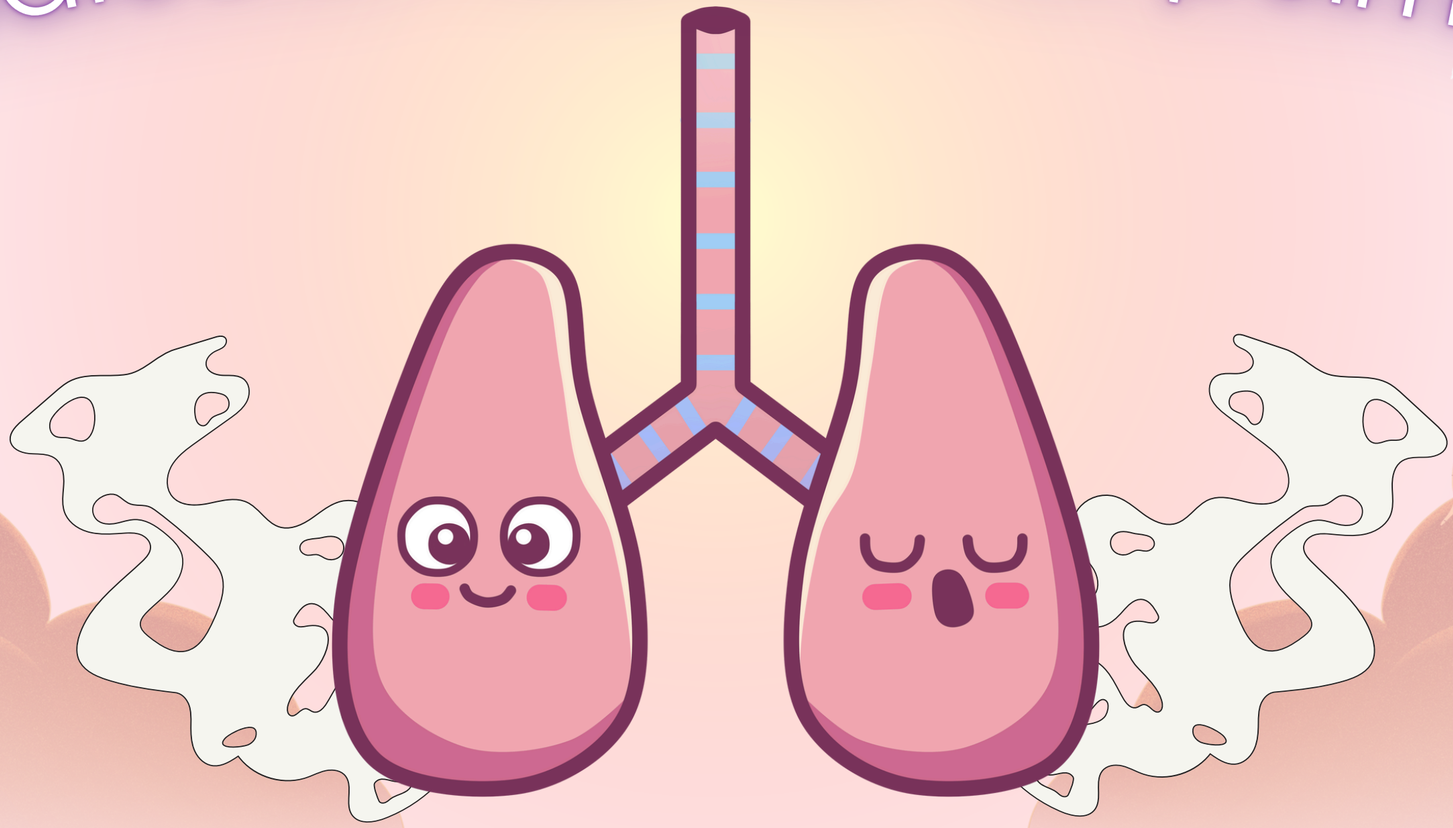
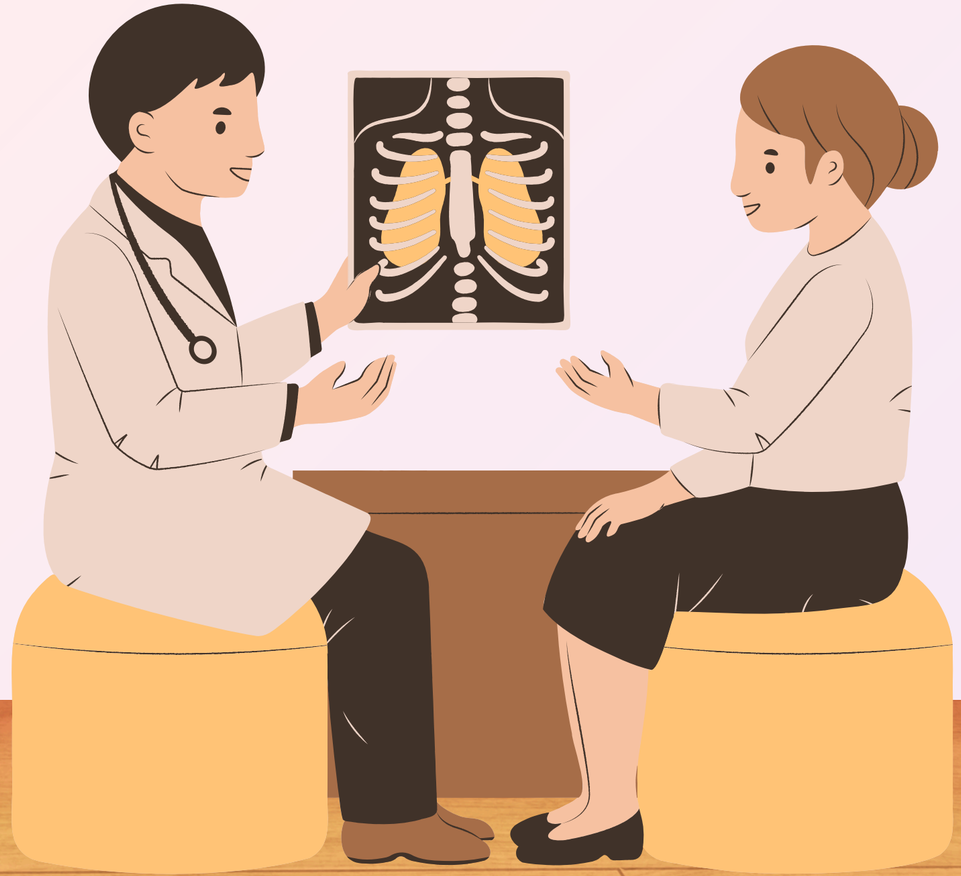
PROYECTO MACHINE LEARNING:
Predicción cáncer de pulmón



Laura García Brena

CONTEXTO

- Padre fallecido de cáncer de pulmón
- Estudios Biología
- Impulso futuro en Bioinformática



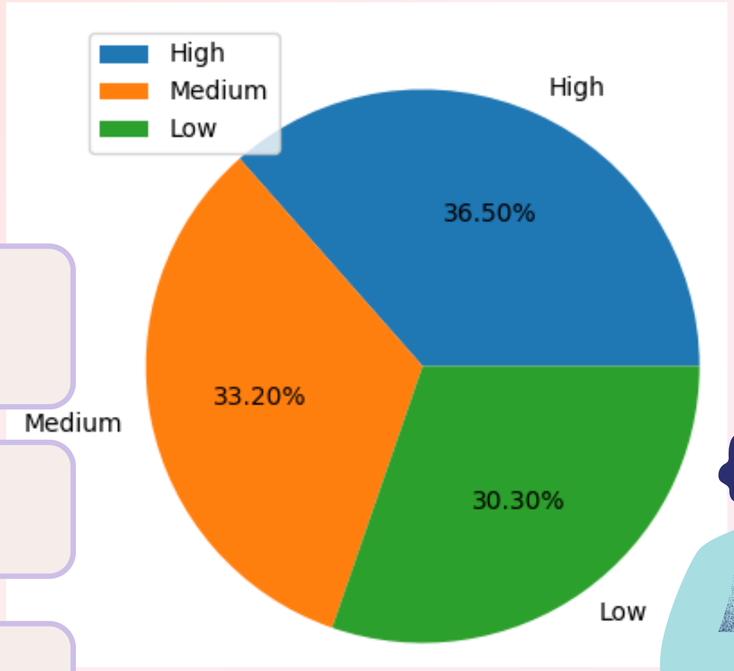
DATASET

- Kaggle

- Edad: Edad del paciente.
- Género: Sexo del paciente.
- Contaminación atmosférica: Nivel de exposición del paciente a la contaminación atmosférica.
- Consumo de alcohol: Nivel de consumo de alcohol del paciente.
- Alergia al polvo: Nivel de alergia al polvo del paciente.
- Riesgos laborales: Nivel de riesgos laborales del paciente.
- Riesgo genético: Nivel de riesgo genético del paciente.
- Enfermedad pulmonar crónica: Nivel de enfermedad pulmonar crónica del paciente.
- Dieta equilibrada: Nivel de dieta equilibrada del paciente.
- Obesidad: Nivel de obesidad del paciente.
- Tabaquismo: Nivel de tabaquismo del paciente.
- Fumador pasivo: Nivel de tabaquismo pasivo del paciente.
- Dolor torácico: Nivel de dolor torácico del paciente.
- Tos con sangre: Nivel de tos con sangre del paciente.
- Fatiga: Nivel de fatiga del paciente.
- Pérdida de peso: Nivel de pérdida de peso del paciente.
- Dificultad para respirar: Nivel de dificultad para respirar del paciente.
- Sibilancias: Nivel de sibilancias del paciente.
- Dificultad para tragar: Nivel de dificultad para tragar del paciente.
- Uñas en palillo de tambor: Nivel de uñas en palillo de tambor del paciente.

PROBLEMA DE NEGOCIO

- Pedecir el grado de daño del cáncer de pulmón
- Problema supervisado
- Clasificador multiclase



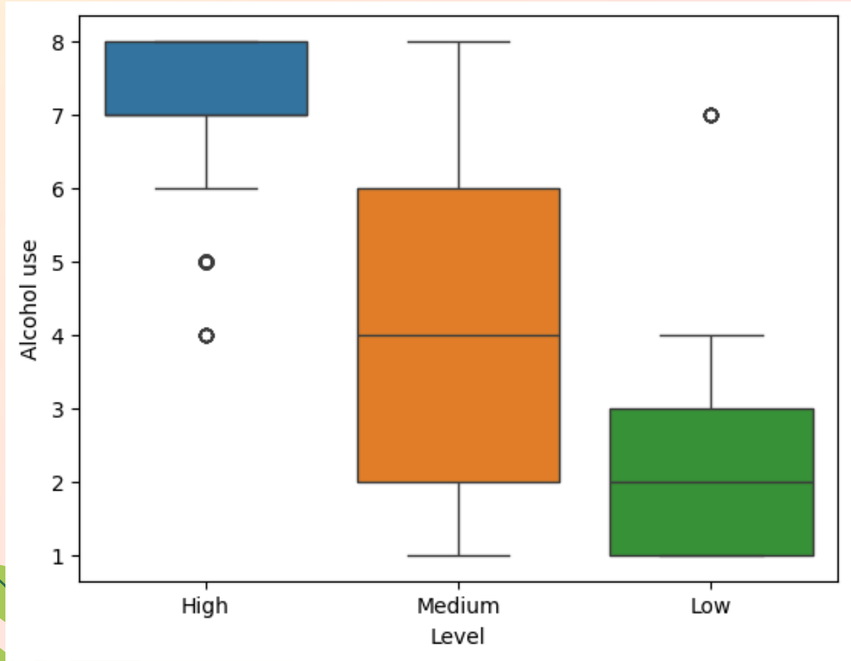
PASOS SEGUIDOS

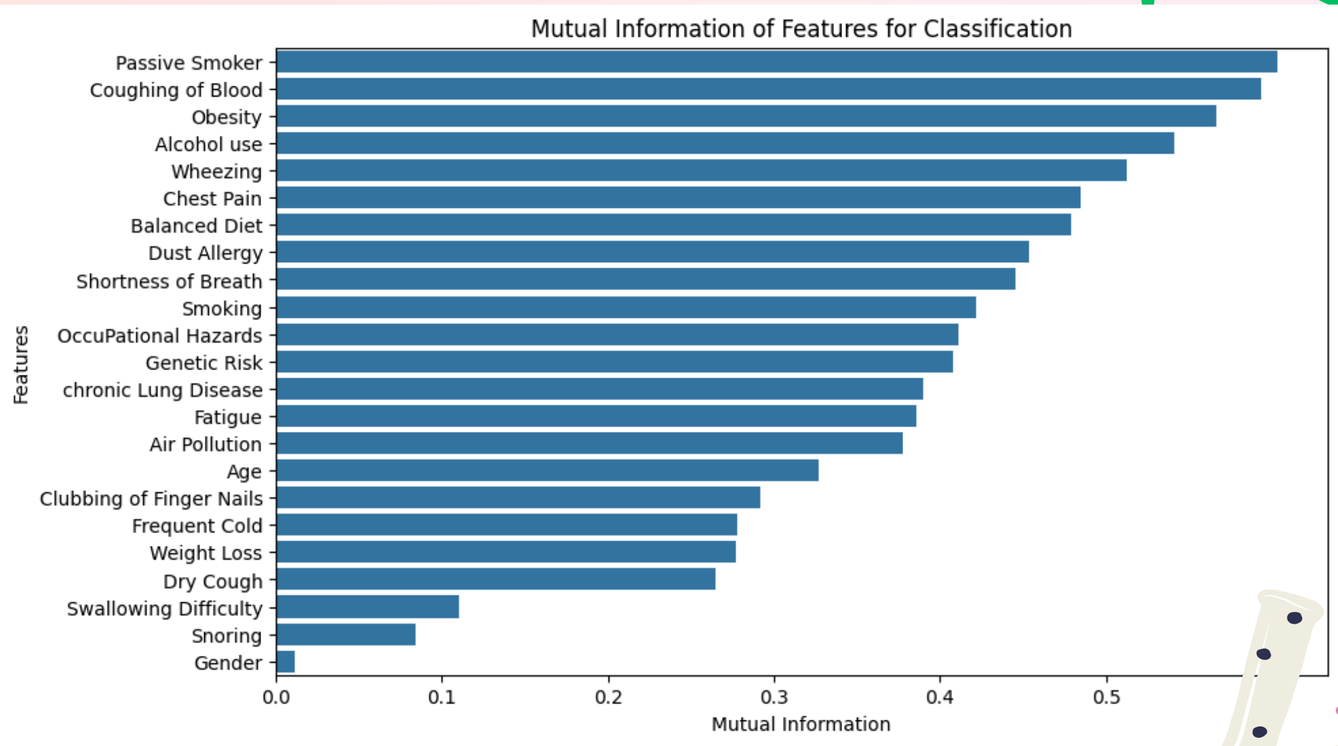
- Importar Librerías
- Cargar Dataset
- Exploración y primera limpieza
- Train/Test Split
- MiniEDA
- Selección Features
- Baseline de Modelos y Validación Cruzada
- Optimización Hiperparámetros
- Evaluación contra test
- Primeras conclusiones
- EXTRA

SELECCIÓN FEATURES

- Selección visual
- Selección Estadística
- Mutual Information
- Selección por modelo
- Seleción manual
- RFE
- SFS
- Hard-Voting

SELECCIÓN FEATURES



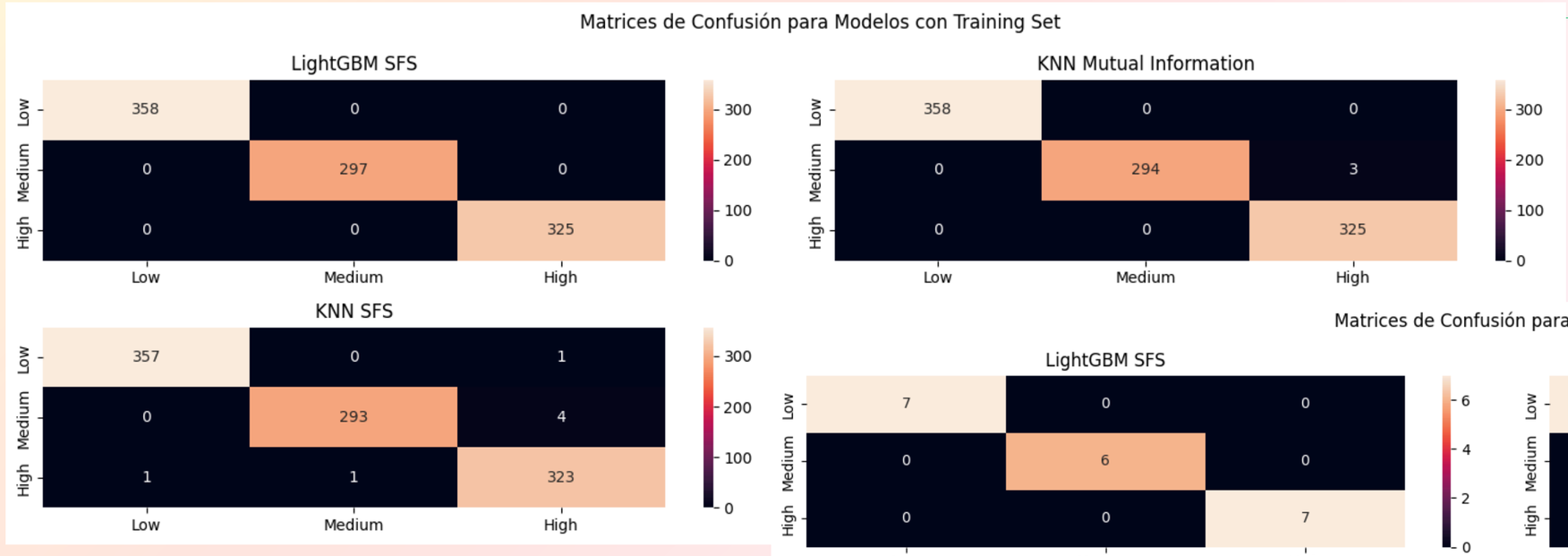


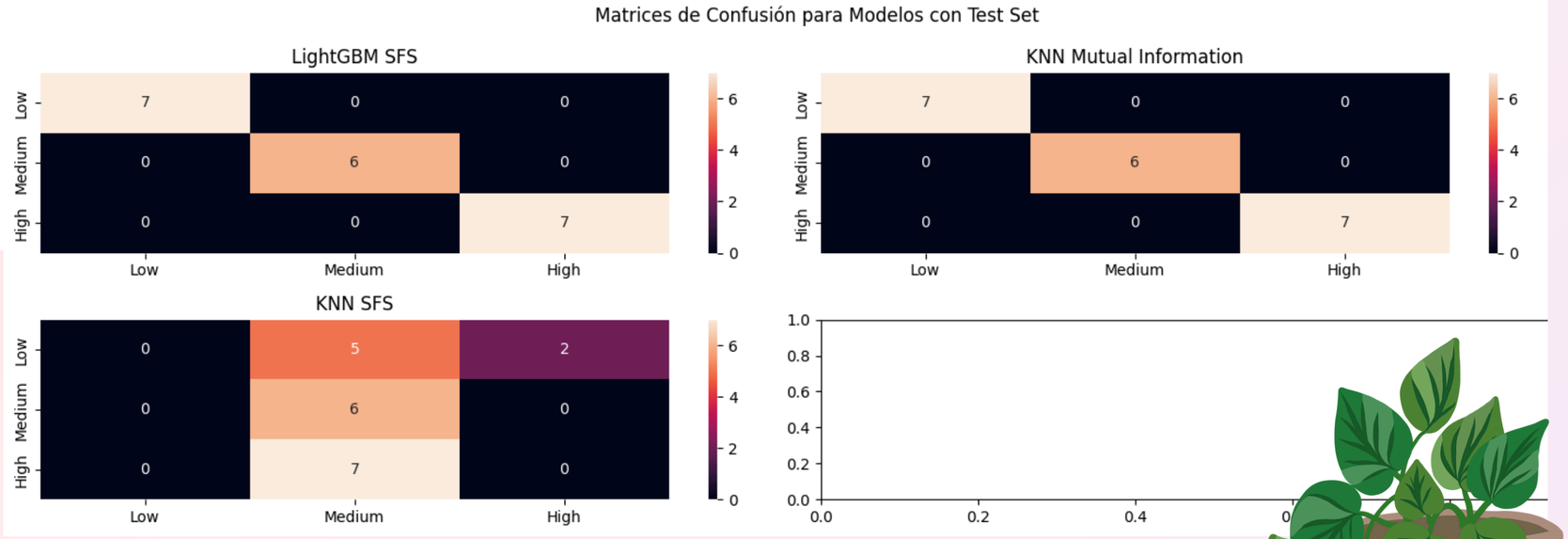
BASELINE MODELOS Y VALIDACIÓN CRUZADA

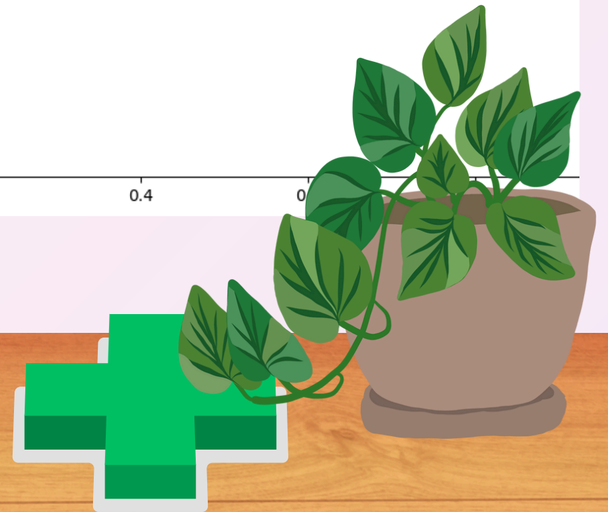
	features_list	model	avg. recall
0	Seleccion estadistico	Random Forest	1.000000
2	Seleccion estadistico	Decission Tree	1.000000
8	Seleccion mutual information	Decission Tree	1.000000
4	Seleccion estadistico	LightGBM	1.000000
5	Seleccion estadistico	Catboost	1.000000
6	Seleccion mutual information	Random Forest	1.000000
13	Seleccion manual	KNN	1.000000
12	Seleccion manual	Random Forest	1.000000
11	Seleccion mutual information	Catboost	1.000000
10	Seleccion mutual information	LightGBM	1.000000
41	Seleccion voting	Catboost	1.000000
40	Seleccion voting	LightGBM	1.000000
29	Seleccion rfe	Catboost	1.000000
30	Seleccion sfs	Random Forest	1.000000
14	Seleccion manual	Decission Tree	1.000000
16	Seleccion manual	LightGBM	1.000000
17	Seleccion manual	Catboost	1.000000
18	Seleccion modelo	Random Forest	1.000000
24	Seleccion rfe	Random Forest	1.000000
19	Seleccion modelo	KNN	1.000000
20	Seleccion modelo	Decission Tree	1.000000
22	Seleccion modelo	LightGBM	1.000000
25	Seleccion rfe	KNN	1.000000
23	Seleccion modelo	Catboost	1.000000
28	Seleccion rfe	LightGBM	1.000000
26	Seleccion rfe	Decission Tree	1.000000
38	Seleccion voting	Decission Tree	1.000000
37	Seleccion voting	KNN	1.000000
36	Seleccion voting	Random Forest	1.000000
35	Seleccion sfs	Catboost	1.000000
32	Seleccion sfs	Decission Tree	1.000000
34	Seleccion sfs	LightGBM	0.998974
1	Seleccion estadistico	KNN	0.997740
7	Seleccion mutual information	KNN	0.996610
31	Seleccion sfs	KNN	0.992509
15	Seleccion manual	AdaBoost	0.971352
21	Seleccion modelo	AdaBoost	0.971352
9	Seleccion mutual information	AdaBoost	0.970753
39	Seleccion voting	AdaBoost	0.970326
27	Seleccion rfe	AdaBoost	0.970326
3	Seleccion estadistico	AdaBoost	0.961181
33	Seleccion sfs	AdaBoost	0.951992

	features_list	model	avg. recall
0	Seleccion estadistico	Logistic Regressor	1.000000
1	Seleccion estadistico	SVC	1.000000
2	Seleccion mutual information	Logistic Regressor	1.000000
3	Seleccion mutual information	SVC	1.000000
4	Seleccion manual	Logistic Regressor	1.000000
5	Seleccion manual	SVC	1.000000
6	Seleccion modelo	Logistic Regressor	1.000000
7	Seleccion modelo	SVC	1.000000
8	Seleccion rfe	Logistic Regressor	1.000000
9	Seleccion rfe	SVC	1.000000
13	Seleccion voting	SVC	1.000000
12	Seleccion voting	Logistic Regressor	1.000000
11	Seleccion sfs	SVC	0.990715
10	Seleccion sfs	Logistic Regressor	0.958561

EVUALANDO CONTRA TEST





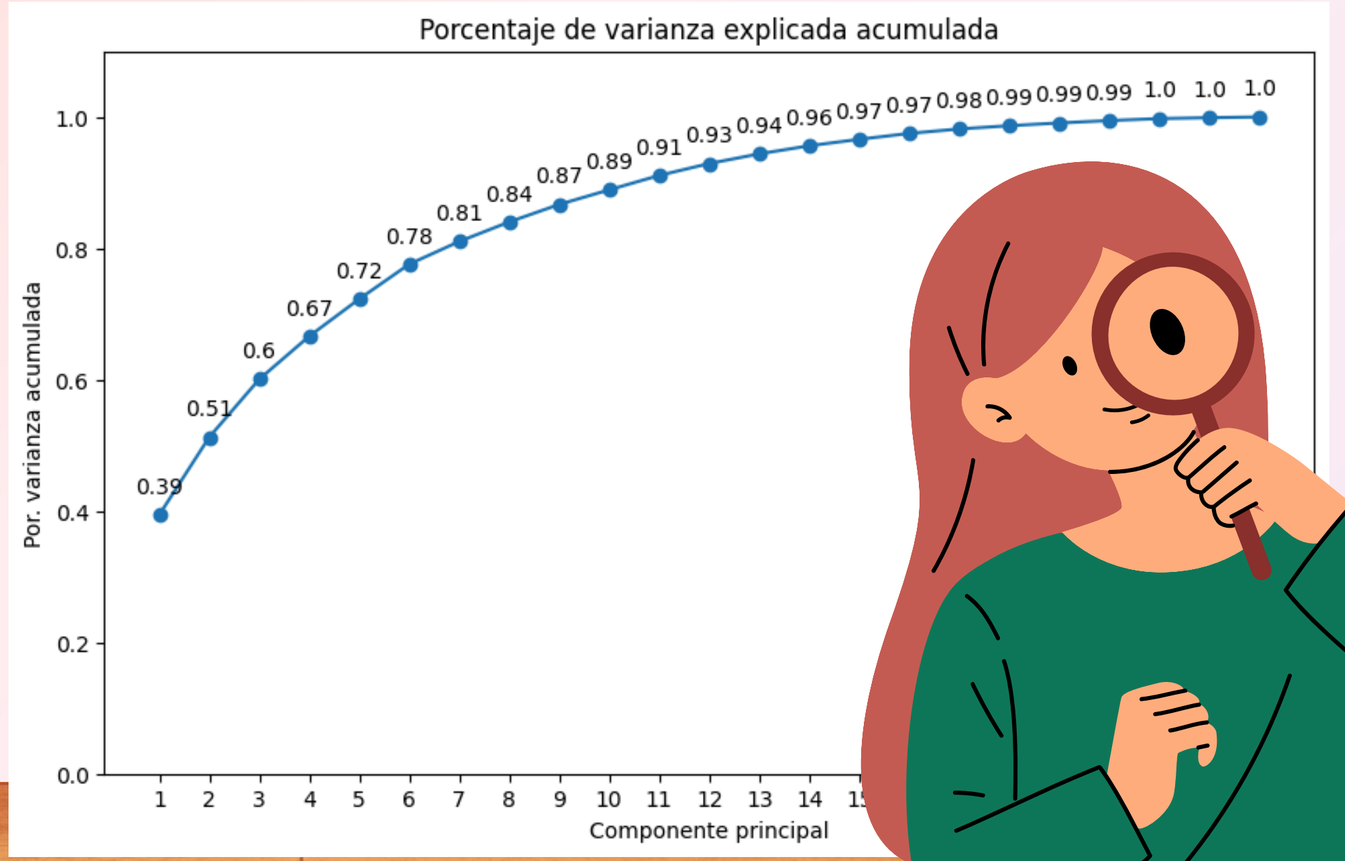


PRIMERAS CONCLUSIONES

- Los modelos en general dan muy buenas métricas
- KNN Selección features Mutual Information
- Sospechas dataset sintético
- Extra

EXTRA: BUSCANDO LAS MÍNIMAS FEATURES

- Sospechas dataset sintético
- PCA
- Selector por umbral de varianza



DOS FEATURES

	features_list	model	avg. recall
6	Seleccion mutual information	Random Forest	0.935722
8	Seleccion mutual information	Decission Tree	0.935722
11	Seleccion mutual information	Catboost	0.935722
10	Seleccion mutual information	LightGBM	0.935722
7	Seleccion mutual information	KNN	0.923081
4	Seleccion estadistico	LightGBM	0.879777
40	Seleccion voting	LightGBM	0.879777
16	Seleccion manual	LightGBM	0.879777
2	Seleccion estadistico	Decission Tree	0.874867
0	Seleccion estadistico	Random Forest	0.874867
14	Seleccion manual	Decission Tree	0.874867
38	Seleccion voting	Decission Tree	0.874867
36	Seleccion voting	Random Forest	0.874867
12	Seleccion manual	Random Forest	0.874867
41	Seleccion voting	Catboost	0.874137
17	Seleccion manual	Catboost	0.874137
5	Seleccion estadistico	Catboost	0.874137
1	Seleccion estadistico	KNN	0.869625
13	Seleccion manual	KNN	0.869625
37	Seleccion voting	KNN	0.869625
39	Seleccion voting	AdaBoost	0.835505
3	Seleccion estadistico	AdaBoost	0.835505
15	Seleccion manual	AdaBoost	0.835505
22	Seleccion modelo	LightGBM	0.826651
20	Seleccion modelo	Decission Tree	0.821010
18	Seleccion modelo	Random Forest	0.821010
23	Seleccion modelo	Catboost	0.821010
19	Seleccion modelo	KNN	0.812325
26	Seleccion rfe	Decission Tree	0.733874
24	Seleccion rfe	Random Forest	0.733874
28	Seleccion rfe	LightGBM	0.733874
29	Seleccion rfe	Catboost	0.729650
27	Seleccion rfe	AdaBoost	0.703844
25	Seleccion rfe	KNN	0.698796
30	Seleccion sfs	Random Forest	0.679418
32	Seleccion sfs	Decission Tree	0.677742
34	Seleccion sfs	LightGBM	0.672443
35	Seleccion sfs	Catboost	0.665124
31	Seleccion sfs	KNN	0.653226
21	Seleccion modelo	AdaBoost	0.646574
9	Seleccion mutual information	AdaBoost	0.635185
33	Seleccion sfs	AdaBoost	0.479467

	features_list	model	avg. recall
3	Seleccion mutual information	SVC	0.922542
1	Seleccion estadistico	SVC	0.872085
5	Seleccion manual	SVC	0.872085
13	Seleccion voting	SVC	0.872085
4	Seleccion manual	Logistic Regressor	0.752798
0	Seleccion estadistico	Logistic Regressor	0.752798
12	Seleccion voting	Logistic Regressor	0.752798
9	Seleccion rfe	SVC	0.730626
7	Seleccion modelo	SVC	0.689324
6	Seleccion modelo	Logistic Regressor	0.685104
2	Seleccion mutual information	Logistic Regressor	0.662217
8	Seleccion rfe	Logistic Regressor	0.595226
11	Seleccion sfs	SVC	0.498287
10	Seleccion sfs	Logistic Regressor	0.426731

- PCA: 4 empieza malos resultados
- Real: 2 features
- Dos peores features : age & gender
- Dos mejores features: Passive Smoker y Coughing of Blood

CONCLUSIONES FINALES Y PERSONALES

- Dataset artificial
- Problema con la “categorización features”
- Buena bateria de selectores de features
- Buena batería de modelos

GRACIAS POR LA ATENCIÓN