



Estudio del abandono desde la perspectiva de los foros

David Castillejo Camacho

MASTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS

U.N.E.D.

Junio 2022

Tutor: Álvaro Rodrigo Yuste



ÍNDICE

01. INTRODUCCIÓN

02. LECTURA Y ANÁLISIS DE LOS DATOS

03. MODELOS DE CLASIFICACIÓN BASADOS SÓLO EN TEXTO

04. MODELOS DE CLASIFICACIÓN BASADOS EN INFORMACIÓN SOCIAL

05. CONCLUSIONES Y PASOS FUTUROS

FOROS DE ALUMNOS DE LA U.N.E.D



**Registro de la
comunicación**

**Registro de
las notas de los
alumnos**



ESTUDIAR A DISTANCIA.

- Incremento de la Educación a distancia.
- La UNED es pionera en la modalidad de educación eLearning.
- Desventaja: Pérdida de la relación cara a cara.
- Foros: vía de transmisión del conocimiento.

FOROS = REGISTRO DE LA ACTIVIDAD DEL CURSO

Para este proyecto se han utilizado 4 colecciones de mensajes correspondientes a los cursos entre 2017 y 2020.



**DETECCIÓN DEL
ABANDONO**

=

**OPTIMIZACIÓN
DE RECURSOS**

OBJETIVO DEL TFM

OBJETIVO DEL TFM.

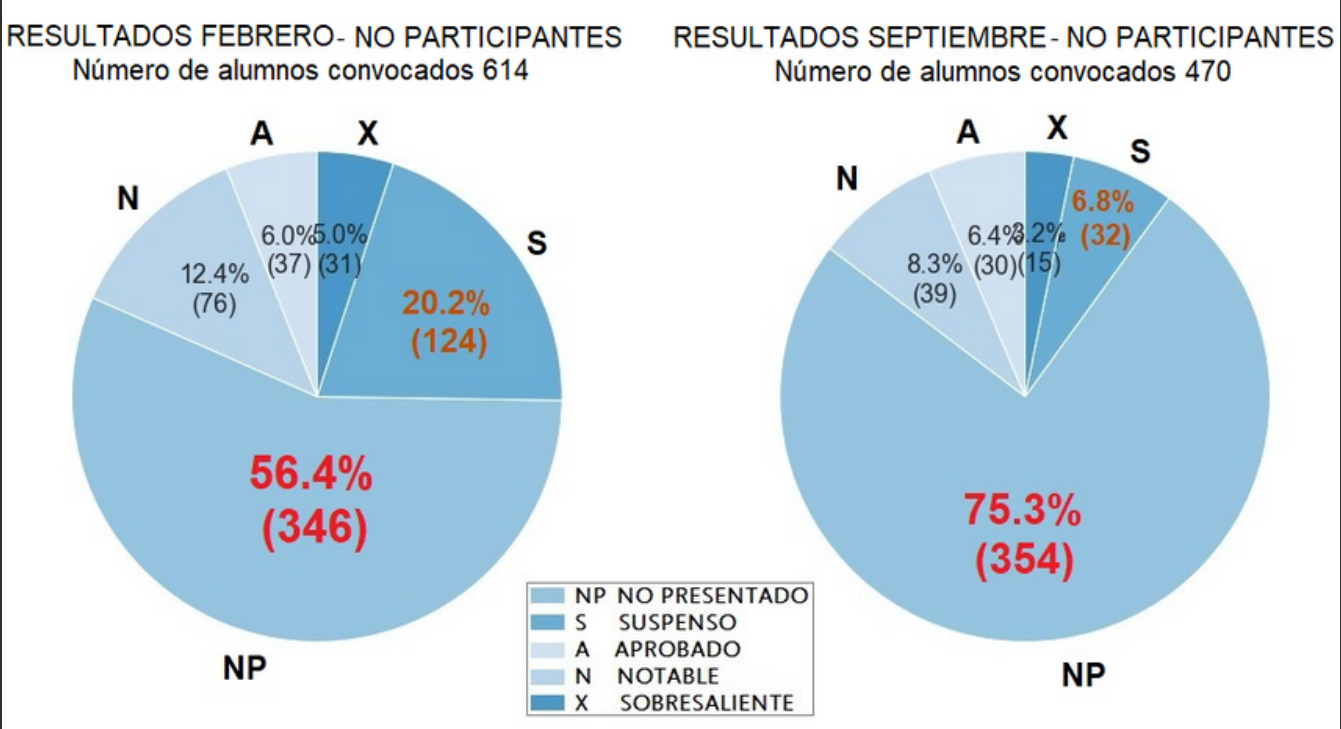
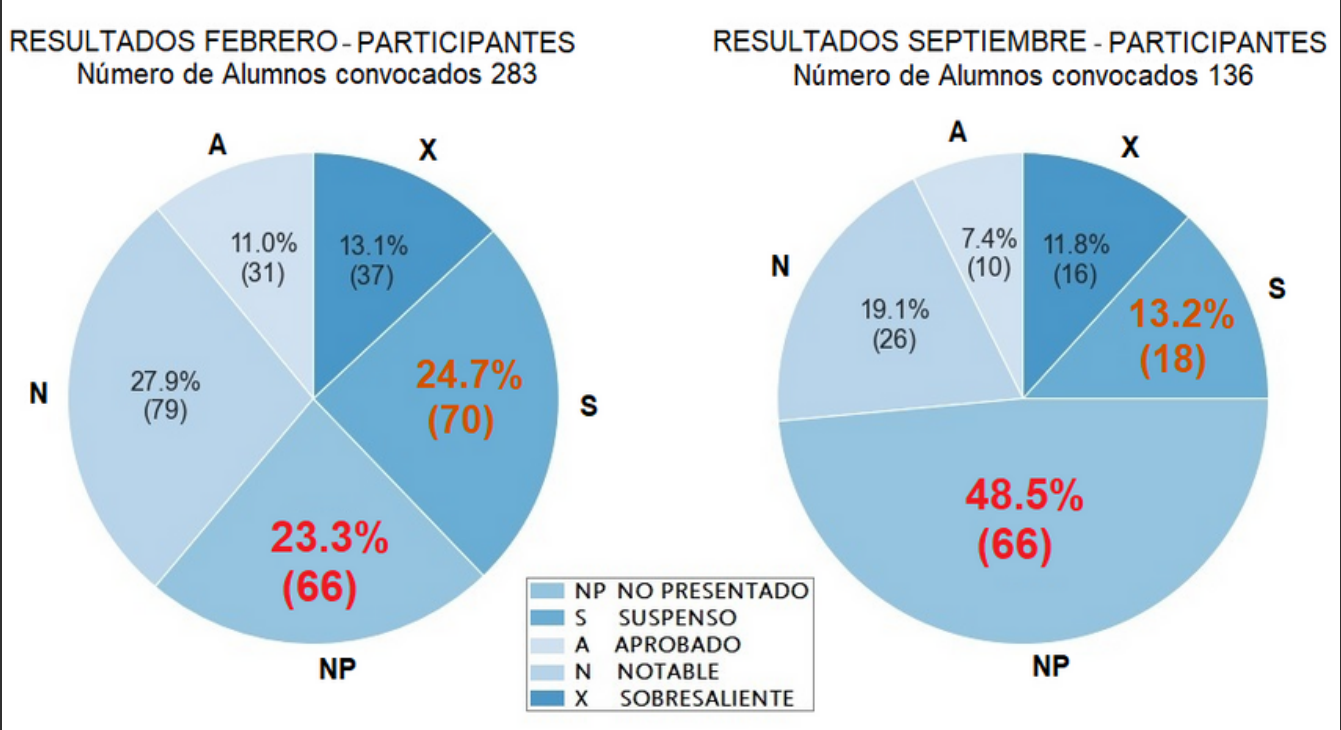
Evaluar si es posible la detección del abandono de un alumno a través del estudio de sus intervenciones en los foros dentro del ámbito de la educación a distancia.

METODOLOGÍA DE TRABAJO

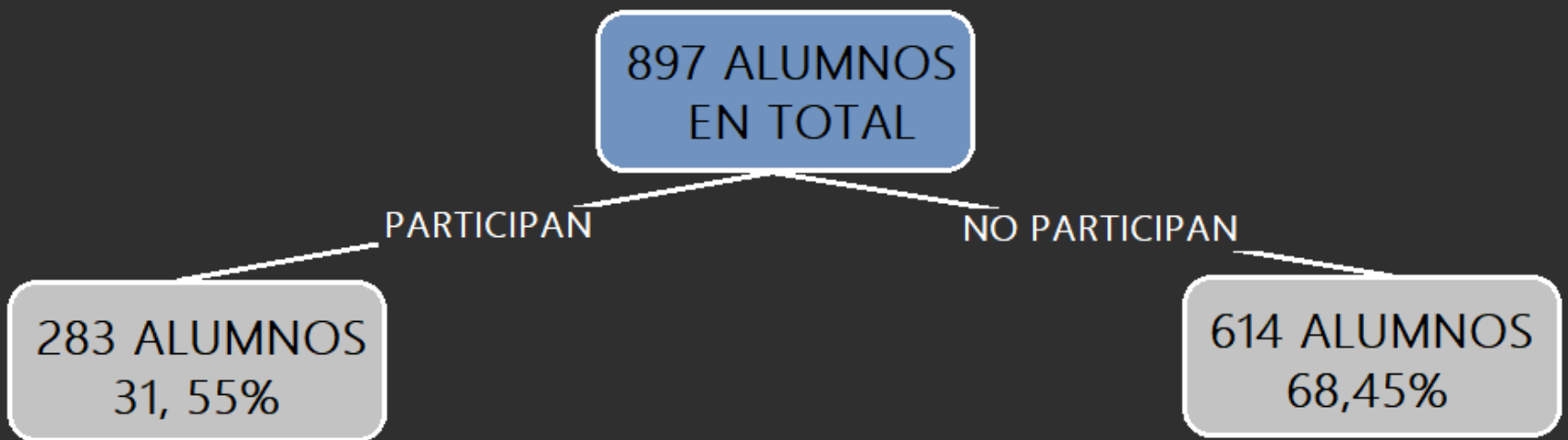
- Analizar y cuantificar
- Extraer información adicional.
- Detectar patrones. Identificación del Abandono o Suspenso

02. LECTURA Y ANÁLISIS DE LOS DATOS

ANÁLISIS DE DATOS



¿Influye la participación en los foros en el abandono de los alumnos?



En Febrero

Participantes

- 23,3% de Abandono
- 24,7% de Suspense

No Participantes:

- 56,4% de Abandono
- 20,2% de Suspensos

En Septiembre

Participantes

- 48,5% de Abandono
- 13,2% de Suspense

No Participantes:

- 75,3% de Abandono
- 6,8% de Suspensos

02. LECTURA Y ANÁLISIS DE LOS DATOS

Autor	n_mensajes	n_foros	n_hilos_ini	n_hilos_ini	n_res	words_avg	FEB	SEP	FIN
37143987	6	2	3	3	3	73.500000	10.0	0.0	10.0
6475899	21	6	2	2	19	152.714286	10.0	0.0	10.0
9032498	2	1	0	0	2	53.500000	10.0	0.0	10.0
0229960	3	1	3	3	0	90.333333	0.7	10.0	10.0
7109741	1	1	0	0	1	31.000000	10.0	0.0	10.0
8482158	1	1	0	0	1	39.000000	10.0	0.0	10.0
7727924	2	1	1	1	1	14.500000	0.6	10.0	10.0
3124949	2	2	2	2	0	82.500000	9.9	0.0	9.9
0117794	1	1	1	1	0	138.000000	0.0	9.9	9.9
8083044	1	1	0	0	1	7.000000	9.9	0.0	9.9
3878710	10	3	2	2	8	49.400000	9.9	0.0	9.9
9767478	16	4	7	7	9	70.687500	9.9	0.0	9.9
7590882	3	2	1	1	2	44.333333	9.9	0.0	9.9
0009986	3	1	1	1	2	44.333333	0.0	9.9	9.9
2852338	1	1	0	0	1	84.000000	9.9	0.0	9.9
8991892	6	3	0	0	6	71.666667	9.9	0.0	9.9
6485661	58	8	12	12	46	128.534483	9.9	0.0	9.9

ANÁLISIS DE DATOS

Extracción de nuevas características

Pasar de un conjunto de datos estrictamente texto a medidas de la actividad a través de variables numéricas.



Dos tipos de características:

- Características de conteo
- Características de ratio

¿Influye la participación en los foros en la nota de los alumnos?

Existen **alumnos** que con una **baja participación** en los foros obtiene una **nota muy elevada**.

02. LECTURA Y ANÁLISIS DE LOS DATOS

ANÁLISIS DE DATOS

	n_mensaje	n_foros	n_hilos	n_hilos_ini	pub_ratio	f_t01_rat	f_t02_rat	f_t03_rat	N_FEB	N_SEP	N_FINAL
n_mensajes	1,0000	0,7794	0,9422	0,8669	0,8299	0,6452	0,5338	0,6762	0,1730	-0,0273	0,1620
n_foros	0,7794	1,0000	0,8483	0,7291	0,8011	0,6099	0,5901	0,5734	0,2134	-0,0461	0,1778
n_hilos	0,9422	0,8483	1,0000	0,8379	0,8416	0,6654	0,6096	0,6340	0,1795	-0,0132	0,1766
n_hilos_ini	0,8669	0,7291	0,8379	1,0000	0,7962	0,6012	0,4907	0,6739	0,1810	0,0443	0,2188
n_res	0,9813	0,7441	0,9164	0,7546	0,7847	0,6169	0,5130	0,6297	0,1578	-0,0531	0,1287
f_t01	0,8096	0,6978	0,8202	0,6782	0,6683	0,8855	0,4962	0,3923	0,2237	-0,0477	0,2000
f_t02	0,7033	0,6574	0,7483	0,6018	0,6047	0,5729	0,8318	0,3517	0,1714	-0,0020	0,1609
f_t03	0,4355	0,3678	0,4302	0,5013	0,5739	0,1932	0,2211	0,7816	0,1793	-0,0311	0,1589
f_t04	0,6350	0,4013	0,5270	0,5717	0,4696	0,2482	0,1572	0,3302	0,0449	0,0310	0,0843
sem_1	0,3089	0,2261	0,2408	0,3052	0,2275	0,0880	0,0436	0,2118	0,0706	-0,0695	0,0151
sem_2	0,6336	0,3965	0,5239	0,5507	0,4427	0,2648	0,1574	0,3365	0,0162	0,0550	0,0851
sem_3	0,6071	0,5949	0,6426	0,5436	0,5787	0,3145	0,5511	0,4854	0,0781	-0,0261	0,0513
sem_4	0,6575	0,5324	0,6147	0,5718	0,5363	0,4199	0,3658	0,4564	0,1180	-0,0723	0,0808
sem_5	0,4444	0,3739	0,5028	0,5138	0,5573	0,3250	0,2127	0,4138	0,1450	-0,0433	0,1187
sem_6	0,6624	0,5595	0,7150	0,5064	0,5258	0,5499	0,3984	0,3877	0,1523	-0,0809	0,1014
sem_7	0,3319	0,4396	0,3975	0,2788	0,4467	0,2466	0,3851	0,2994	0,1022	-0,0379	0,0589
sem_8	0,4999	0,3528	0,3811	0,4528	0,4055	0,2646	0,2726	0,3119	0,0910	-0,0103	0,0872
sem_9	0,6377	0,4573	0,5192	0,5450	0,5182	0,4069	0,2833	0,4356	0,0803	0,0288	0,1047
sem_10	0,6915	0,5195	0,6080	0,5848	0,5046	0,5543	0,4476	0,4021	0,1163	0,0270	0,1434
sem_11	0,6361	0,4758	0,6506	0,5053	0,4690	0,5811	0,3491	0,3658	0,1267	0,0431	0,1544
sem_12	0,2656	0,2302	0,3072	0,2380	0,2910	0,3283	0,2387	0,2516	0,1729	-0,0348	0,1498
sem_13	0,2184	0,2831	0,2790	0,1920	0,2812	0,2378	0,2276	0,2441	0,1219	-0,0011	0,1202
sem_14	0,2225	0,2345	0,2552	0,2193	0,1800	0,1930	0,0575	0,0255	0,0622	0,0141	0,0878
sem_15	0,7232	0,4412	0,6225	0,5556	0,4573	0,3473	0,2526	0,4958	0,0729	-0,0785	0,0230
sem_16	0,7962	0,4737	0,6321	0,6044	0,4975	0,3968	0,2682	0,4969	0,0420	-0,0172	0,0254
words_avg	0,1858	0,2233	0,2466	0,3474	0,2506	0,2352	0,2097	0,2294	0,1472	0,0254	0,1590
avg_w_len	0,0129	0,1208	0,0422	0,0429	0,0658	0,0138	-0,0018	0,1221	0,0640	-0,0311	0,0732
noun_avg	0,1729	0,2248	0,2355	0,3292	0,2327	0,2358	0,2394	0,1909	0,1676	0,0129	0,1774
adj_avg	0,1846	0,2354	0,2539	0,3619	0,2719	0,2494	0,2035	0,2246	0,1549	-0,0153	0,1415
verb_avg	0,1567	0,1741	0,2082	0,2878	0,2033	0,1731	0,1817	0,2159	0,1473	0,0412	0,1754
pub_ratio	0,8299	0,8011	0,8416	0,7962	1,0000	0,6144	0,5590	0,6790	0,2268	-0,0295	0,2116
f_t01_rat	0,6452	0,6099	0,6654	0,6012	0,6144	1,0000	0,4996	0,3076	0,2266	-0,0209	0,2115
f_t02_rat	0,5338	0,5901	0,6096	0,4907	0,5590	0,4996	1,0000	0,2905	0,1747	0,0206	0,1819
f_t03_rat	0,6762	0,5734	0,6340	0,6739	0,6790	0,3076	0,2905	1,0000	0,1508	-0,0178	0,1302
f_t04_rat	0,4924	0,3326	0,4273	0,4886	0,4277	0,1745	0,1068	0,2296	0,0386	0,0563	0,1003
words_ratio	0,1677	0,2546	0,2124	0,2746	0,2576	0,2583	0,2255	0,1758	0,2183	-0,0022	0,2274
len_w_ratio	-0,0089	0,0624	0,0199	0,0063	-0,0020	-0,0217	-0,0130	0,0899	0,0435	-0,0178	0,0680
nouns_ratio	0,1554	0,2486	0,2047	0,2629	0,2397	0,2550	0,2432	0,1438	0,2198	-0,0151	0,2222
adjs_ratio	0,1785	0,2758	0,2356	0,2948	0,2741	0,2707	0,2128	0,1602	0,1964	-0,0532	0,1669
verbs_ratio	0,1387	0,1896	0,1768	0,2232	0,2122	0,1836	0,1844	0,1692	0,1878	0,0294	0,2218

¿Existe correlación estrecha entre alguna de las características y la nota?

NO se aprecia correlación directa

¿Existe correlación entre las características de cada registro?

SI se aprecia correlación entre algunos conjuntos de variables.

03. MODELOS BASADOS SÓLO EN TEXTO

MEDIDAS DE CALIDAD DE LAS PREDICCIONES

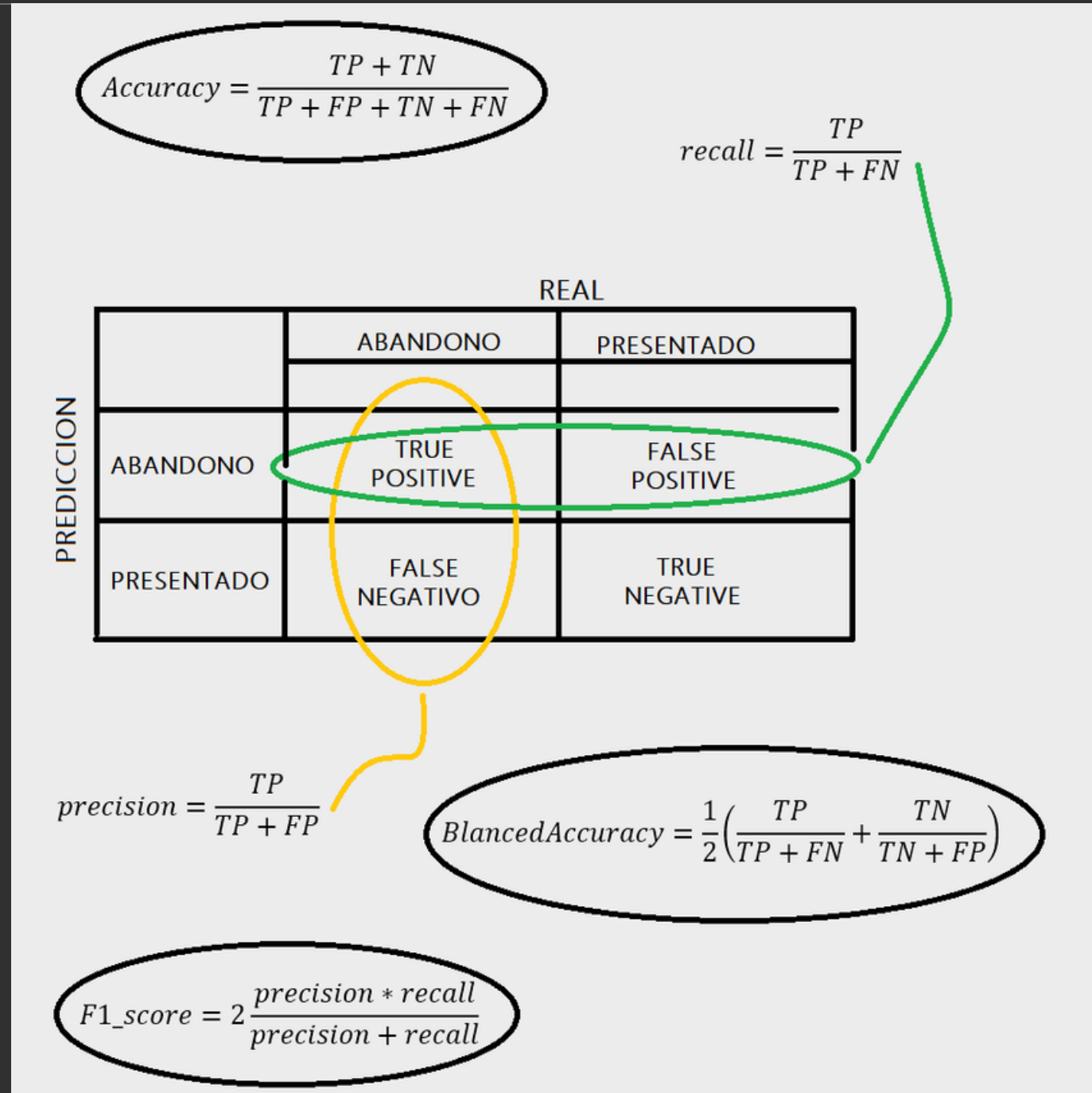
MEDIDAS DE CALIDAD

Uso de la Matriz de confusión para calcular medidas de calidad de las predicciones.

- Precisión
- Sensibilidad (recall)

Conjunto de entrenamiento fuertemente desbalanceado:

- Balanced Accuracy
- F1_score



03. MODELOS BASADOS SÓLO EN TEXTO

MODELOS DE CLASIFICACIÓN DE APRENDIZAJE AUTOMÁTICO

REGRESIÓN LOGÍSTICA MODELO BASE

Método de partida, basado en la resgresión lineal y uno de los métodos simples más usados en clasificación

ÁRBOLES DE DECISIÓN MODELO NO LINEAL

Metódo no linear de clasificación que forma un árbol binario en base a la elección de una característica.

ADABOOST

Boosting adaptativo

LIGHTGMB

Intensificación del Gradiente

BOSQUES ALEATORIOS

Boosting Aggregating

XGBOOST

Intensificación del Gradiente

MODELOS "ENSEMBLED"

WEAK MODEL

03. MODELOS BASADOS SÓLO EN TEXTO

3 TIPOS DE EXPERIMENTOS



E1. CLASIFICACIÓN DE ALUMNOS SEGÚN LA NOTA FINAL DEL CURSO

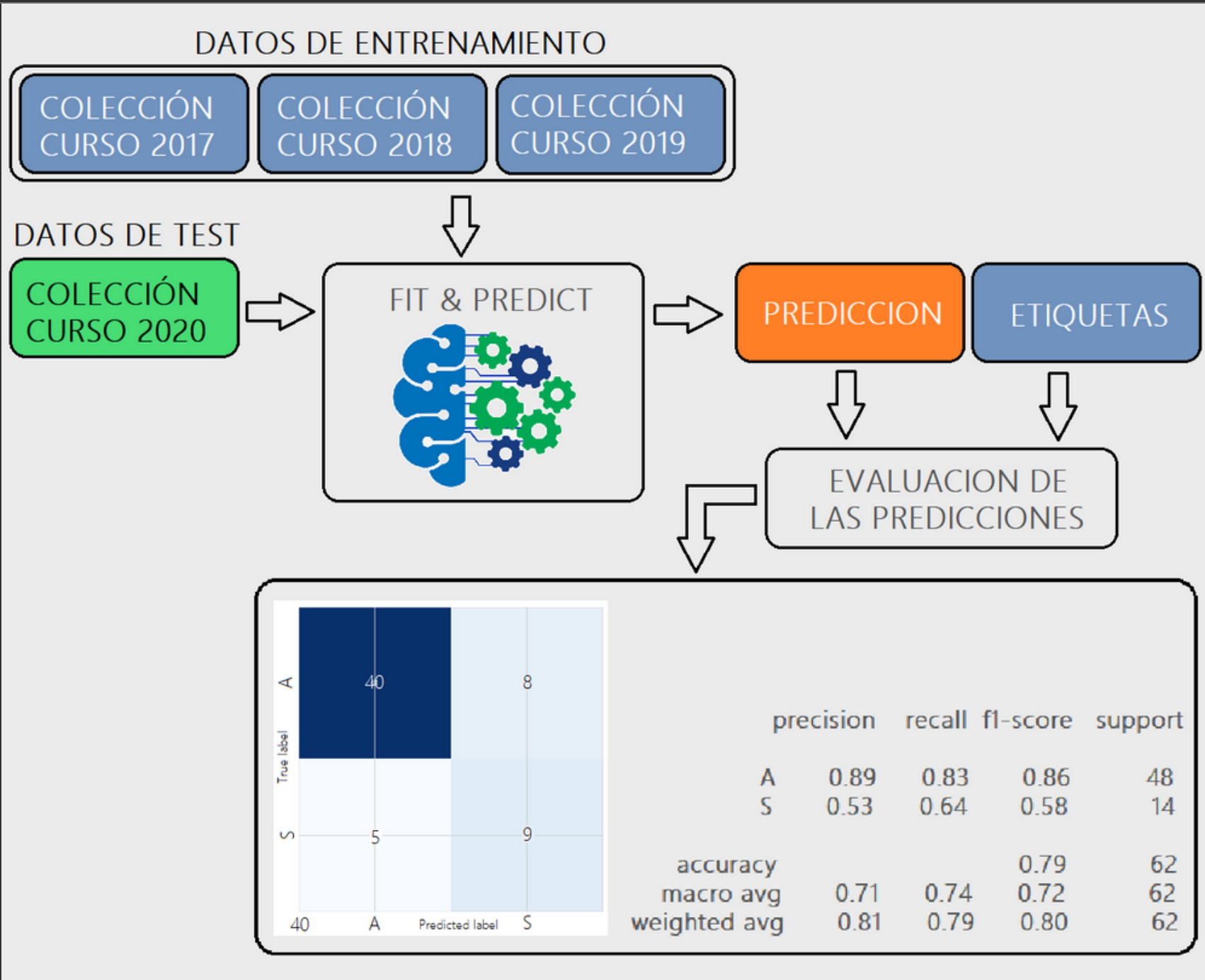
Clasificación de grano fino: multiclase

E2. CLASIFICACIÓN DE ALUMNOS SEGÚN SI HAN APROBADO O NO LA ASIGNATURA

Clasificación binaria

E3. CLASIFICACIÓN DE ALUMNOS SEGÚN SI HAN ABANDONADO O NO LA ASIGNATURA

Clasificación binaria



01. GRID SEARCH

Ajuste de los hiperparámetros

02. VALIDACIÓN CRUZADA

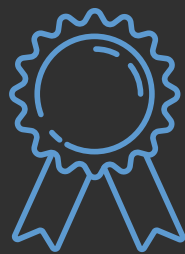
Evaluación del clasificador resultado del ajuste anterior.

03. TEST DE GENERALIZACIÓN

Comprobamos como generaliza cada modelo ajustado.

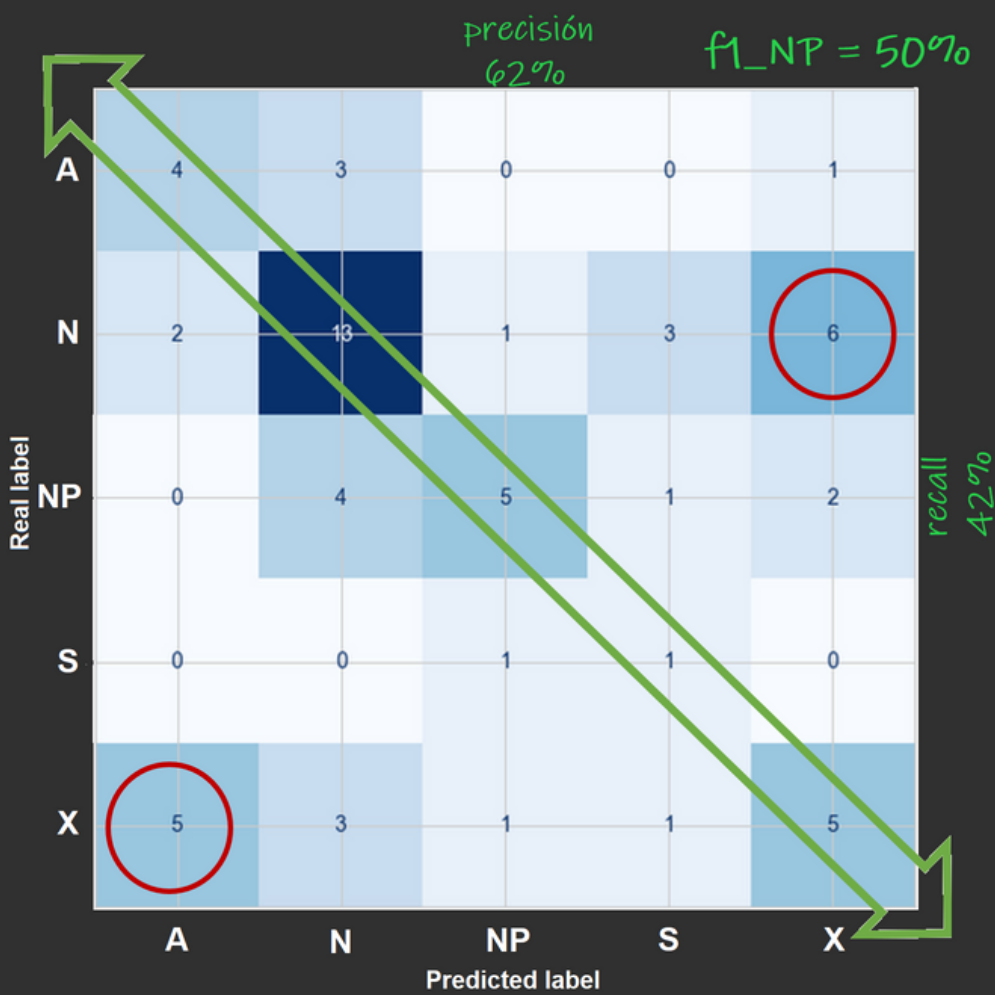
0.4 TECNICA DE BALANCEO

Balanced Bagging



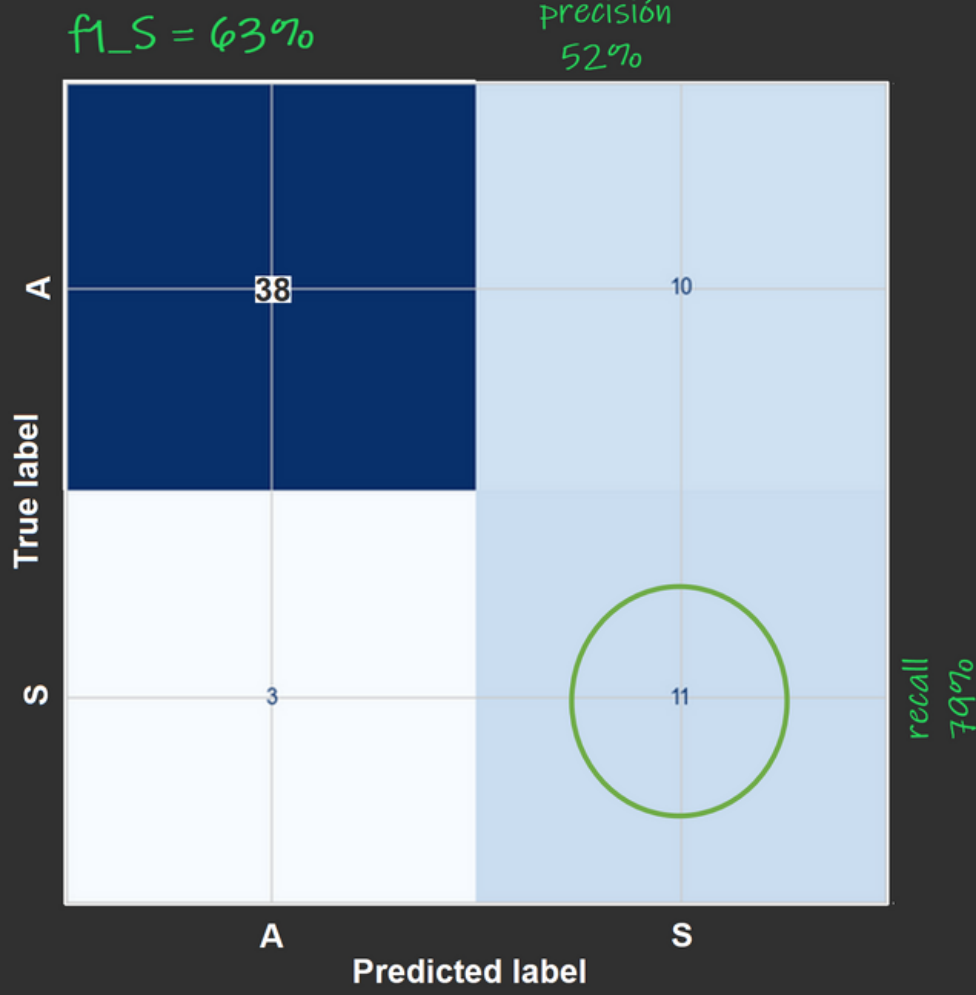
E1. ADABOOST + Balanced Bagging

Clasificación multiclase



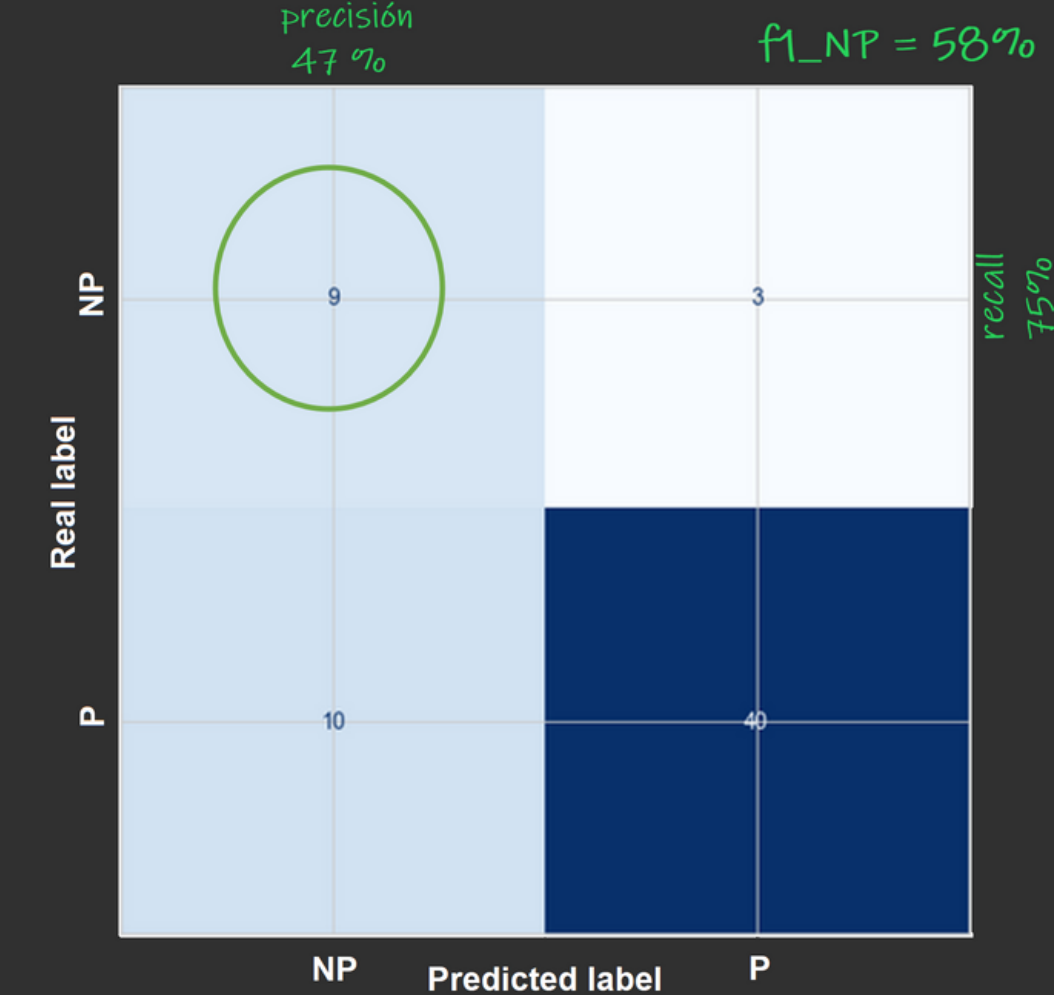
E2. ARBOLES DE DECISIÓN + Balanced Bagging

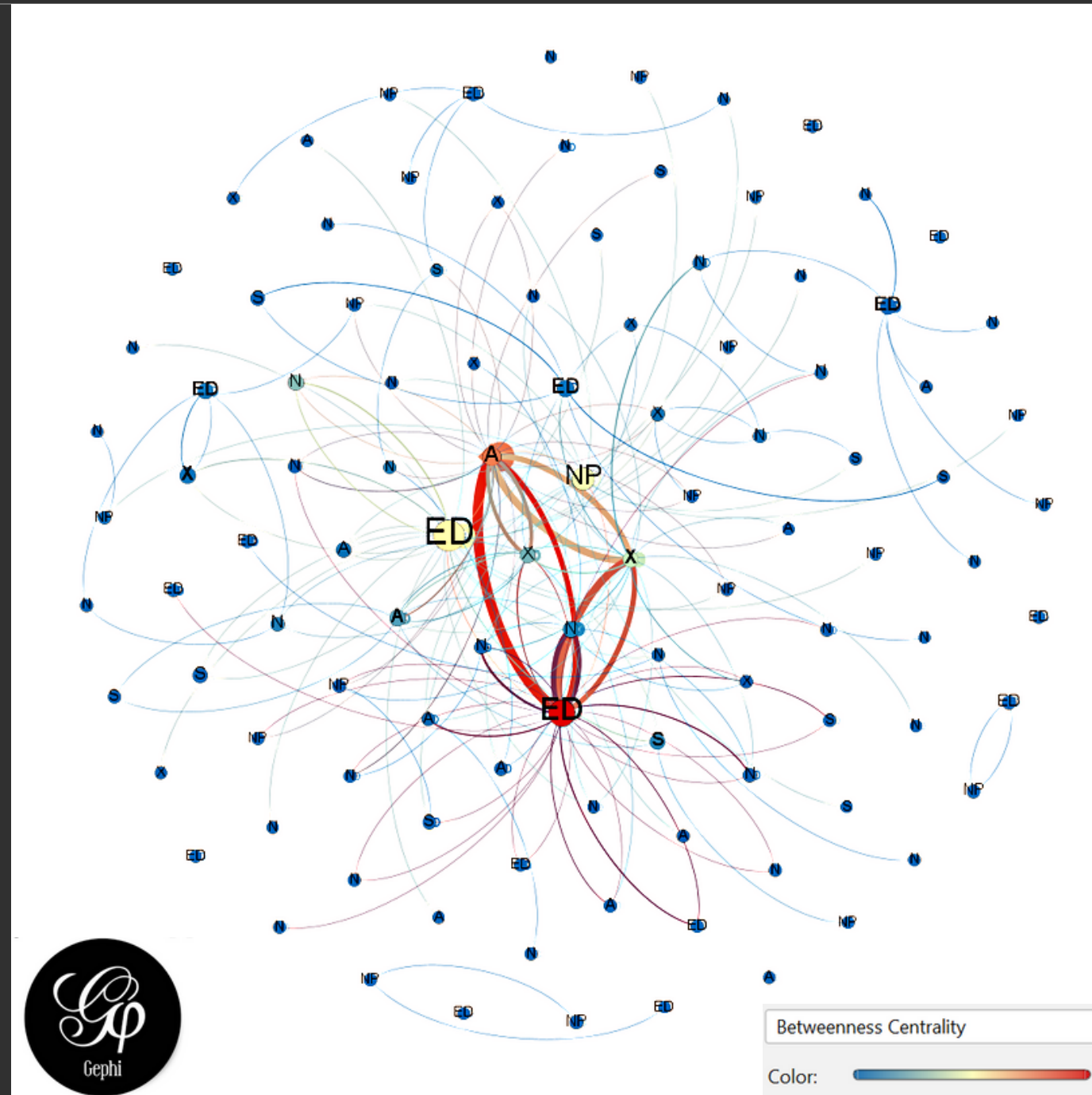
Clasificación binaria. Alumnos con suspenso



E3. LightGMB + Balanced Bagging

Clasificación binaria. Abandono





TRANSFORMACIÓN DE LOS DATOS

Convertir los mensajes en un grafo dirigido. No se considera el contenido de los mensajes sino las interacciones entre autores.

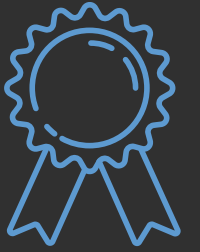
- **Aristas:** mensajes
- **Nodos:** autores

OBTENCIÓN DE CARACTERÍSTICAS TÍPICAS DE LA TEORÍA DE GRAFOS.

Usamos métricas típicas del estudio de grafos para obtener nuevas características que describan la actividad de los alumnos.

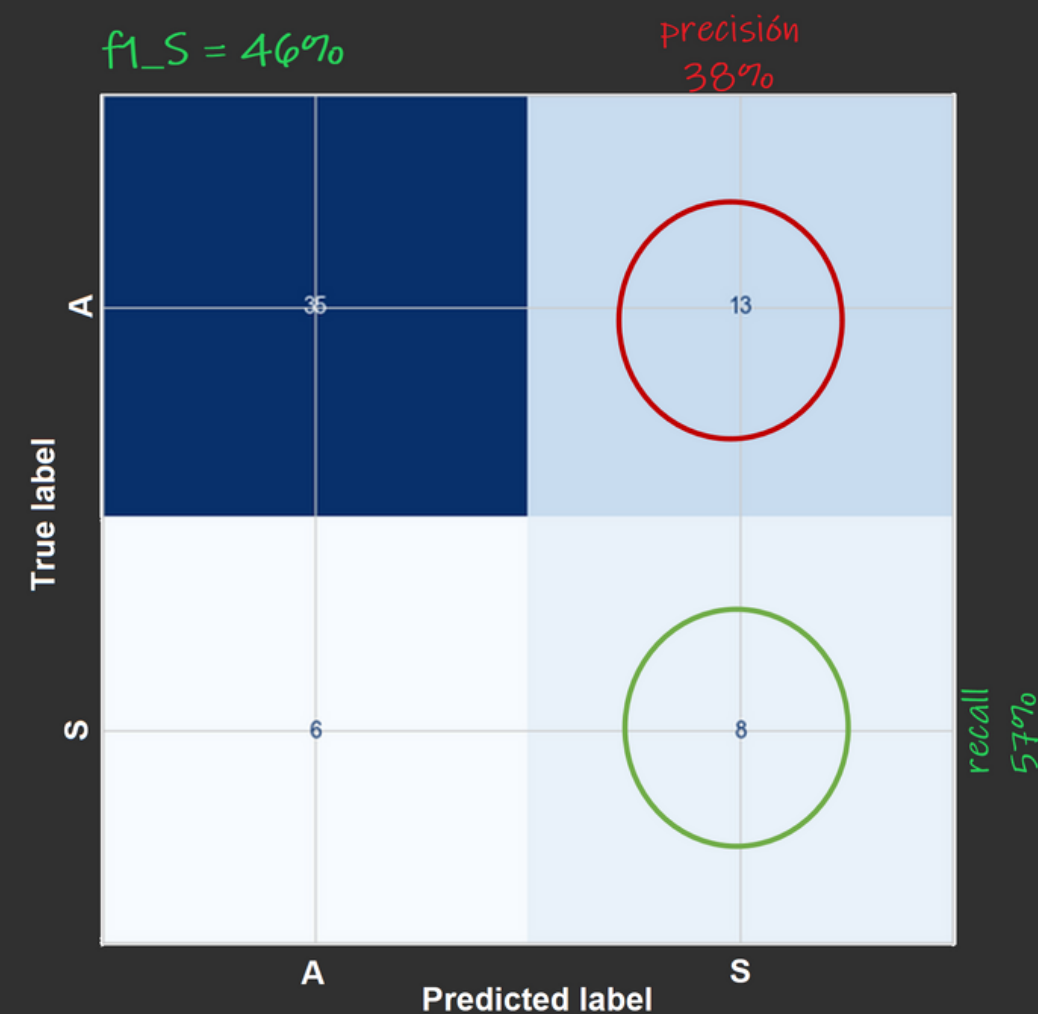
Tipos de Métricas:

- Centralidad
- Grado
- Prestigio



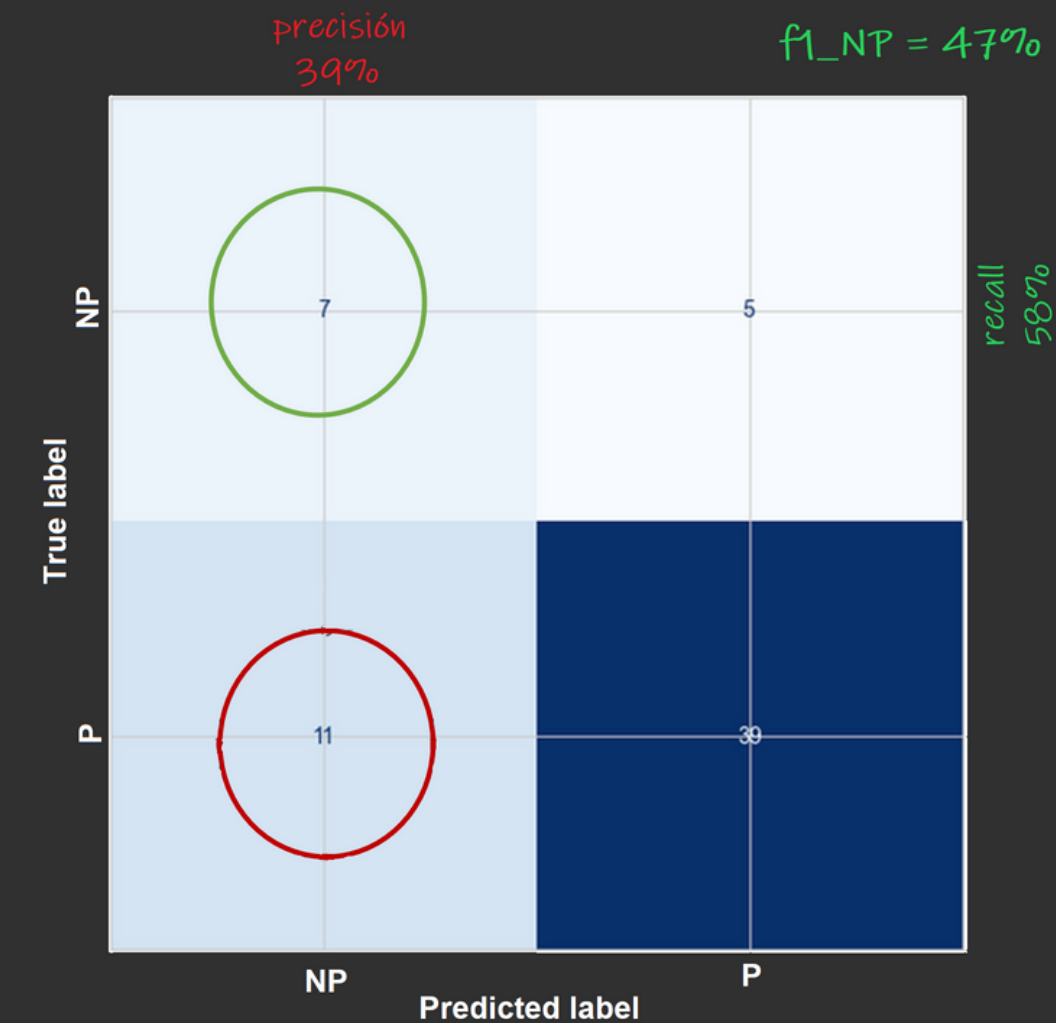
E4_1. Bósques Aleatorios + Balanced Bagging

Igualado con AdaBoost



E4_2. Árboles de decisión

Adaboost mejora el valor de f1 con una mayor precisión valor más bajo de recall



05. CONCLUSIONES Y PASOS FUTUROS



CONCLUSIONES

CON LOS MÉTODOS UTILIZADOS SOBRE EL CONJUNTO DE DATOS SUMINISTRADO, NO SE PUEDE IDENTIFICAR CON PRECISIÓN A LOS ALUMNOS QUE ABANDONAN O SUSPENDEN LA ASIGNATURA

Conjunto de datos

- Menos del 32% de los alumnos matriculados participan en los foros.
- La base de conocimiento por asignatura para el entrenamiento es pequeña.
- Los datos están muy desbalanceados para los casos de detección del abandono y suspenso de la asignatura

Entrenamiento de modelos

- El modelo de Regresión logística, en general es capaz de detectar el mayor numero de casos de abandono y supensos a consta de tener una mala precisión
- Los modelos de Bósques Aleatorios y Adaboost obtienen mejores resultados ponderados una vez aplicado tecnicas de balanceo.
- Usando características obtenidas del estudio de los datos como RR.SS. da resultados similares.

05. CONCLUSIONES Y PASOS FUTUROS



PROPUESTA DE MEJORA Y PASOS FUTUROS

Extracción de nueva características

- Investigar en la extracción de nuevas características de los datos suministrados para el estudio.
- Nuevas características desde la plataforma de cursos virtuales de la UNED.

Predicción en la primera convocatoria de Febrero

Estudiar el comportamiento de las predicciones para la primera convocatoria.

Investigación de modelos y métricas alternativas

- Analizará los resultados bajo nuevas métricas de evaluación
- Investigar el comportamiento de clasificadores añadidos.

Gracias

VISUALIZACIÓN DE RESULTADOS
11.1%

AUTOMATIZACIÓN DEL ENTRENAMIENTO DE MODELOS
22.2%

Agradecimientos especiales a
Álvaro Rodrigo Yuste, tutor de este proyecto
Mi familia y amigos por su apoyo ayuda continua.

