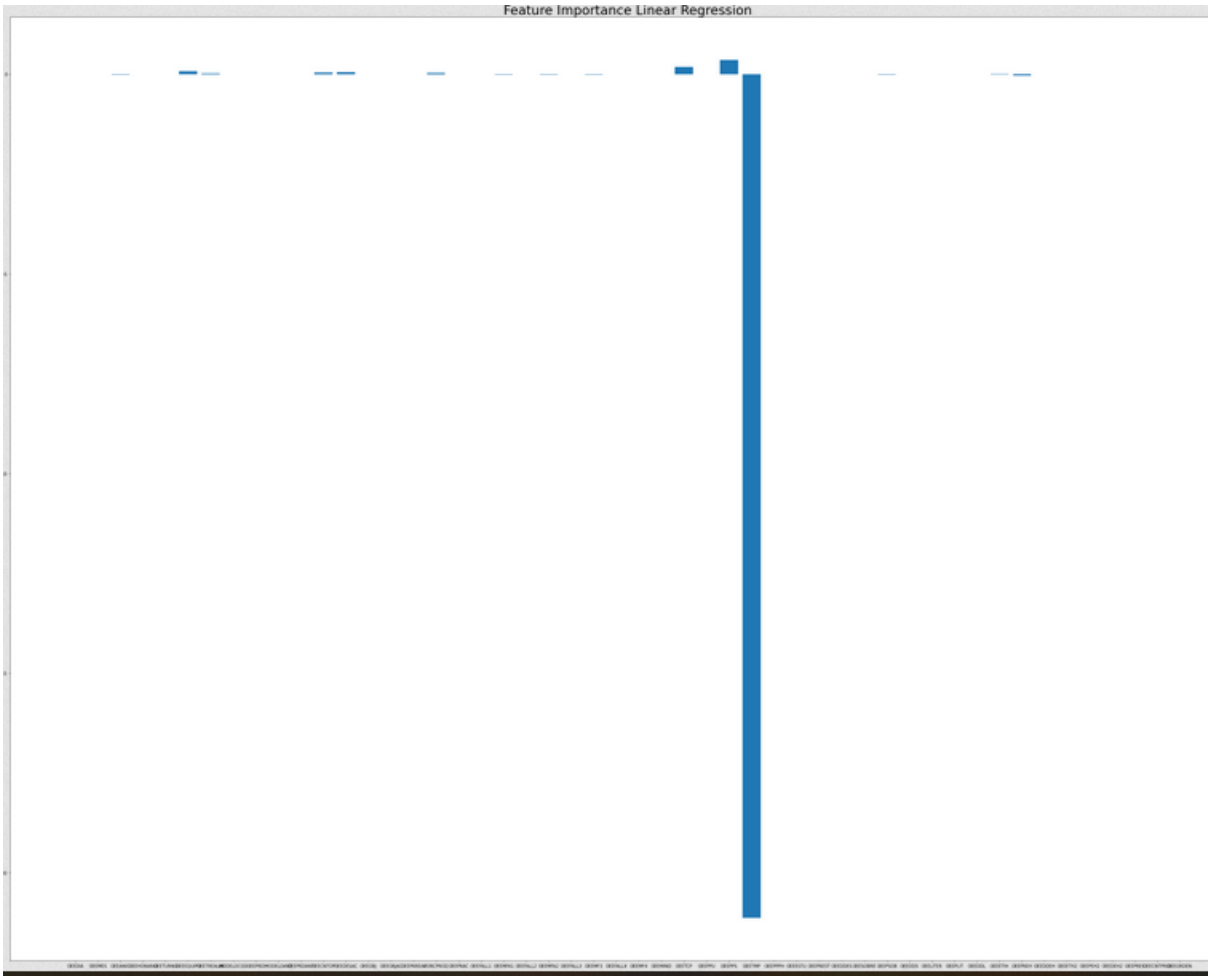
Based on feature permutation



Detectó los siguientes parámetros como importantes:

PORCPROD	0.200028
OEEMIND	0.001807
OEEMFA1	0.000551
OEEPPPH	0.000224
OEEEQUIPO	0.000112
OEEPRREAL	0.000093

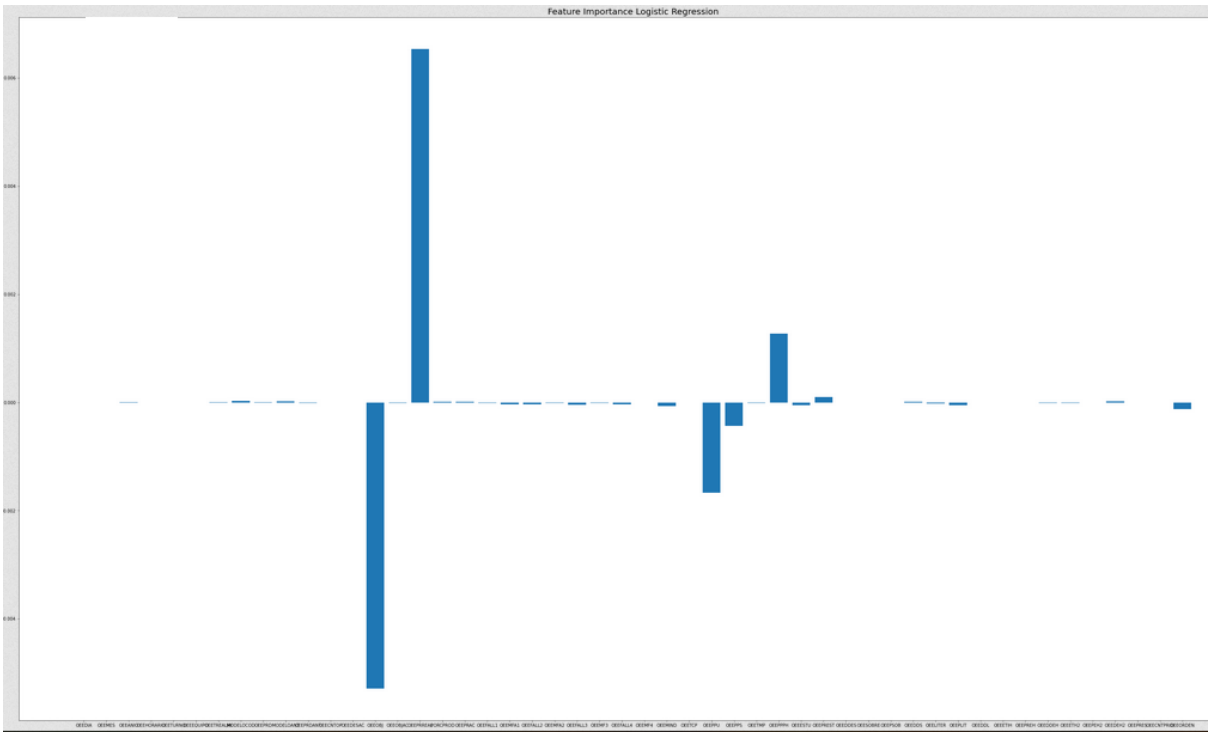
Linear Regression



Detectó los siguientes parámetros como importantes:

```
OEEPPS          3.516569e-01
OEETCP          1.755161e-01
OEEEQUIPO       7.447531e-02
OEEDESAC        4.613333e-02
OEECNTOP        3.583522e-02
PORCPROD        2.996024e-02
OEETREALMI      1.290065e-02
```

Logistic Regression

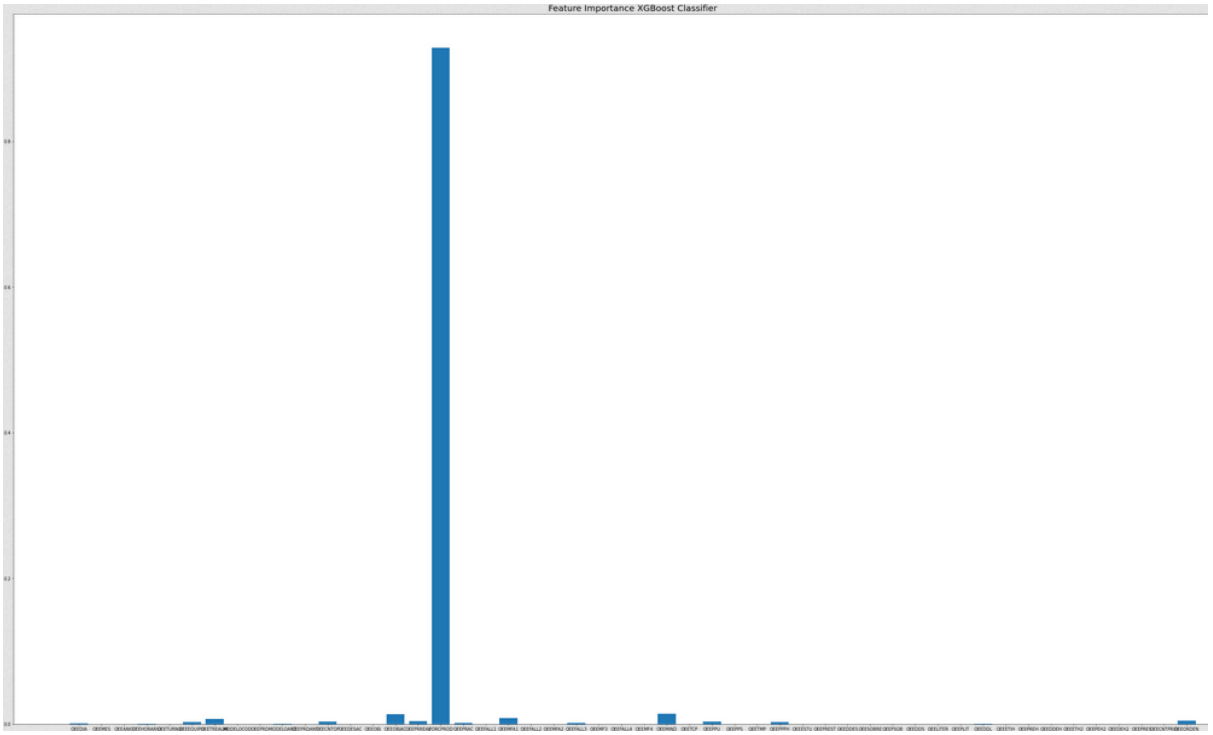


Detectó los siguientes parámetros como importantes:

OEEPRREAL	6.535447e-03
OEEPPPH	1.271367e-03
OEEPREST	9.851704e-05
MODELOCOD	2.558476e-05
MODELOANT	2.510702e-05
OEEDEH2	2.119819e-05
PORCPROD	1.234641e-05
OEEPRAC	1.185208e-05
OEEDDS	1.008422e-05

XGBoost Classifier

```
R2 Score  0.997107907222253
mean_squared_log_error:  0.00020392742526239446
Accuracy: 99.96%
```

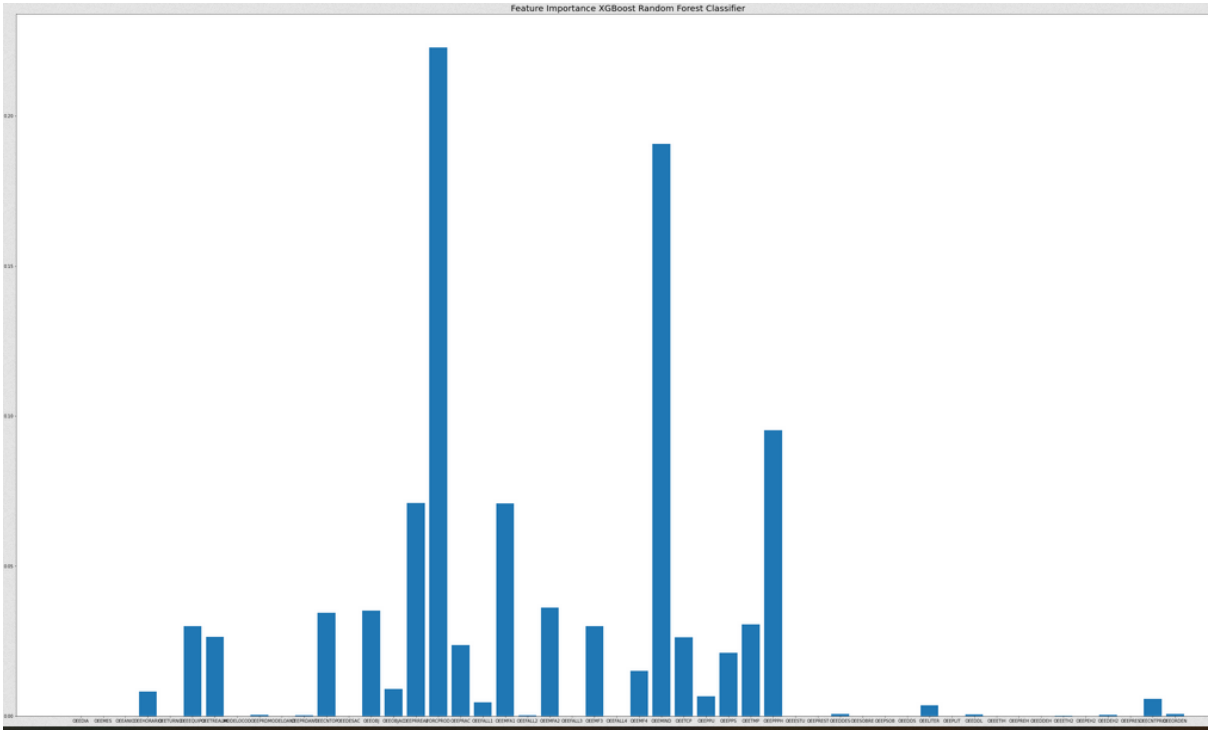


Detectó los siguientes parámetros como importantes:

```
PORCPROD      0.928785
OEEMIND       0.014527
OEEOBJAC      0.013737
OEEMFA1       0.008005
OEETREALMI    0.007296
OEEORDEN      0.004821
OEEPRREAL     0.003961
OEECNTOP      0.003508
OEEPPU        0.003346
OEEEQUIPO     0.002928
OEEPPPH       0.002831
OEEPRAC       0.001730
OEEFALL3      0.001635
OEEDIA        0.001353
MODELOANT     0.000582
OEEHORARIO    0.000580
OEEDDL        0.000349
OEEPPS        0.000027
```

XGBoost Random Forest Classifier

```
Mean Accuracy:
0.9970265349395874
0.0011651128908497595
R2 Score  0.9758992268521076
mean_squared_log_error:  0.0016993952105199537
Random Forest Accuracy: 99.65%
```

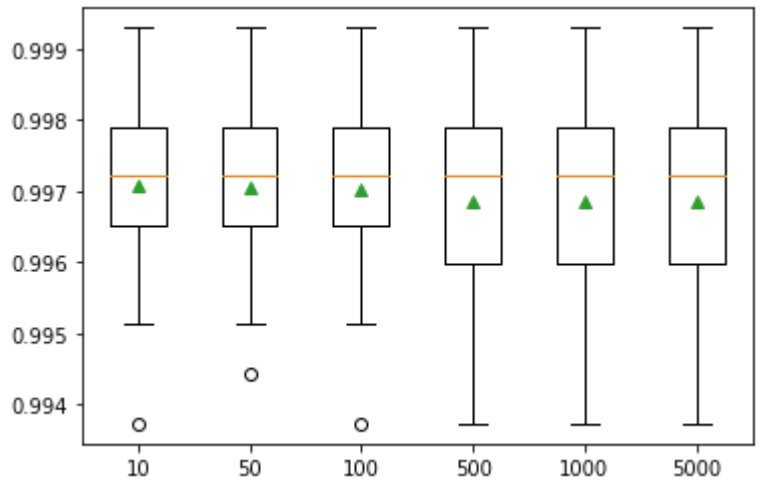


Detectó los siguientes parámetros como importantes:

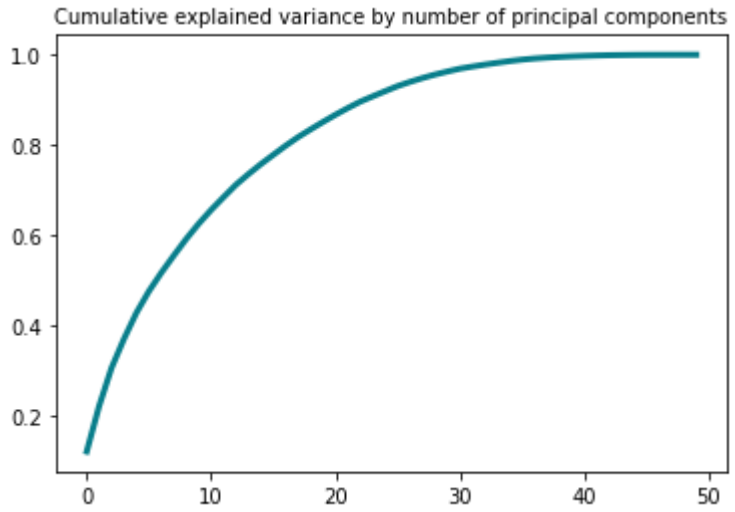
PORCPROD	0.222878
OEEMIND	0.190710
OEEPPPH	0.095206
OEEPRREAL	0.071011
OEEMFA1	0.070829
OEEMFA2	0.036063
OEEOBJ	0.035092
OEECNTOP	0.034472
OEETMP	0.030504
OEEEQUIPO	0.029999
OEEMF3	0.029917
OEETREALMI	0.026401
OEETCP	0.026223
OEEPRAC	0.023667
OEEPPS	0.021034

XGBoost Random Forest Hyperparameters

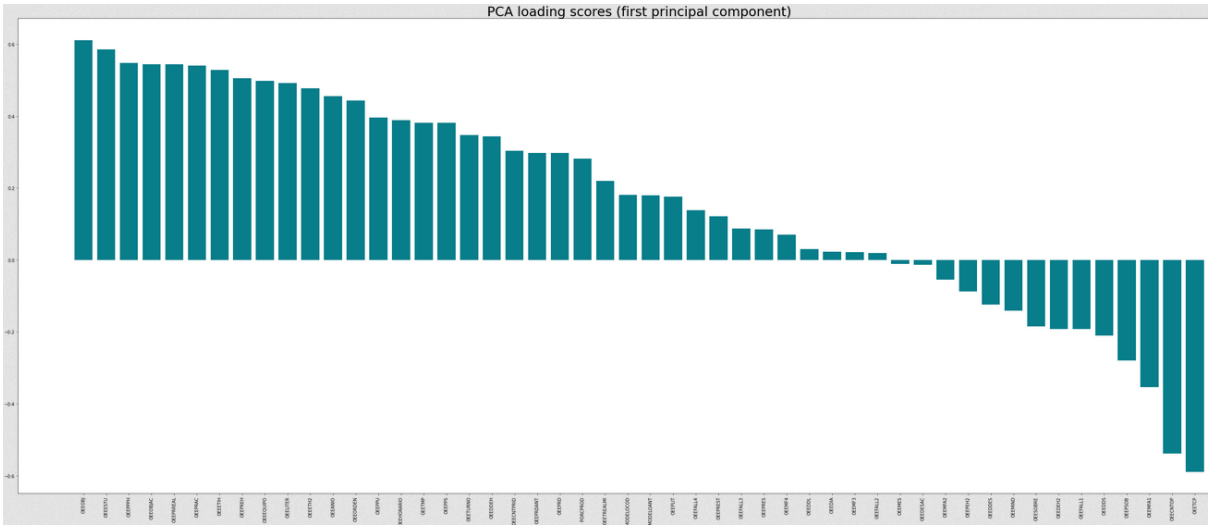
```
XGBoost Random Forest Hyperparameters
>10 0.997 (0.001)
>50 0.997 (0.001)
>100 0.997 (0.001)
>500 0.997 (0.001)
>1000 0.997 (0.001)
>5000 0.997 (0.001)
```



PCA - Principal component analysis

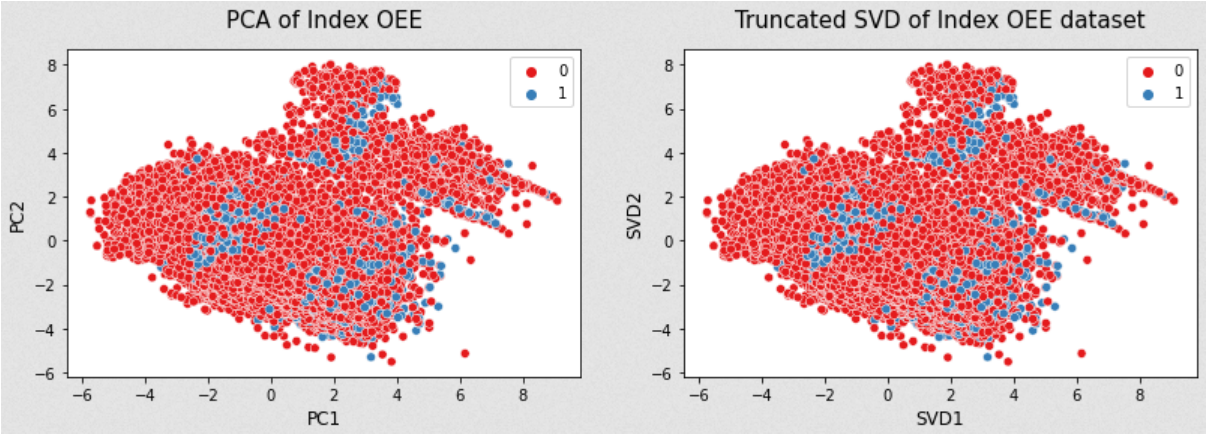


Podemos decir que los primeros 15 componentes contienen aproximadamente el 75% de la varianza

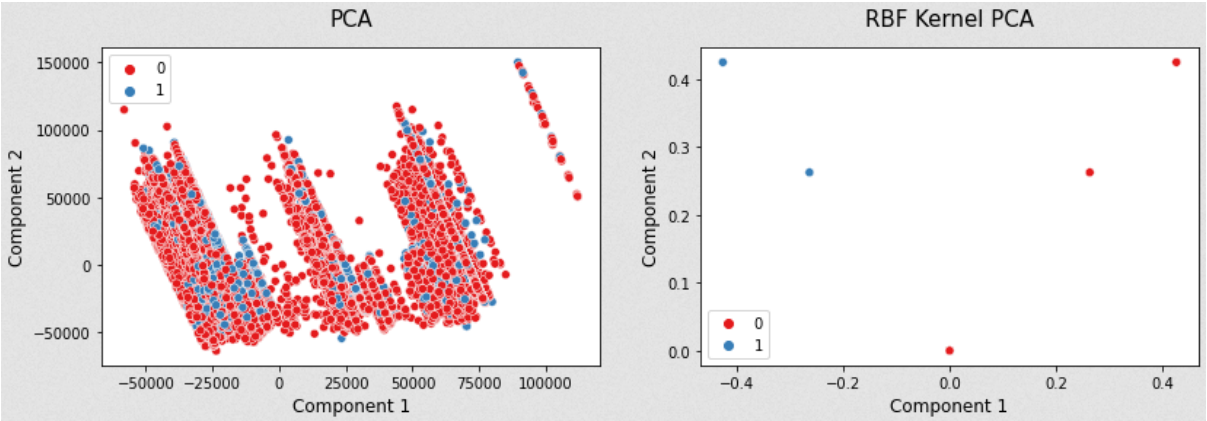


	Attribute	CorrelationWithPC1
0	OEEOBJ	0.611265
1	OEEESTU	0.584848
2	OEEPPPH	0.547162
3	OEEOBJAC	0.543746
4	OEEPRREAL	0.543492
5	OEEPRAC	0.540552
6	OEEETIH	0.527820
7	OEEPREH	0.505521
8	OEEEQUIPO	0.497914
9	OEELITER	0.491976
10	OEEETH2	0.477742
11	OEEANIO	0.455272
12	OEEORDEN	0.442567
13	OEEPPU	0.395885
14	OEEHORARIO	0.389038
15	OEETMP	0.380777

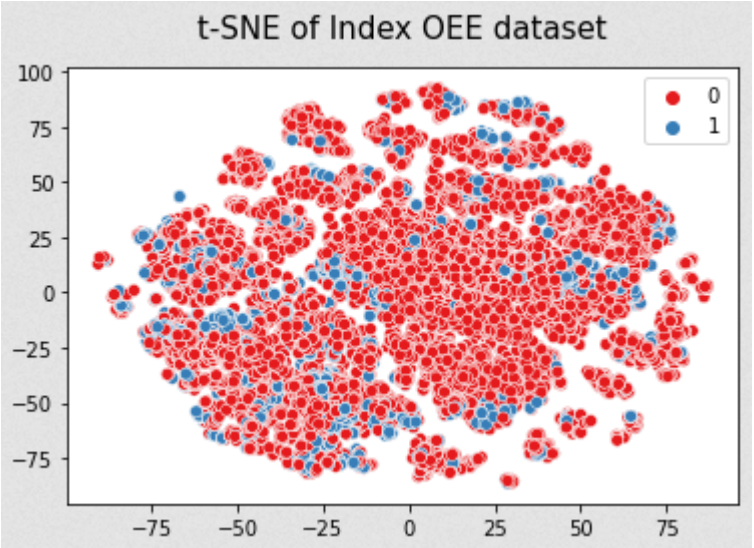
Truncated Singular Value Decomposition (SVD)



Kernel PCA



t-distributed Stochastic Neighbor Embedding (t-SNE)



Modelos

Me basé en los features selections que obtuve y seleccioné el siguiente subjconjunto:

'OEEEQUIPO',
'OEETREALMI',
'MODELOCOD',
'OEEPRD',
'MODELOANT',
'OEECNTOP',
'OEEDESAC',
'OEEPRREAL',
'PORCPROD',
'OEEMFA1',
'OEEMFA2',
'OEEMF3',
'OEEMF4',
'OEEMIND',
'OEETCP',
'OEEPPU',
'OEEPPS',
'OEETMP',
'OEEPPPH',
'OEEPREH',
'OEECNTPRID',
'OEEORDEN'

LinearRegression

Precisión del Nuevo modelo Reducido en TRAIN:
0.4727158072174794
Precisión del Nuevo modelo Reducido en TEST:
0.47838577311516106
DATOS DEL MODELO REGRESIÓN LINEAL MULTIPLE

Valor de las pendientes o coeficientes "a":
[-3.52499210e-02 1.09236756e-02 1.74778013e-05 -5.18982046e-07
  1.21588904e-04 1.09427700e-02 8.51416274e-03 -4.61754783e-05
  3.72292672e-02 -1.50391010e-02 -1.75513985e-02 -1.99649434e-02
 -2.10319387e-02 3.77583855e-03 -3.44633073e-01 6.58536267e-05
  3.84007466e-01 -2.30748591e+01 6.06318386e-05 -8.45018148e-03
 -2.96109932e-07 -2.57102847e-06]
Valor de la intersección o coeficiente "b":
1.252534341275608

StatsModel

Stats Models OLS Regression Results
==========
========== Dep. Variable: y R-squared (uncentered): 0.563 Model: OLS Adj. R-squared (uncentered): 0.562 Method: Least Squares F-statistic: 838.9 Date: Sun, 07 Aug 2022 Prob (F-statistic): 0.00 Time: 23:41:08 Log-Likelihood: -1827.3 No. Observations: 14349 AIC: 3699. Df Residuals: 14327 BIC: 3865. Df Model: 22 Covariance Type: nonrobust
==========
======== coef std err t P>|t| [0.025 0.975]
---------- x1 -0.0329 0.006 -5.092 0.000 -0.046 -0.020 x2 0.0110 0.001 18.796 0.000 0.010 0.012 x3 -2.644e-05 8.89e-05 -0.297 0.766 -0.000 0.000 x4 3.265e-08 1.97e-08 1.659 0.097 -5.92e-09 7.12e-08 x5 0.0001 8.9e-05 1.202 0.230 -6.75e-05 0.000 x6 0.0155 0.006 2.436 0.015 0.003 0.028 x7 0.0124 0.032 0.382 0.703 -0.051 0.076 x8 -4.673e-05 3.38e-06 -13.812 0.000 -5.34e-05 -4.01e-05 x9 0.0375 0.003 11.620 0.000 0.031 0.044 x10 -0.0150 0.002 -9.146 0.000 -0.018 -0.012 x11 -0.0176 0.002 -10.613 0.000 -0.021 -0.014 x12 -0.0200 0.002 -11.922 0.000 -0.023 -0.017 x13 -0.0211 0.002 -12.139 0.000 -0.024 -0.018 x14 0.0038 0.001 2.551 0.011 0.001 0.007 x15 -0.3610 0.038 -9.510 0.000 -0.435 -0.287 x16 6.89e-05 7.79e-06 8.849 0.000 5.36e-05 8.42e-05 x17 0.3806 0.195 1.954 0.051 -0.001 0.762 x18 -22.8724 11.689 -1.957 0.050 -45.785 0.040 x19 6.287e-05 1.58e-05 3.967 0.000 3.18e-05 9.39e-05 x20 -0.0139 0.003 -4.285 0.000 -0.020 -0.008 x21 -3.307e-07 9.48e-08 -3.488 0.000 -5.17e-07 -1.45e-07 x22 -4.85e-06 4.49e-06 -1.080 0.280 -1.36e-05 3.95e-06
==========
======== Omnibus: 384.598 Durbin-Watson: 1.994 Prob(Omnibus): 0.000 Jarque-Bera (JB): 305.789 Skew: 0.278 Prob(JB): 3.97e-67 Kurtosis: 2.550 Cond. No. 1.14e+10
==========
========

Logistic Regression

Precisión del Nuevo modelo Reducido en TRAIN:
0.8374799637605408
Precisión del Nuevo modelo Reducido en TEST:
0.8293718166383701
DATOS DEL MODELO REGRESIÓN LINEAL MULTIPLE

Valor de las pendientes o coeficientes "a":
[[-7.59317893e-07 -3.21642520e-06 2.02680166e-05 -1.32669288e-06
   2.01831381e-05 5.78968774e-07 -3.34903703e-09 3.13617706e-04
   3.74696809e-06 -9.72084495e-06 -4.26034105e-06 -2.80786810e-06
  -1.38299267e-06 -2.22294331e-05 1.61428925e-07 -9.89276439e-04
  -2.44108131e-04 -4.06850150e-06 8.82490072e-05 -7.11645633e-08
   1.77752059e-06 -9.09084411e-05]]
Valor de la intersección o coeficiente "b":
[-9.35749746e-10]