Proyecto8: Predicción Retención Empleados

ALUMNO: RAFAEL CASTELLOT DE MIGUEL

INTRODUCCIÓN

MODELOS 1 Y 2

- 1-eda-y-gestion-de-nulos
- 2-encoding
- 3-estandarizacion
- 4-gesition-de-outliers
- 5-Modelos

MODELOS 2 Y 3

- 1-eda-y-gestion-de-nulos
- 2-encoding
- 3-estandarizacion
- 4-gesition-de-outliers
- 5-desbalanceo
- 6-Modelos

Distinción de modelos

- **MODELO 1**
  - No se eliminan duplicados ni se trata el desbalanceo
- **MODELO 2**
  - Se eliminan duplicados, no se trata el desbalanceo.
- **MODELO 3**
  - Se eliminan duplicados y se trata el desbalanceo

Métricas

- **MODELO 1**

	accuracy	precision	recall	f1	kappa	auc	modelo
0	0.862622	0.842180	0.862622	0.830323	0.266828	0.811597	logistic_regression
1	0.843249	0.816180	0.843249	0.801269	0.198315	0.812602	logistic_regression
2	0.958214	0.957370	0.958214	0.956898	0.830346	0.977996	tree
3	0.933638	0.931342	0.933638	0.931228	0.746309	0.934939	tree
4	0.999714	0.999714	0.999714	0.999714	0.998904	1.000000	random_forest
5	0.997712	0.997718	0.997712	0.997705	0.991777	1.000000	random_forest
6	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	gradient_boosting
7	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	gradient_boosting
8	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	xgboost
9	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	xgboost

Métricas

- **MODELO 2**

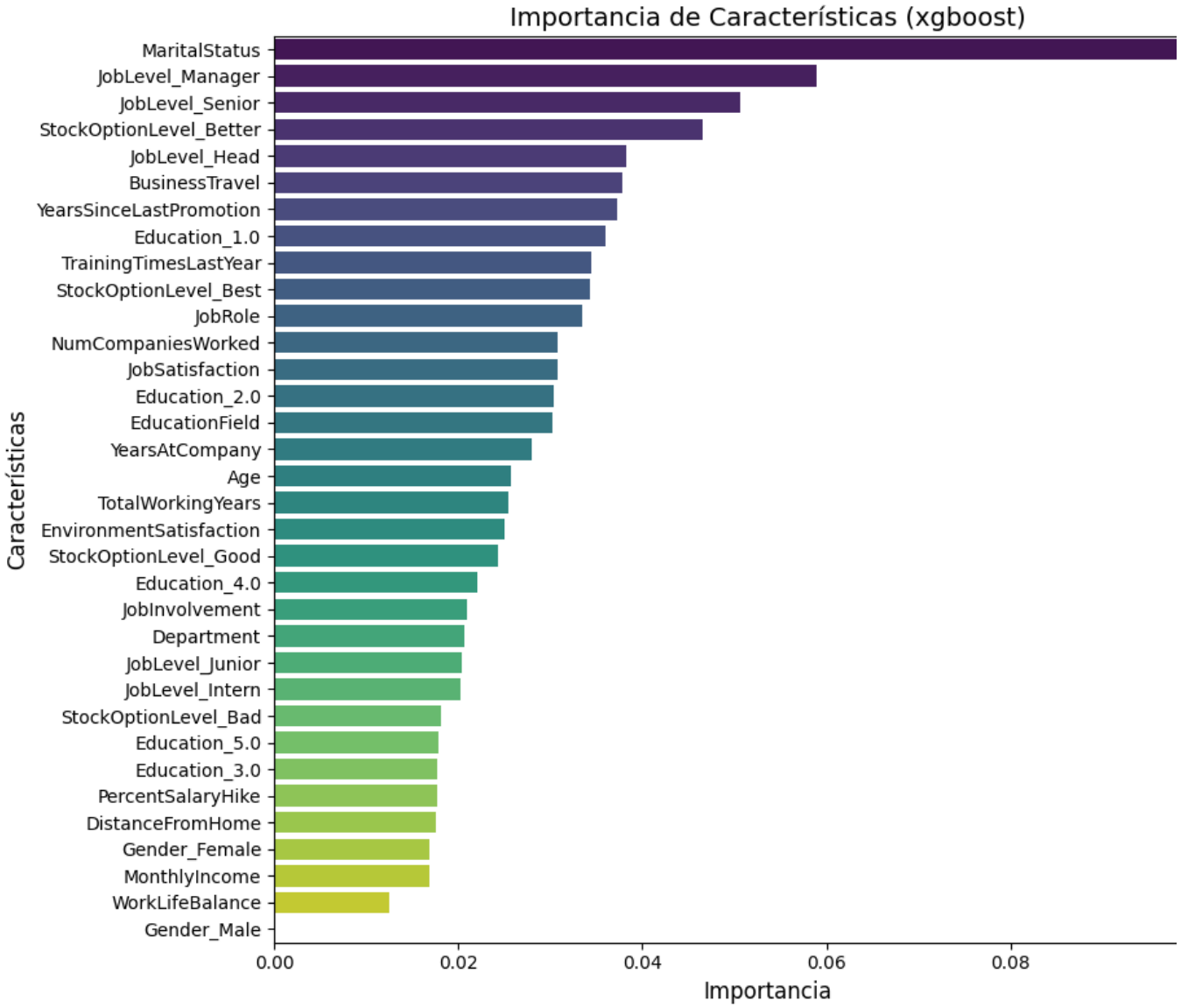
	accuracy	precision	recall	f1	kappa	auc	modelo
0	0.864366	0.842966	0.864366	0.831259	0.253543	0.812105	logistic_regression
1	0.849359	0.833113	0.849359	0.819942	0.323367	0.831124	logistic_regression
2	0.861958	0.839706	0.861958	0.824414	0.218604	0.693877	tree
3	0.836538	0.810656	0.836538	0.801726	0.251974	0.683071	tree
4	0.999197	0.999198	0.999197	0.999197	0.996861	1.000000	random_forest
5	0.858974	0.861140	0.858974	0.824910	0.335913	0.824079	random_forest
6	0.892456	0.898666	0.892456	0.867770	0.417123	0.943053	gradient_boosting
7	0.852564	0.842898	0.852564	0.819802	0.318907	0.834473	gradient_boosting
8	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	xgboost
9	0.858974	0.844709	0.858974	0.839126	0.404993	0.810059	xgboost

Métricas

- **MODELO 3**

	accuracy	precision	recall	f1	kappa	auc	modelo
0	0.836725	0.837051	0.836725	0.836684	0.673445	0.916441	logistic_regression
1	0.839319	0.839447	0.839319	0.839308	0.678649	0.903173	logistic_regression
2	0.905821	0.905851	0.905821	0.905820	0.811643	0.972101	tree
3	0.807183	0.808027	0.807183	0.807039	0.614327	0.854395	tree
4	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	random_forest
5	0.911153	0.914874	0.911153	0.910962	0.822337	0.958412	random_forest
6	0.999527	0.999527	0.999527	0.999527	0.999053	1.000000	gradient_boosting
7	0.909263	0.913271	0.909263	0.909052	0.818559	0.950057	gradient_boosting
8	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	xgboost
9	0.916824	0.918003	0.916824	0.916770	0.833664	0.964909	xgboost

Importancia variables con xgboost



Propuestas de mejora y next steps

- **- Corregir el overfitting.**
- **- Analizar si al corregir el desbalanceo, se han introducido sesgos en la variable respuesta (por ejempl,o si los "No" generados artificialmente tienen características muy parecidas entre ellos, o diferentes a los que ya existían)**
- **- Hacer un modelo predictivo basado en los duplicados que se eliminaron y comparar conclusiones con este modelo.**
- **- Introducir mejoras en el preprocesamiento (probar un tratamiento diferente de nulos y outliers y probar otros métodos de estandarización)**