APRENDIZAJE DE MÁQUINA I

TRABAJO PRÁCTICO GRUPAL

ALAN CHURICHI
JUAN PABLO ALIANAK

ENTENDIENDO EL PROBLEMA

Las personas son el mayor activo que tiene una empresa, gestionar los recursos humanos de manera efectiva resulta imprescindible para evitar su fuga. En este trabajo analizaremos el dataset “HR-Employee-Attrition” con un modelo de clasificación binaria.

EMPLEADOS A GUSTO

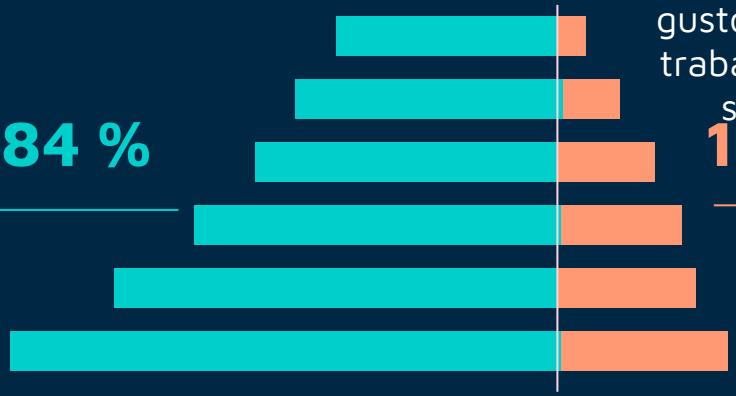
La mayor parte de los empleados se encuentra a gusto.

84 %

EMPLEADOS CON DESGASTE

Si bien la mayor parte de los empleados se encuentran a gusto, es importante detectar y trabajar sobre las personas que se encuentran desgastadas.

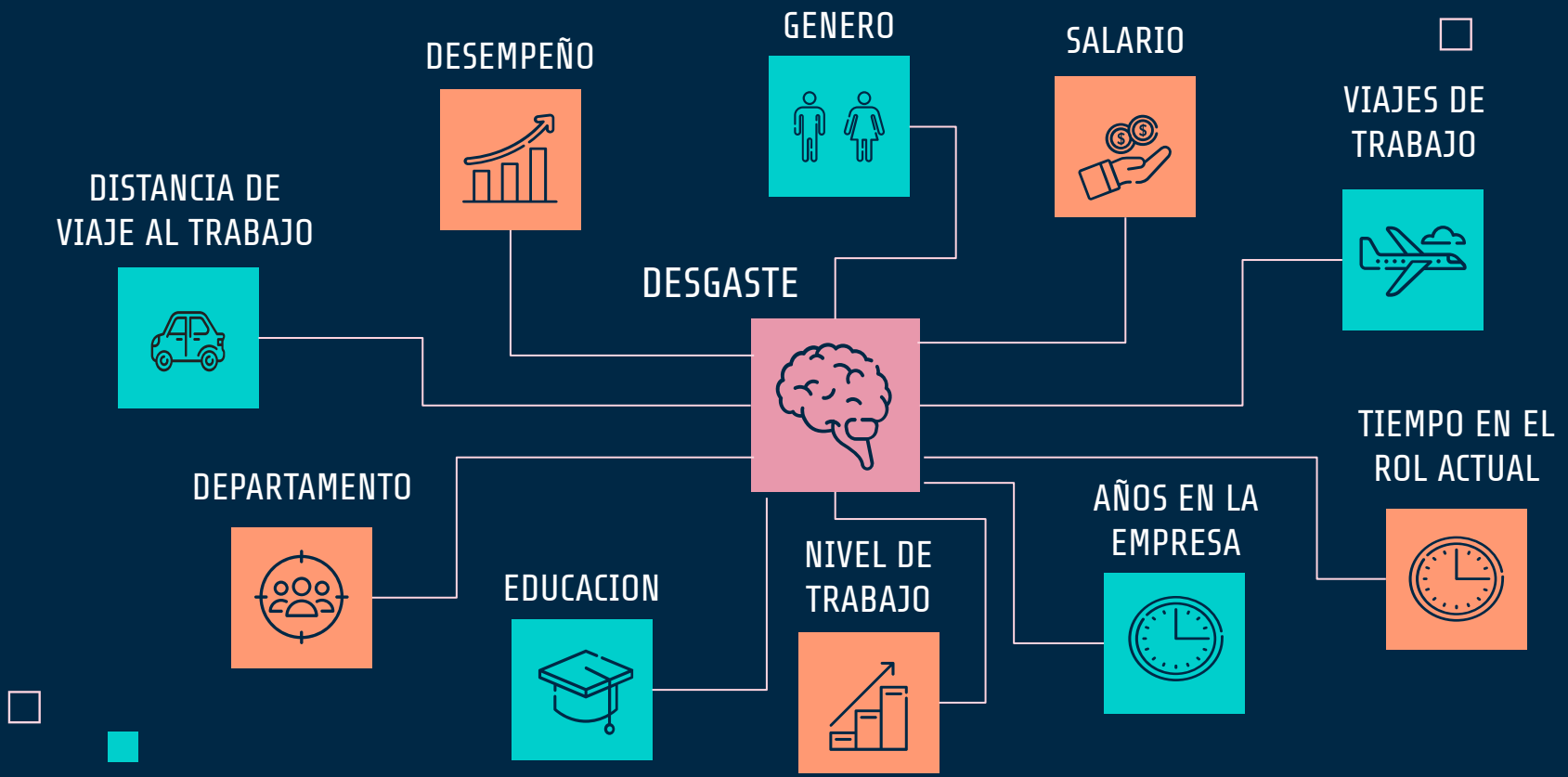
16 %



CONTENIDO DEL TRABAJO PRÁCTICO

1. Análisis preliminar de las feature del dataset.
2. Análisis de la distribución de cada una de las feature (categóricas y numéricas).
3. Reducción de feature en base al análisis anterior.
4. Transformación de los datos para ingresarlos a los modelos.
5. Implementación de los diferentes modelos:
    - **Decision Tree**
    - **Random Forest**
    - **SVM linear**
    - **SVM non-linear**
    - **Logistic Regression**
    - **KNN**
6. Análisis de resultados.
7. Optimización de hiperparámetros de los modelos con mejor performance.
8. Conclusiones

ALGUNAS DE LAS FEATURE CONTENIDAS EN EL DATASET



COMPOSICION DEL DATASET

Tamaño inicial del dataset = 34 feature x 1470 muestras
Variable Objetivo = "Attrition"

19

Feature categoricas

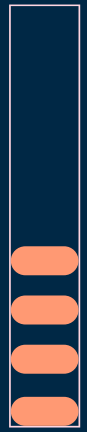
15

Feature numericas

Después de analizar y graficar las distintas feature, descartamos las constantes o cuasi constantes

16

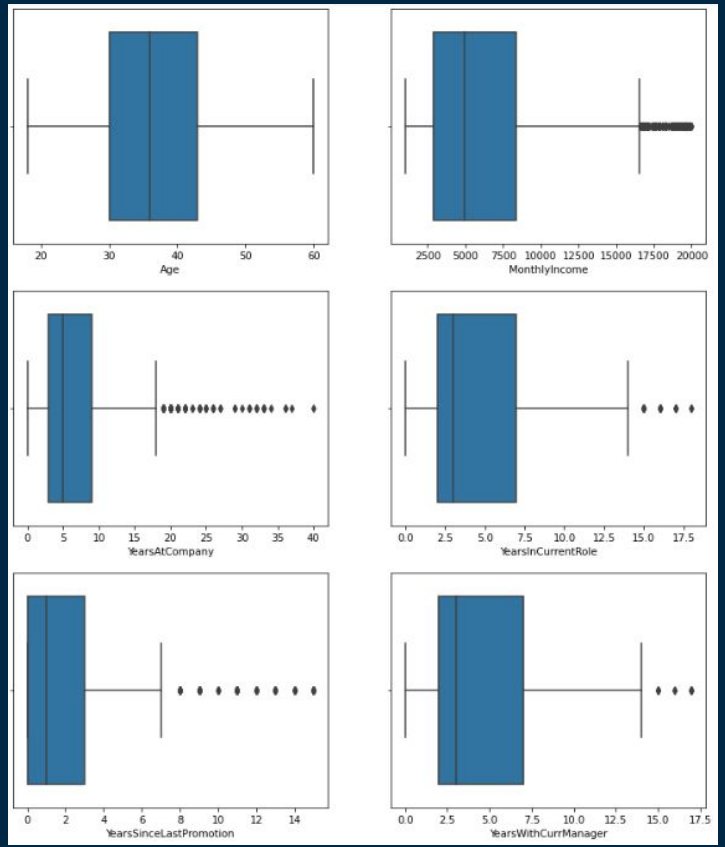
Feature categoricas

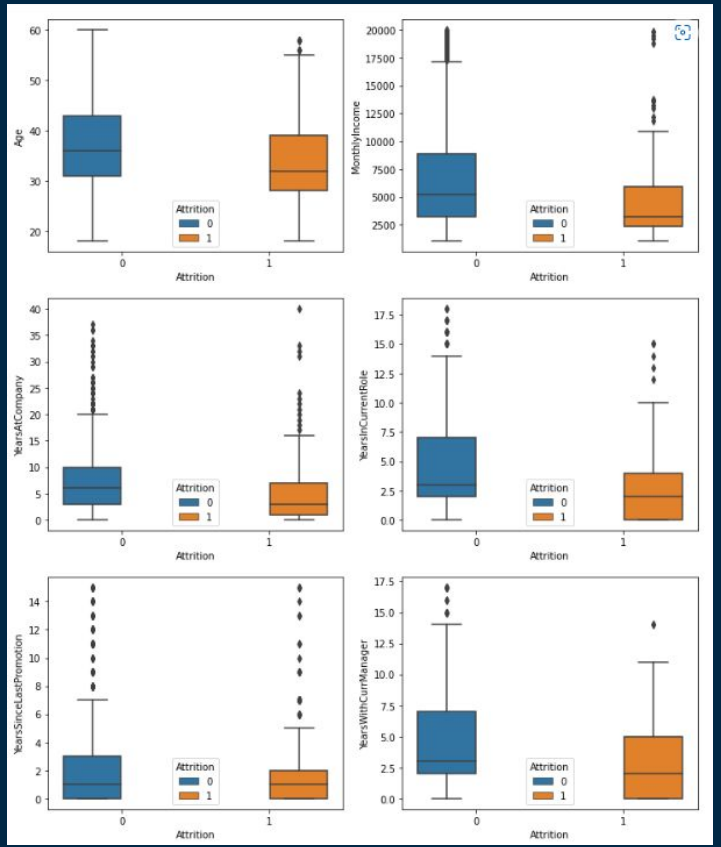


11

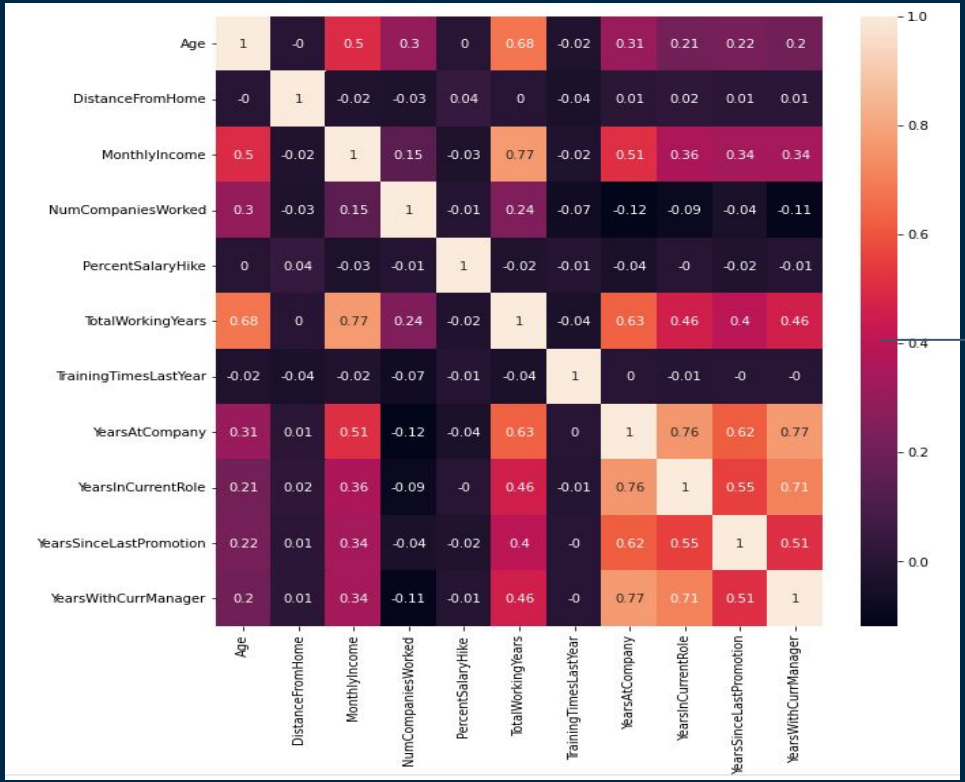
Feature numericas

BOXPLOT DE ALGUNAS FEATURES NUMÉRICAS



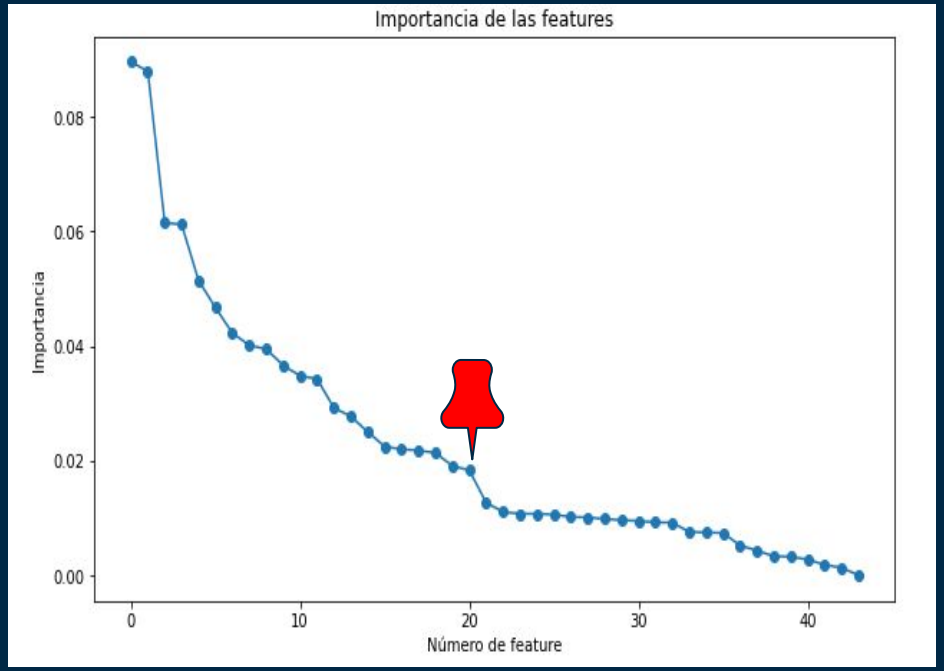


MATRIZ DE CORRELACIÓN DE LAS FEATURE NUMÉRICAS



Si bien se observan algunas feature con correlaciones altas, no vemos que ninguna sea un transformación lineal de otra, por lo tanto decidimos incluir todas las features dentro del modelo.

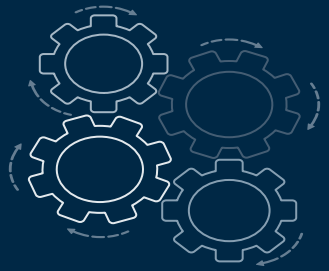
ANÁLISIS DE IMPORTANCIA DE LAS FEATURES



Analizando la gráfica de importancia de las features decidimos quedarnos con las 20 de mayor importancia a la hora de predecir. Las primeras son:

1. MonthlyIncome
2. Age
3. DistanceFromHome
4. YearsAtCompany
5. TotalWorkingYears
6. YearsWithCurrManager
7. NumCompaniesWorked
8. OverTime
9. YearsInCurrentRole
10. TrainingTimesLastYear
11. ……

MODELOS IMPLEMENTADOS



01 SVM LINEAR

02 SVM NON-LINEAR

03 KNN

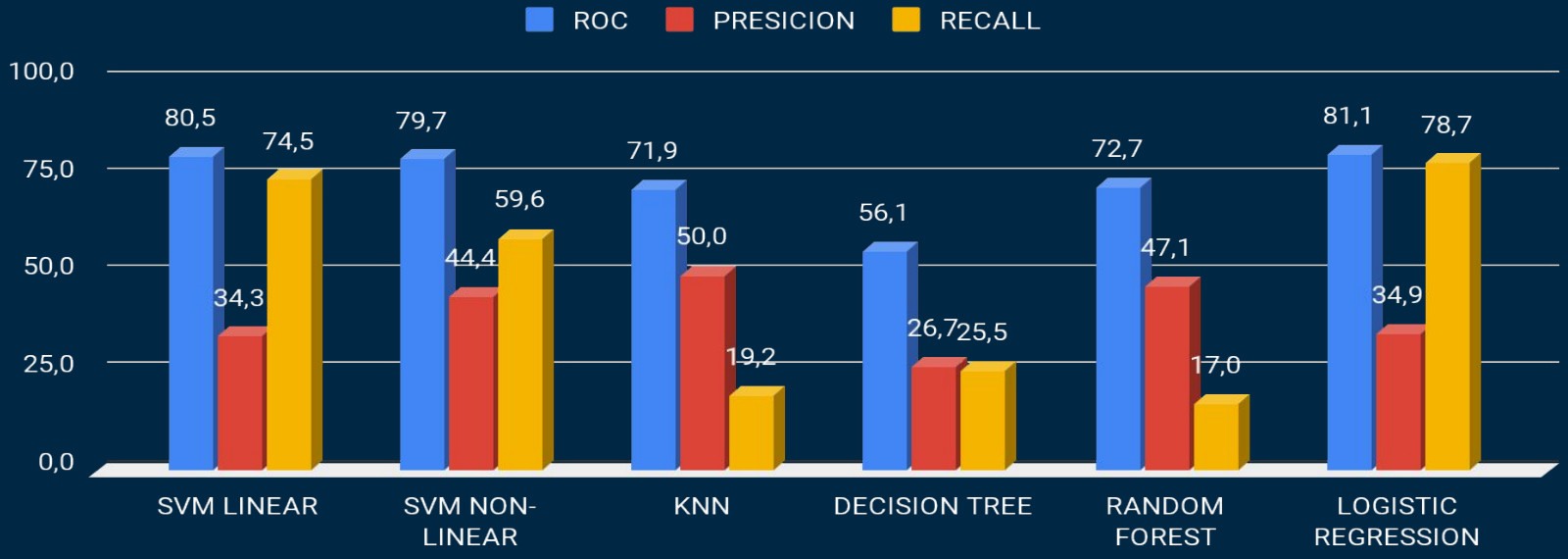
04 DECISION TREE

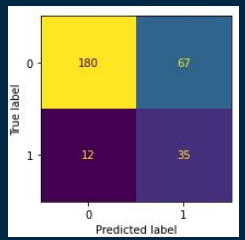
05 RANDOM FOREST

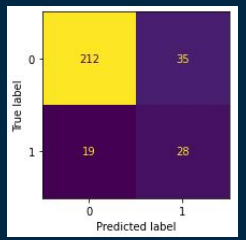
06 LOGISTIC REGRESSION

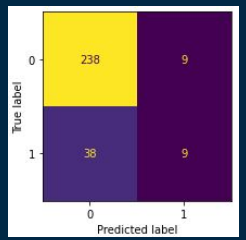
$$\text{RECALL} = \frac{TP}{TP + FN}$$

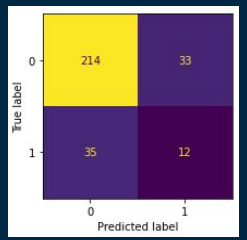
PRIMEROS RESULTADOS

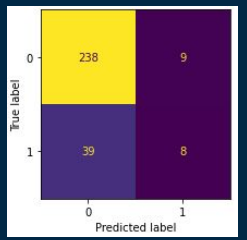


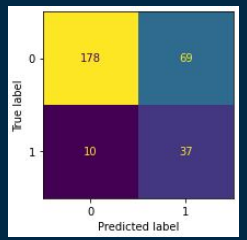




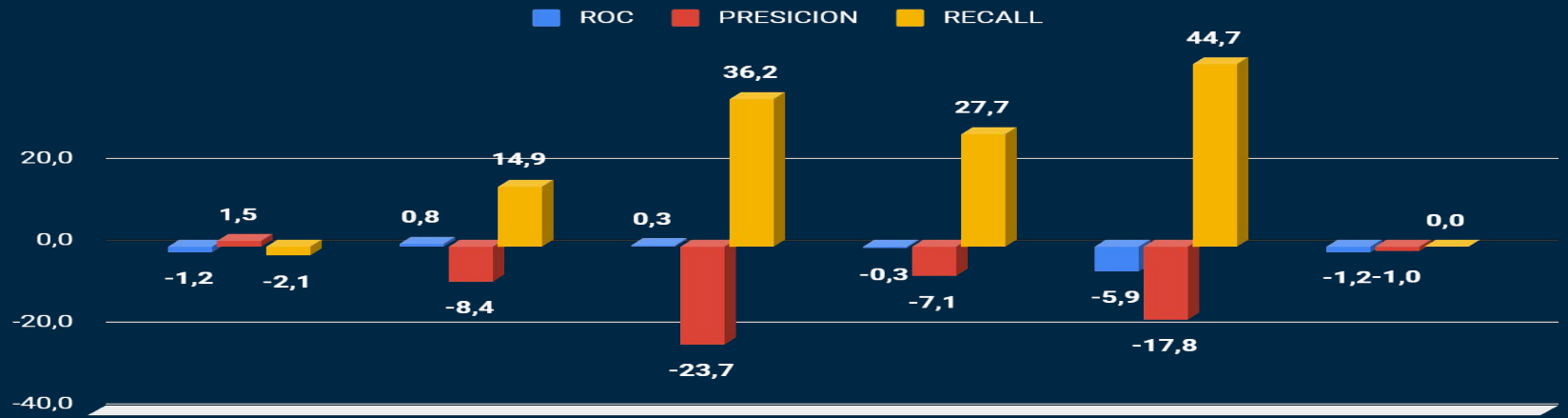


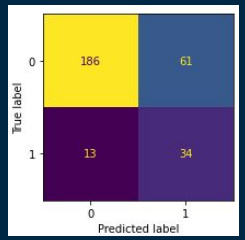


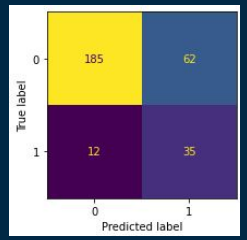


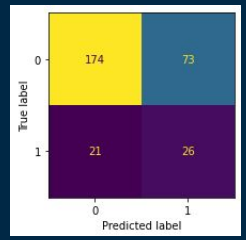


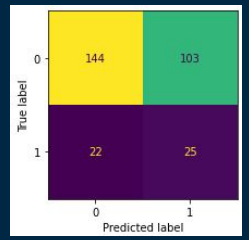
RESULTADOS CON UNDER SAMPLING

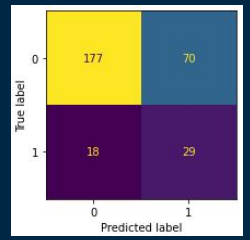


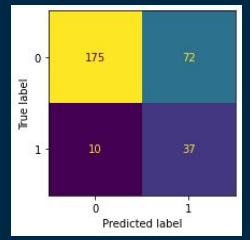






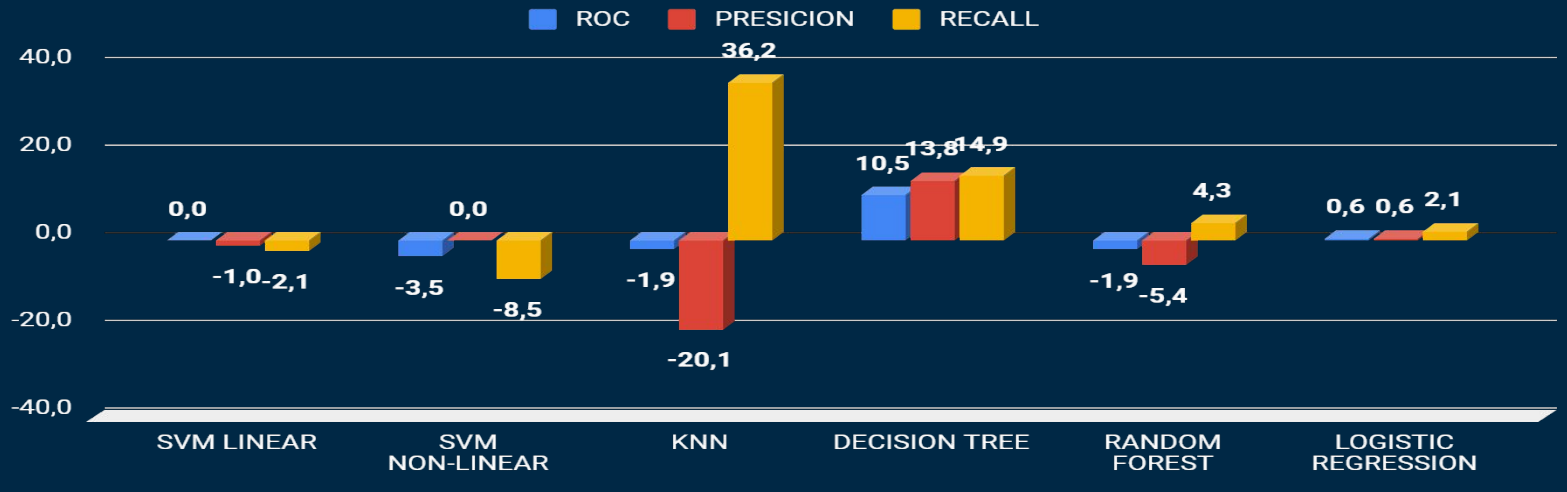


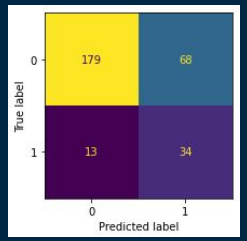




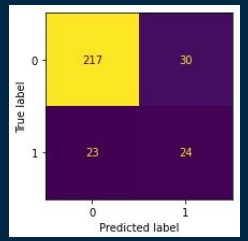
RC = 72.34 | RC = 74.47 | RC = 55.32 | RC = 46.81 | RC = 61.70 | RC = 78.72

RESULTADOS CON OVER SAMPLING

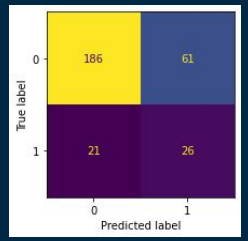




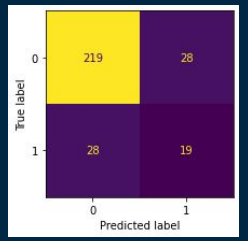
RC = 72.34



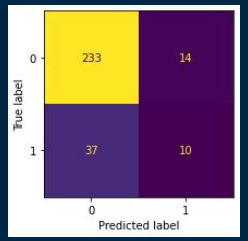
RC = 51.06



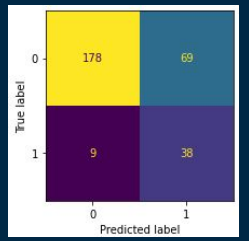
RC = 55.32



RC = 36.17



RC = 21.28



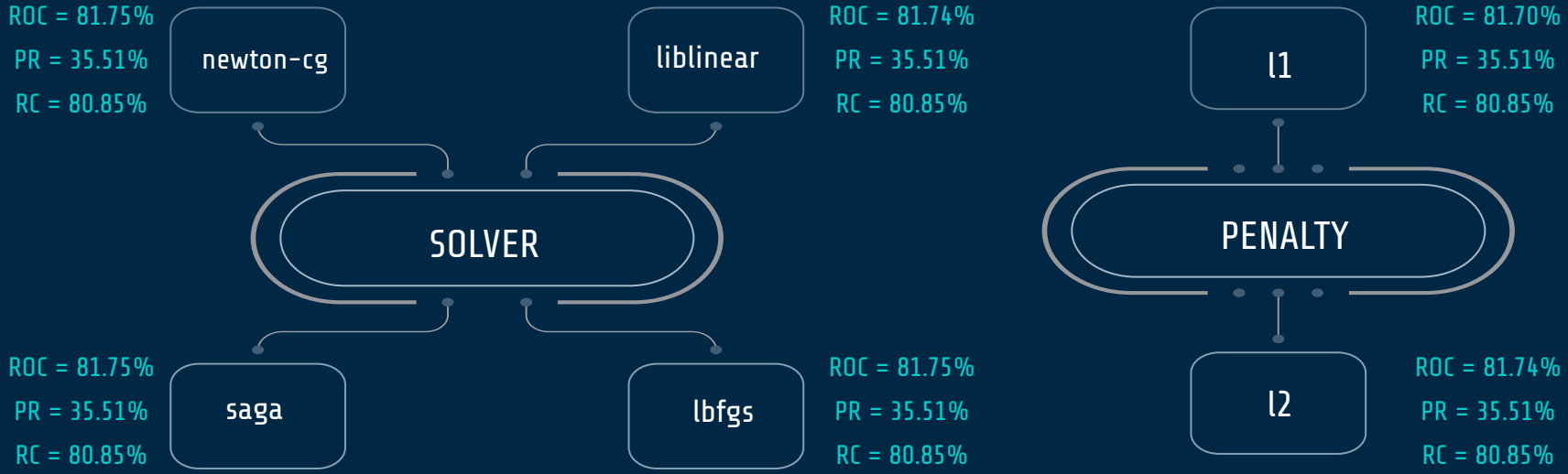
RC = 80.85

MODELOS CON MEJOR PERFORMANCE

	ROC	PRECISION	RECALL	DATOS ENT.
LOGISTIC REGRESSION	81.75%	35.51%	80.85%	OVER SAMPLING
SVM LINEAR	80.49%	34.31%	74.47%	SIN BALANCEAR

OPTIMIZACIÓN DE HIPERPARAMETROS

MODELO: Logistic Regression



Las diferencias observadas en las métricas son despreciables, el modelo no es sensible a estos hiperparametros

CONCLUSIONES

Concluimos en que el mejor modelo para predecir Attrition es el de Regresión Logística utilizando oversampling en los datos de entrenamiento para tratar el problema de las clases desbalanceadas. El segundo modelo que mejores resultados da es el SVM con kernel lineal.

Durante en análisis de datos vimos que la variable a predecir Attrition no presentaba fuerte correlación con las demás features, intuíamos que no iba a ser un problema fácil de resolver. Sin embargo, logramos un modelo con un grado de precisión aceptable .

Sin duda este modelo podría aplicarse en un entorno real para solucionar un problema como el planteado.

GRACIAS