



"E-Detector"

Optimización en la Selección de Personal Destacado

TC2004B.200

"Análisis de Ciencia de Datos"

Profa. Ma. Angelina Alarcón Romero

Profa. María de los Angeles Constantino González

```
$(function(){cards();});  
$(window).on('resize', function(){  
    cards();  
    var width = $(window).width();  
    if(width < 750){  
        cardssmallscreen();  
    } else {  
        cardsbigscreen();  
    }  
});  
function cardssmallscreen(){  
    var cards = $('.card');  
    var height = 0;  
    var card2 = 2;  
    var card1 = 1;  
    var of_type = 'c';
```

Equipo 007



Andres Saldaña
Rodriguez

A01721193

Rol: Científico
de Datos /
Ing. de Datos



Karen González
Ugalde

A01411597

Rol: Científico
de Datos /
Ing. de Datos



Fernando
Gonzalez Rosas

A01253694

Rol: Científico
de Datos /
Ing. de Datos



José Eugenio
Morales Ortiz

A01734612

Rol: Científico de
Datos /
Project Manager



Daniel Sánchez
Villarreal

A01197699

Rol: Científico
de Datos /
Ing. de Datos

• Problemática

La empresa Ternium realiza su proceso de selección de nuevos aplicantes destacados de manera manual y mecánica, lo cual consume mucho tiempo.

• Solución Propuesta

Diseñar una herramienta para poder optimizar la selección del personal destacado y minimizar el tiempo de identificación.

Crear una nueva base de datos en la que sea más fácil visualizar la información y una nueva página web que determine si un candidato es destacado o no.

• Objetivo

Disminuir el tiempo de selección e identificación de personal destacado, además de poder asegurar que dicho personal sea sobresaliente.



1

Datos proporcionados por Ternium

```
[ ] df = pd.read_csv("BD_RIIIE_2021V3.csv")
df.head()
```

	País	ID Candidato	Género	Carrera Gestional	Avance	Semestres Totales	Postulados Si/No	Evaluados Si/No	Altamente Recomendado	Operaciones-Calidad	MTTO-DIMA	Comercial-Planeamiento	DIGI-SC
0	México	1.012101e+09	F	Ing. Industrial	NaN	NaN	Si	Si	Si	Highly Recommend	Recommend	Highly Recommend	Do Not Recommend
1	México	1.012101e+09	F	Ing. Industrial	6	9	Si	Si	Si	Highly Recommend	Do Not Recommend	Do Not Recommend	Do Not Recommend
2	México	1.012101e+09	F	Ing. Industrial	8	9	Si	NaN	NaN	NaN	NaN	NaN	NaN
3	México	1.012101e+09	F	Ing. Mecatrónica/Electrónica	6	9	Si	Si	Si	Highly Recommend	Highly Recommend	Do Not Recommend	Recommend
4	México	1.012101e+09	F	Negocios Internacionales	NaN	NaN	Si	Si	NaN	NaN	NaN	NaN	NaN



2

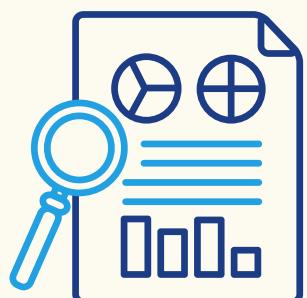
Limpieza y Transformación de datos

- Igualamos valores repetidos.
- Convertimos "ID Candidato" a string.
- Remplazamos ID existentes y asignamos a los que no tenían.
- Nueva columna que junta "Apto" y "Destacado" en una sola.
- Eliminamos columna "País".
- "Avance" y "Semestres Totales" de object a float.
- Cambiamos las 5 columnas de las pruebas pymetrics, de variables categóricas a int.
- Cambiamos "Apto/No Apto", "Destacado Pym", "Ingles", "Ingresados Si/No" de float a int.
- Remplazamos los NaN por 0.

```
df['Apto/No Apto'] = df['Apto/No Apto'].replace({'Apto': 1, 'No Apto' : 0})
```

```
df['Apto/ Destacado'] = df['Apto/ Destacado'].replace({'Apto': 1, 'Destacado' : 2})
```

```
df['Ingresados Si/No'] = df['Ingresados Si/No'].replace({'Si': 1})
df['Ingresados Si/No'] = df['Ingresados Si/No'].fillna(0)
df['Apto/No Apto'] = df['Apto/No Apto'].fillna(0)
df['Ingles'] = df['Ingles'].fillna(0)
df['Apto/ Destacado'] = df['Apto/ Destacado'].fillna(0)
```



Hipótesis

(Confirmación)

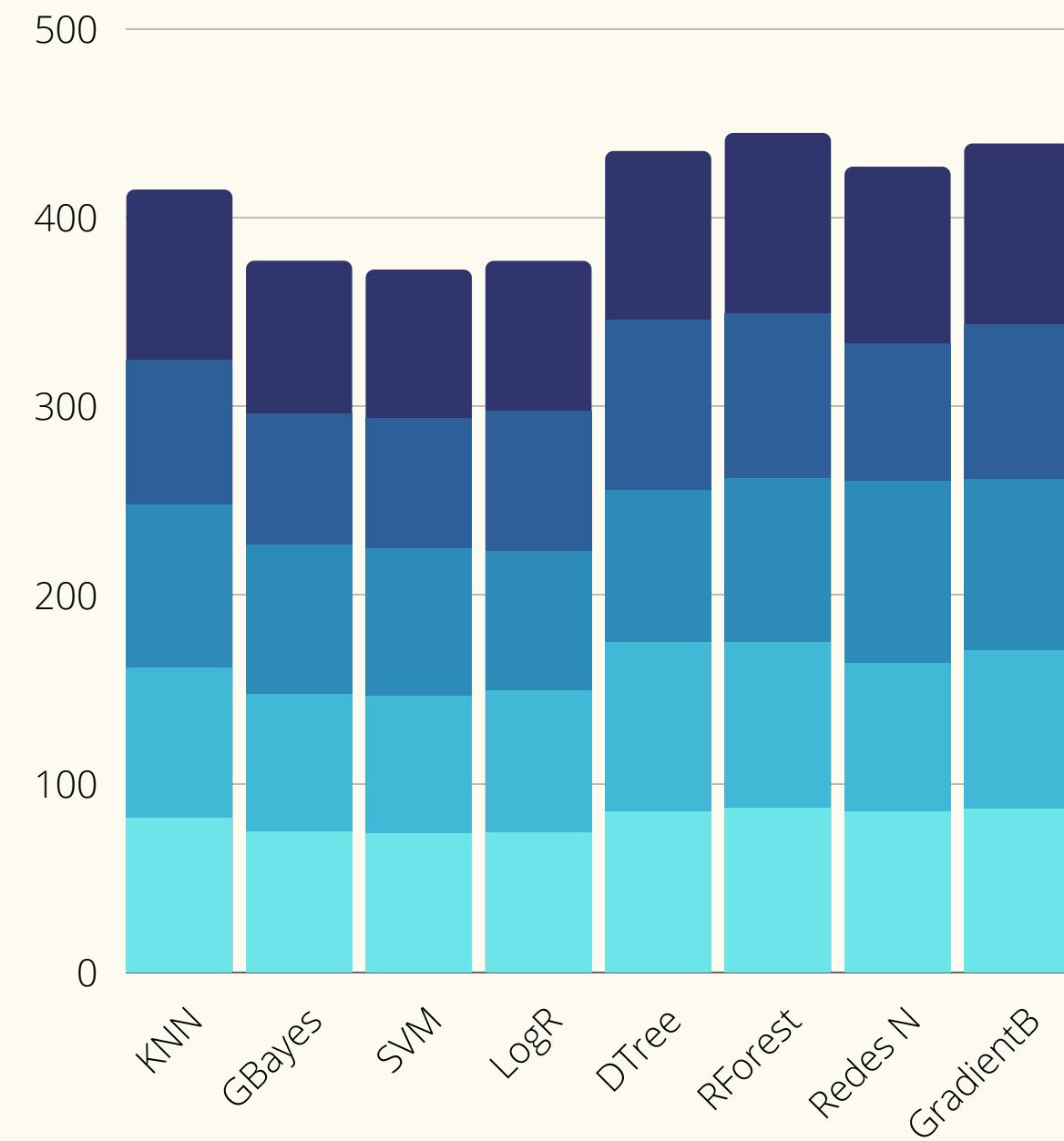
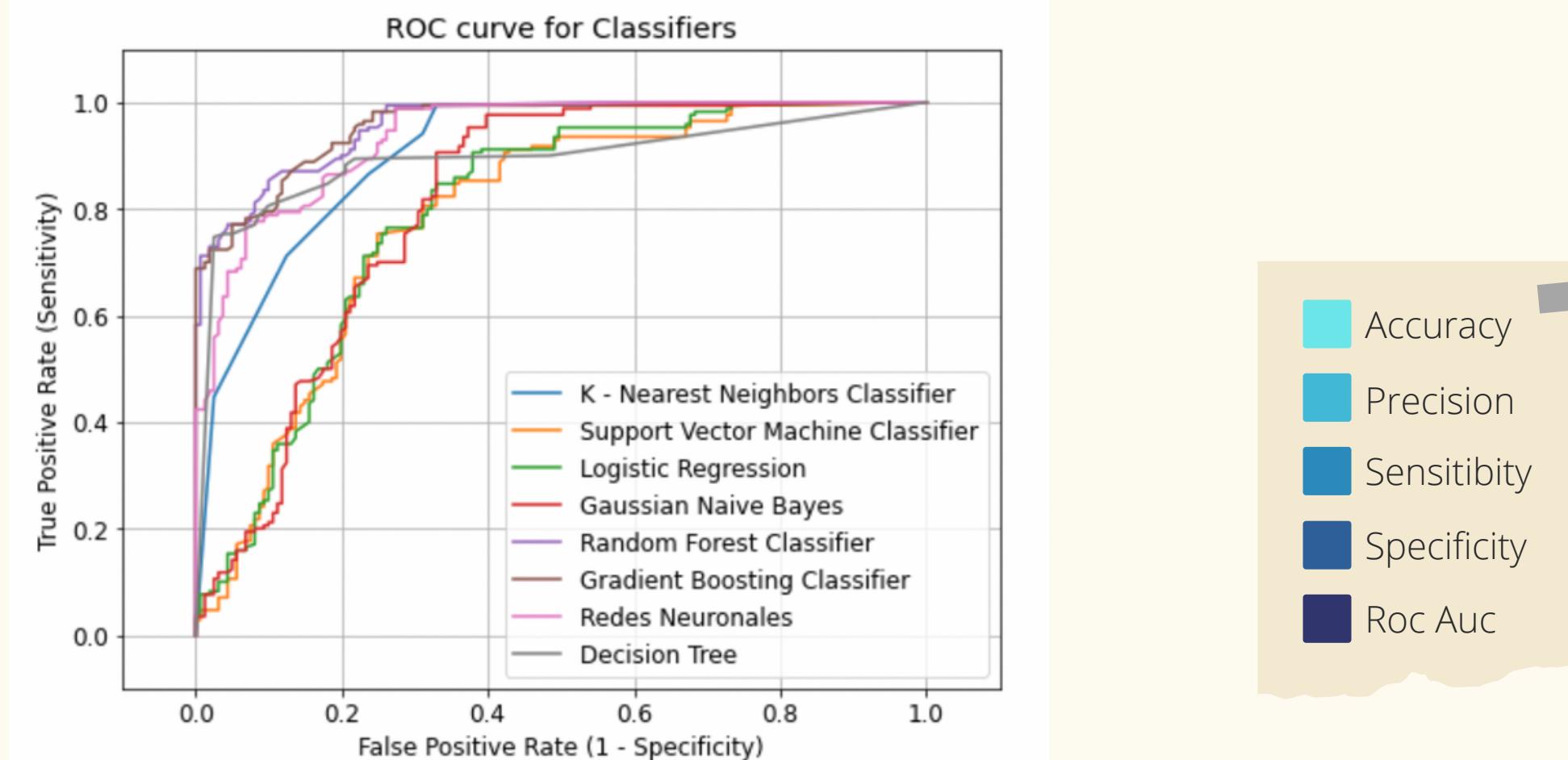
Lo que se busca tener como resultado final es el poder desarrollar un nuevo método para el área de recursos humanos, que les permita poder realizar el proceso de selección de personal nuevo destacado y altamente calificado de una manera más automatizada y sencilla de analizar e interpretar.

Planeamos lograrlo utilizando diferentes conceptos y métodos conocidos en la ciencia de datos, tales como el análisis descriptivo, análisis predictivo, correlación de variables, arquitectura de bases de datos, Business Intelligence; los cuales consideramos que nos pueden ser de utilidad para la principal meta que se busca resolver en esta problemática, la cual es la facilitación en la toma de decisiones en la selección de nuevo personal.

Modelos

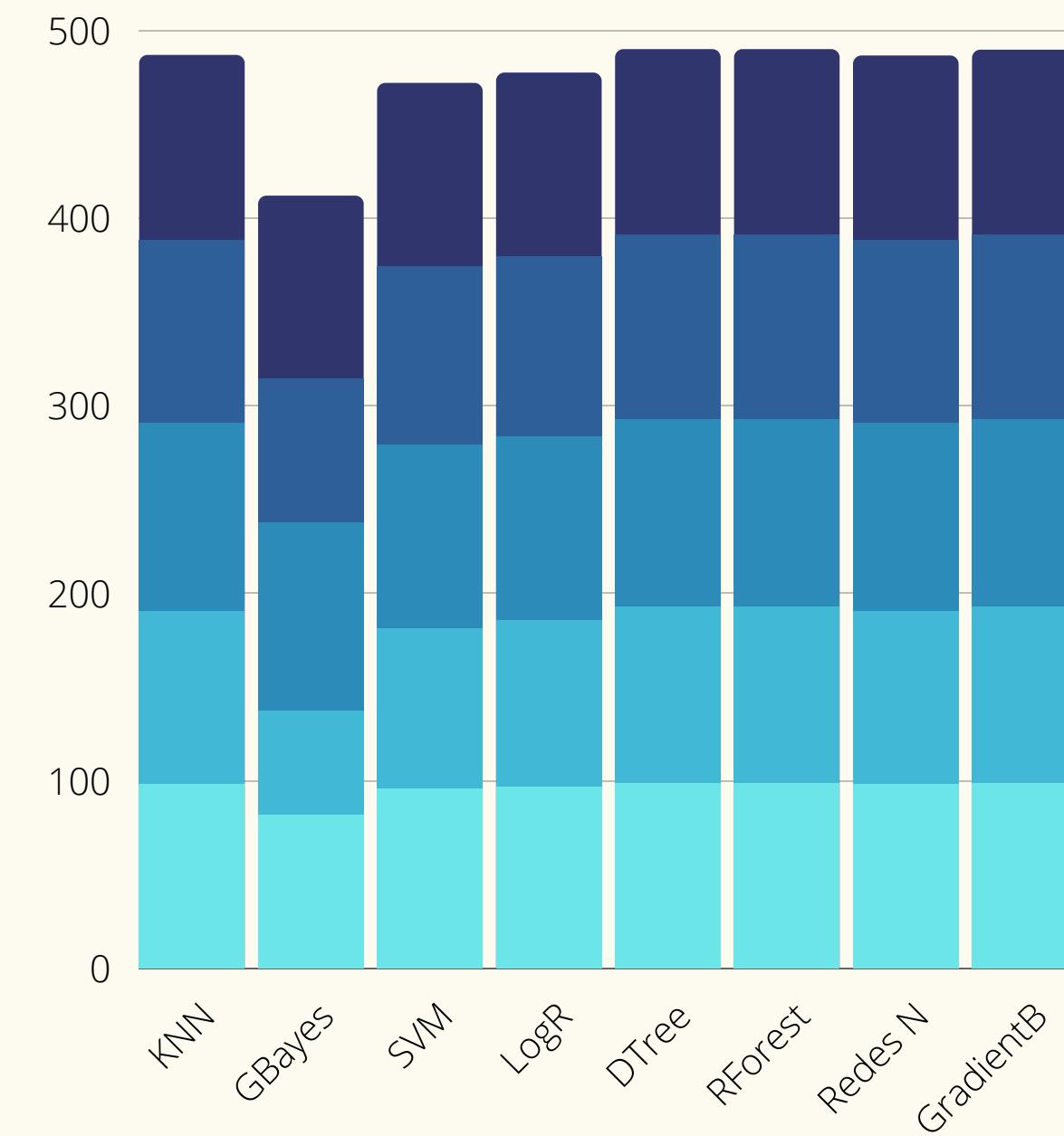
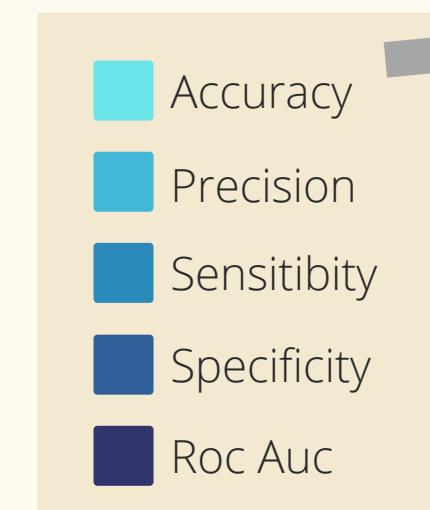
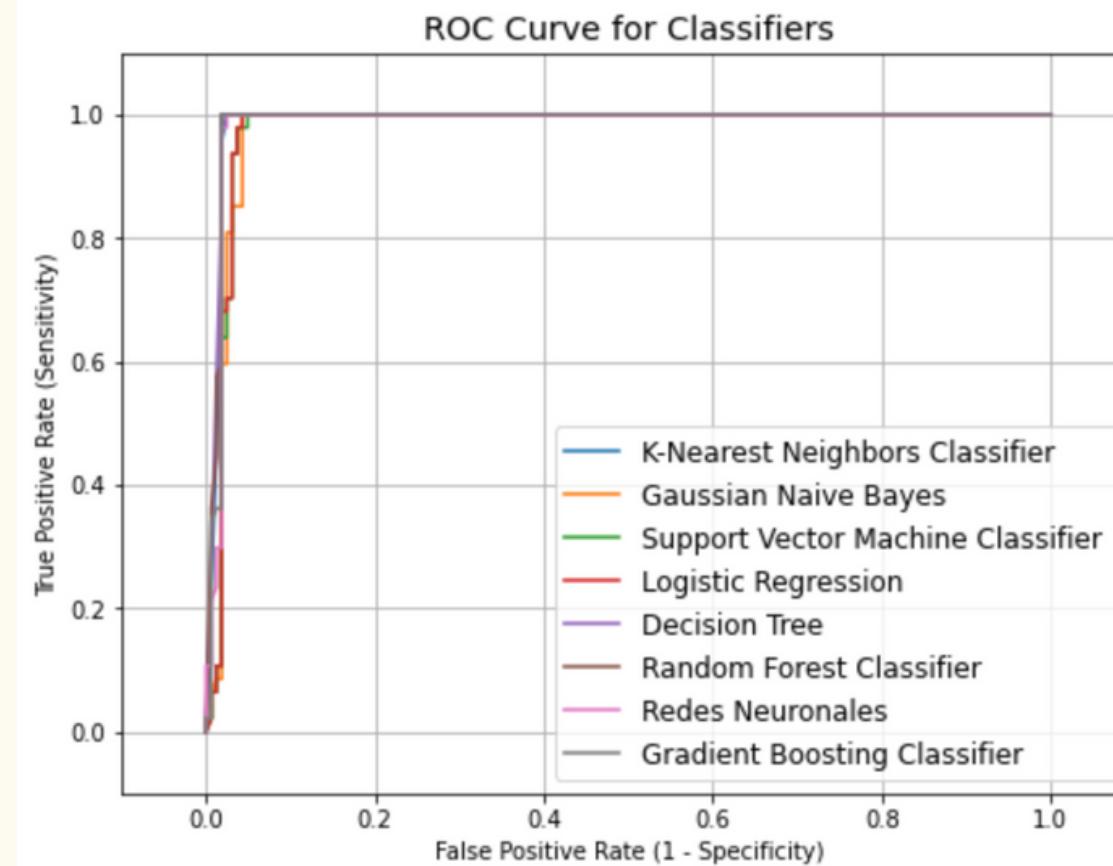
(sin "Apto/No Apto")

	KNN	SVM	LogR	Bayes	RandF	GB	RedesN	Dtree
Accuracy:	0.816	0.737	0.74	0.743	0.87	0.864	0.849	0.852
Precision:	0.795	0.727	0.753	0.732	0.876	0.842	0.788	0.895
Sensitivity:	0.865	0.782	0.735	0.788	0.871	0.906	0.965	0.806
Specificity:	0.764	0.689	0.745	0.696	0.87	0.82	0.727	0.901
Roc Auc Score:	0.905	0.786	0.794	0.809	0.958	0.957	0.937	0.895
Total:	4.145	3.721	3.721	3.768	4.445	4.389	4.266	4.349



Modelos

	KNN	GBayes	SVM	LogR	DTree	RForest	RedesN	GradientB
Accuracy:	0.981	0.819	0.957	0.967	0.986	0.986	0.981	0.986
Precision:	0.922	0.553	0.852	0.885	0.94	0.94	0.922	0.94
Sensitivity:	1	1	0.979	0.979	1	1	1	1
Specificity:	0.975	0.767	0.951	0.963	0.982	0.982	0.975	0.982
Roc Auc Score:	0.988	0.977	0.978	0.978	0.989	0.989	0.985	0.986



Bosques Aleatorios

Ajuste de Hiperparámetros

```
▶ from sklearn.ensemble import RandomForestClassifier
param_grid = {
    'n_estimators': [10,50,100],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [3, 4,5,6,7, 8],
    'criterion' :['gini', 'entropy']
}
rfc =RandomForestClassifier(random_state=29)
rfc_gscv=GridSearchCV(estimator=rfc, param_grid=param_grid, scoring="roc_auc") # cv default = 5
rfc_gscv.fit(X_,y_.values.ravel())

print("Tuned best parameters for random forest: ",rfc_gscv.best_params_)
print("Best score: {}".format(rfc_gscv.best_score_))

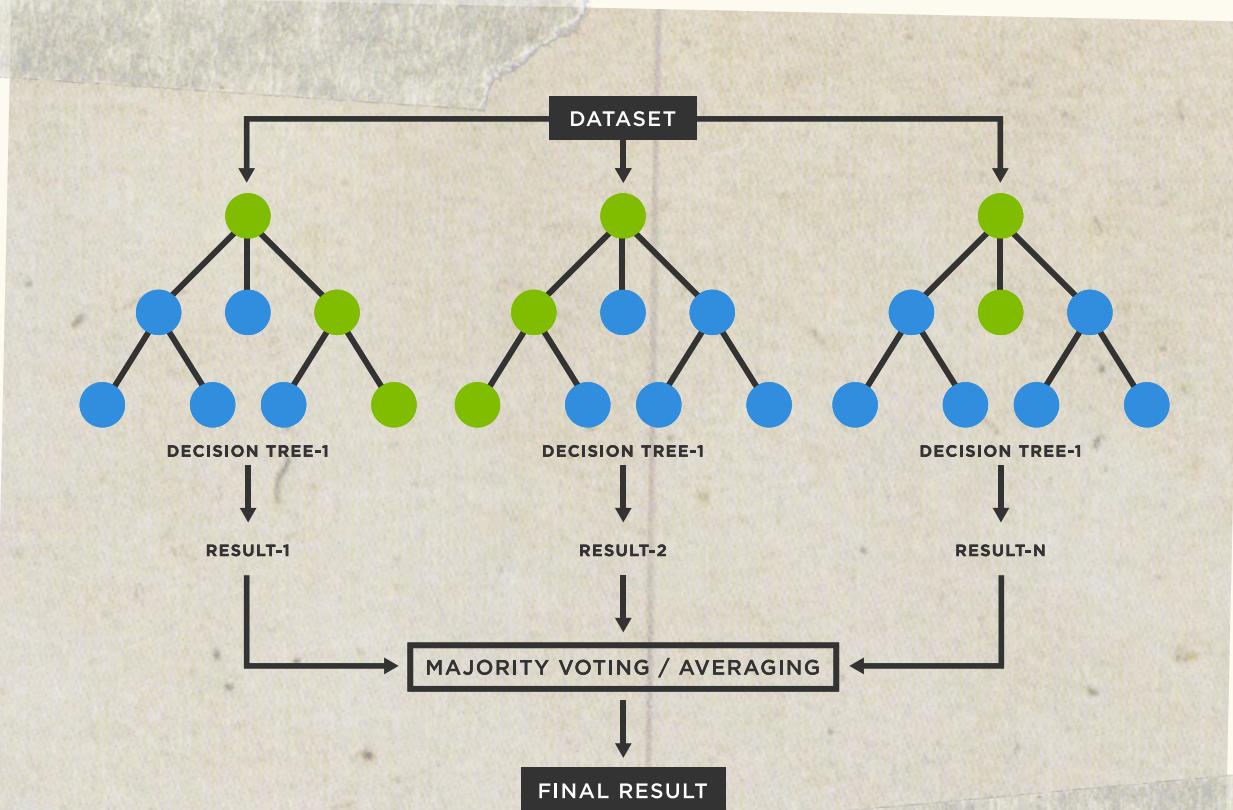
⇒ Tuned best parameters for random forest: {'criterion': 'gini', 'max_depth': 8, 'max_features': 'auto', 'n_estimators': 50}
Best score: 0.9373808401650681
```

Exactitud: 0.8942598187311178
Error de clasificación: 0.10574018126888217
Sensitividad 0.949685534591195
Especificidad: 0.8430232558139535
Falsos positivos: 0.1569767441860465
Precisión: 0.848314606741573
F1: 0.8961424332344213

RESULTADOS

Criterios de Exito

- Algoritmo optimo
- Algoritmo Eficiente
- Algoritmo Preciso
- Algoritmo Sensible
- Fácil de implementar



Hola, favor de introducir los resultados de las pruebas Pymetrics del aplicante y nosotros determinaremos si es un postulado destacado o no

Recuerda que 1 es Destacado y 0 es No Destacado

Introduzca resultados de prueba Operaciones-Calidad

Recommend

Introduzca resultados de prueba MTTO-DIMA

Recommend

Introduzca resultados de prueba Comercial-Planeamiento

Highly Recommend

Introduzca resultados de prueba DIGI-SC

Do Not Recommend

Introduzca resultados de prueba Resto-Soft

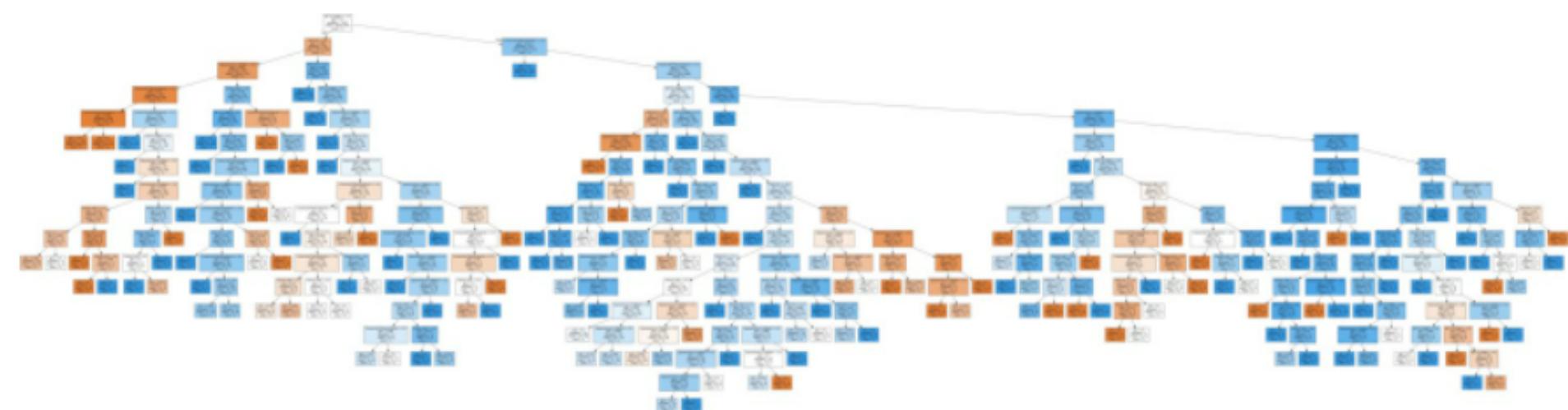
Do Not Recommend

Submit

El aplicante es: 1

Arbol de decisión del modelo

Con ayuda del modelo de arboles de decisión, nos podemos dar una idea de como es que se verían nuestros arboles de decisión



Representación grafica del modelo de arboles de decisión

Prototipo Funcional



Recomendaciones a Socio Formador

- A Investigación y creación de un departamento de análisis de datos
- B Implementación de tecnologías de la industria 4.0
- C Creación de un diccionario para las bases de datos
- D Creación de estándares para la captura de datos



Siguientes Pasos

- A Recolección de más datos
- B Seguimiento al proyecto
- C Explorar nuevos posibles proyectos



¡Gracias!

Por su tiempo
Por su disposición
Por escucharnos
Por prestarnos su base de datos
Por querer trabajar con nosotros

¡Nuestro Aprendizaje!

La gran cantidad de tiempo que toma limpiar los datos
Diferentes modelos de clasificación
Métricas de evaluación
Balanceo de datos
Bases de datos reales