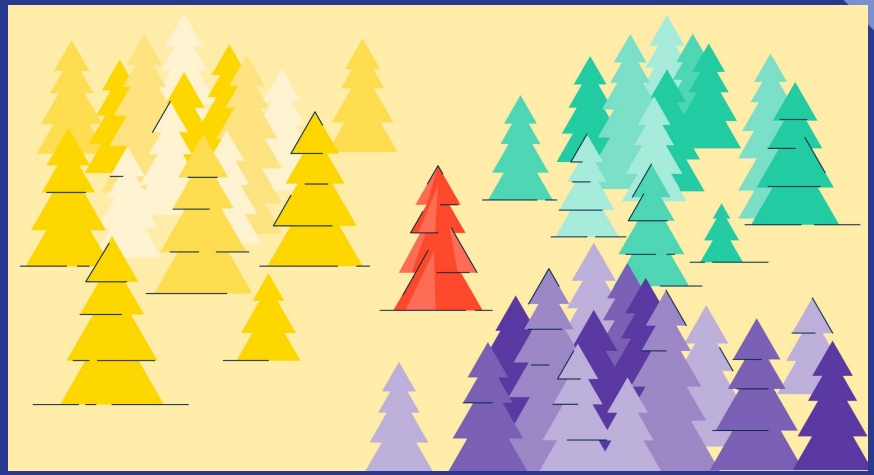
Aprenentatge Computacional

Pràctica 2: Clasificación



Grup GPA603-1530

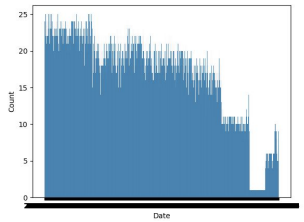
Mario González 1566235
Ferran Bernabé 1564845
Daniel Gutiérrez 1563389

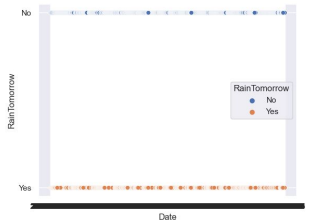
Índice de contenidos

Explicación de la BBDD a analizar

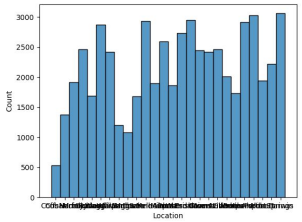
Análisis Base de Datos

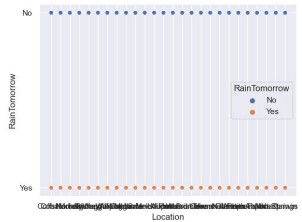
Date (date)



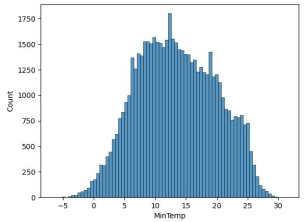


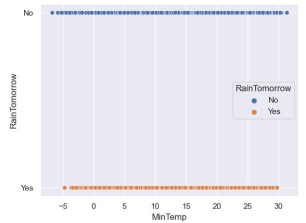
Location (string)





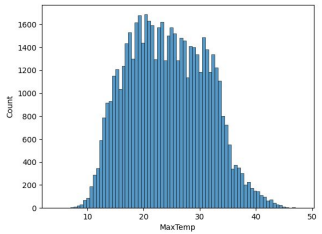
MinTemp (float)

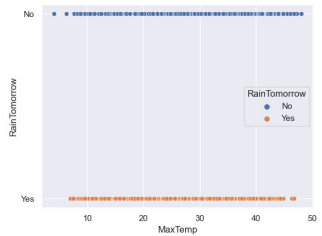




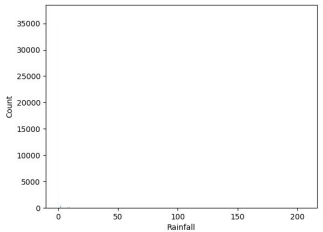
Análisis Base de Datos

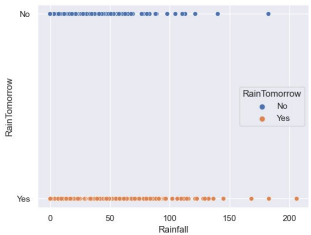
MaxTemp (float)



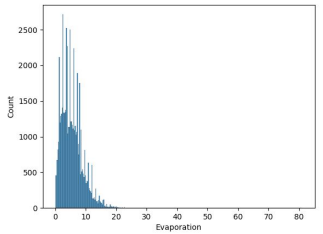


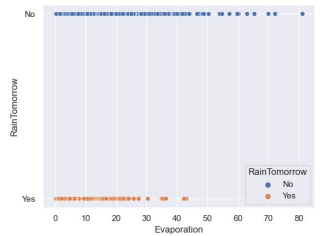
Rainfall (float)





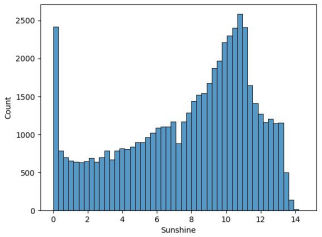
Evaporation (float)



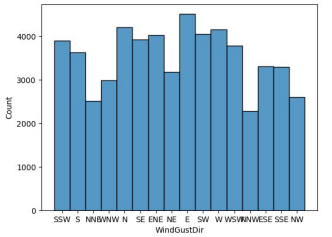


Análisis Base de Datos

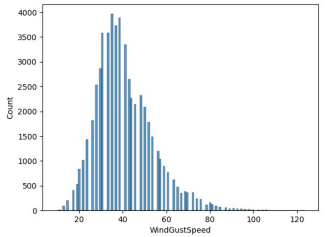
Sunshine (float)



WindGustDir (string)

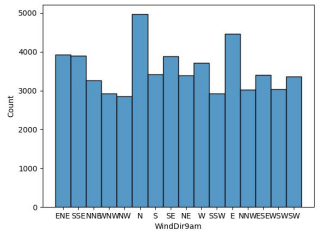


WindGustSpeed (float)

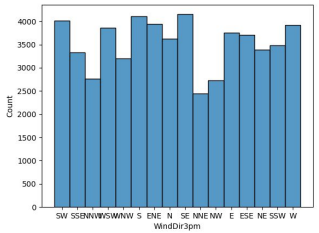


Análisis Base de Datos

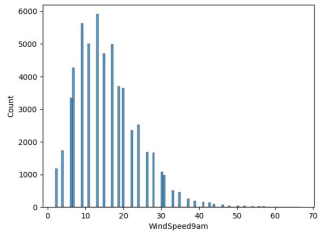
WindDir9am (string)



WindDir3pm (string)

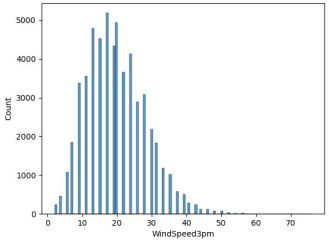


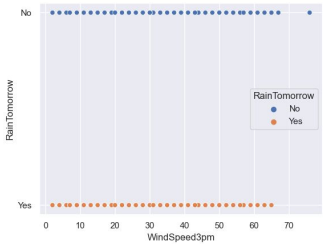
WindSpeed9am (float)



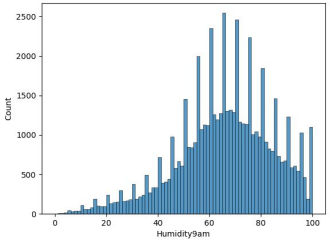
Análisis Base de Datos

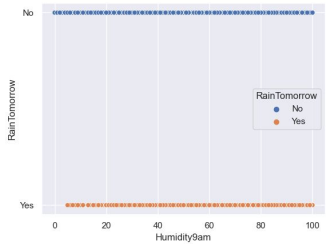
WindSpeed3pm (float)



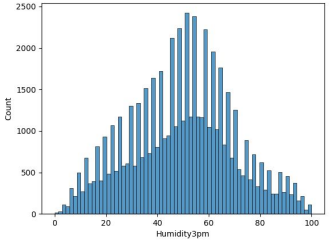


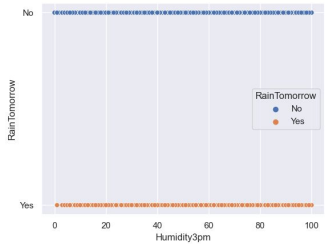
Humidity9am (float)





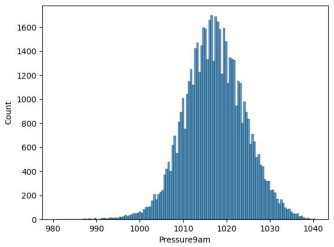
Humidity3pm (float)



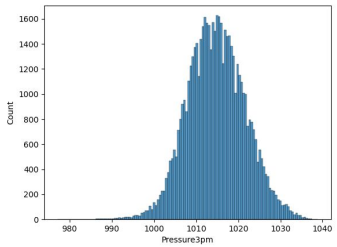


Análisis Base de Datos

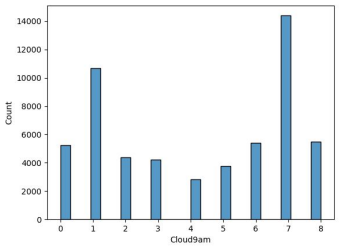
Pressure9am (float)



Pressure3pm (float)

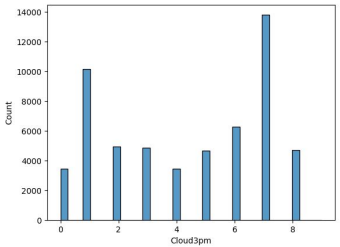


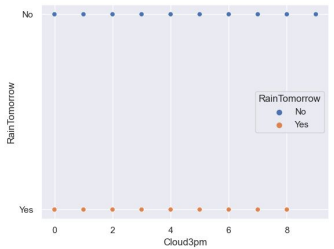
Cloud9am (float)



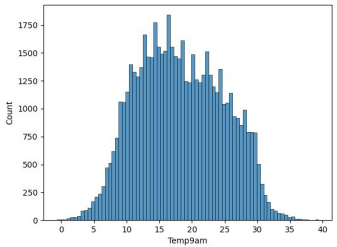
Análisis Base de Datos

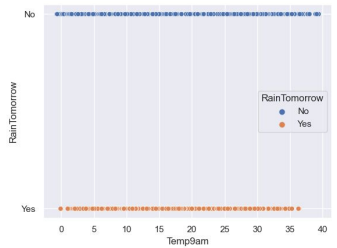
Cloud3pm (float)



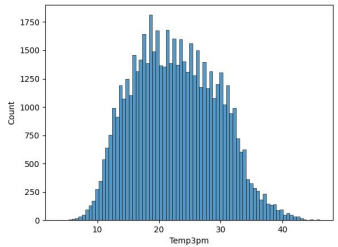


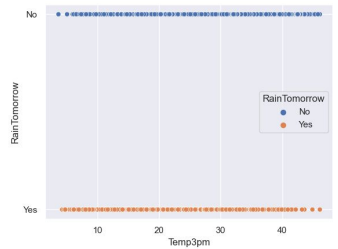
Temp9am (float)





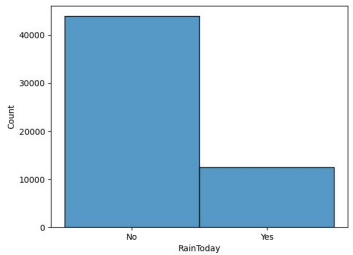
Temp3pm (float)

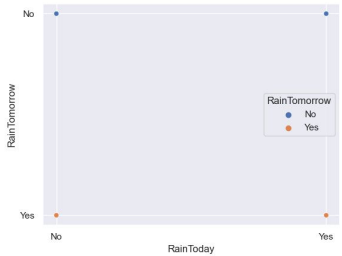




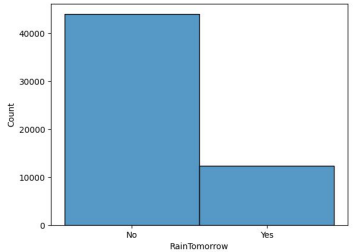
Análisis Base de Datos

RainToday (string)

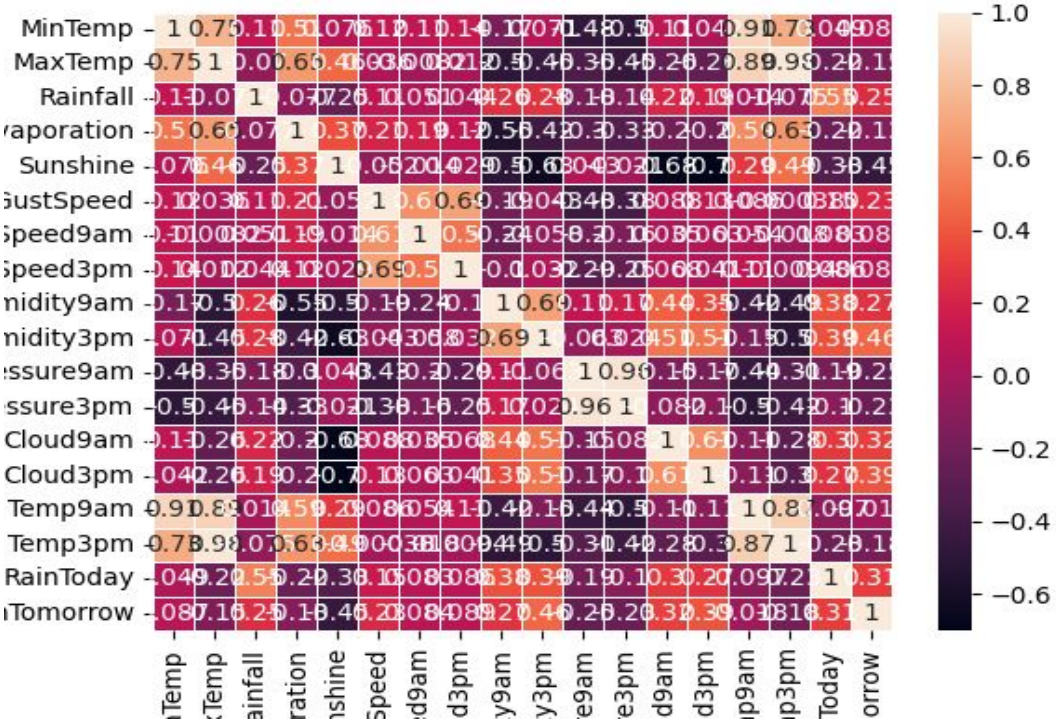




RainTomorrow (string)



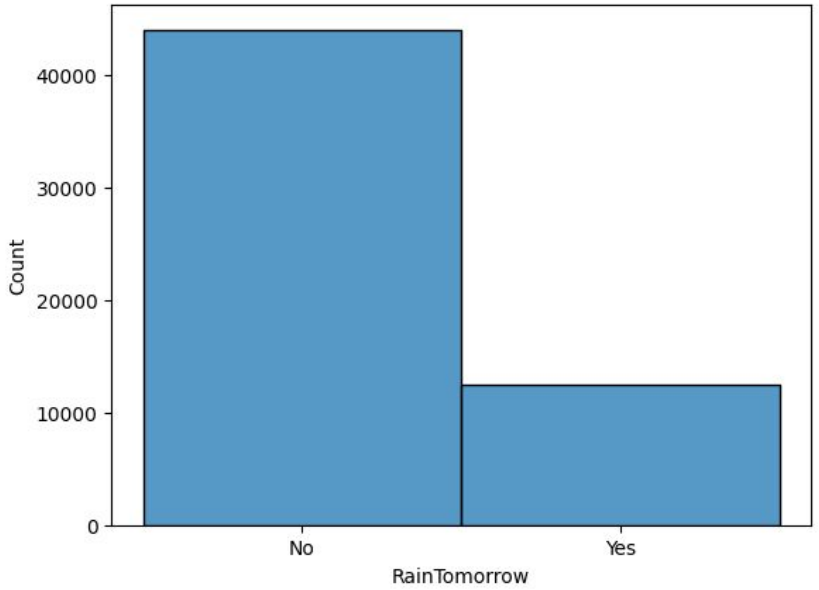
Correlación de los atributos



Etiquetado desbalanceado

No: 0.7797

Yes: 0.2202



Apartado A

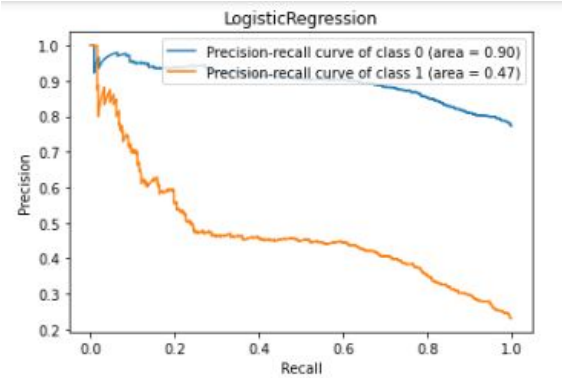
Comparativa de Modelos

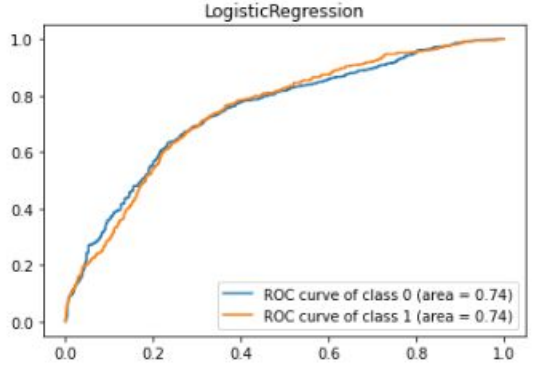
En esta sección hemos aplicado diferentes modelos y métricas para ver cual de ellos da un mejor/peor rendimiento.

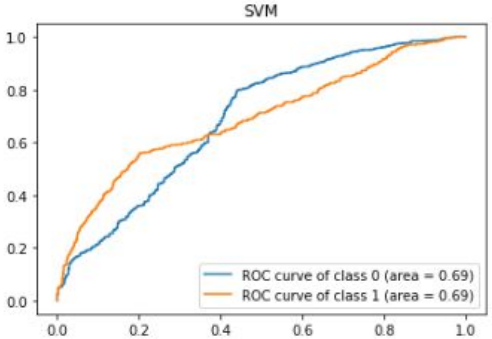
Modelos de selección

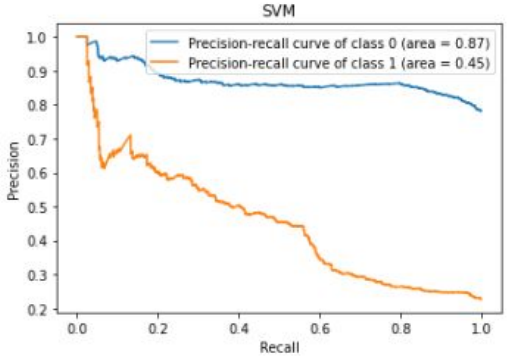
	Logistic Regression	SVM	KNN	Perceptron
50% test 50% validation	0.7892	0.792	0.7644	0.7636
80% test 20% validation	0.794	0.7895	0.772	0.7885
70% test 30% validation	0.7873	0.791	0.7616	0.6683

Resultado de las curvas ROC y PR

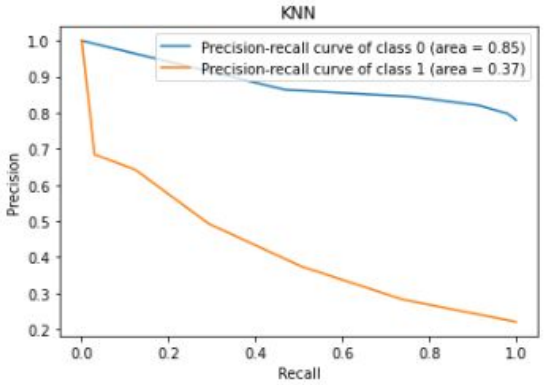


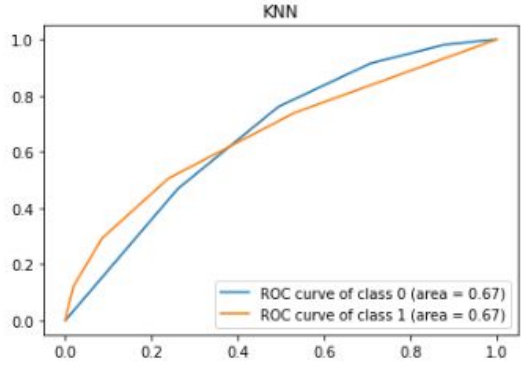


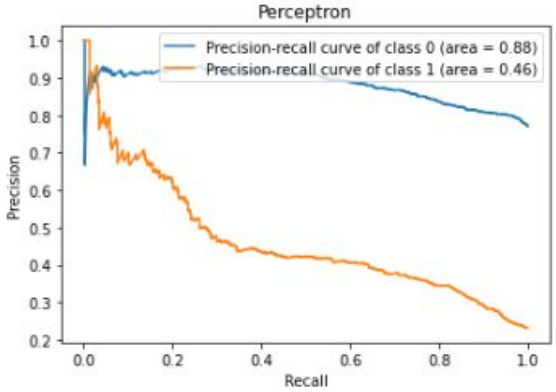


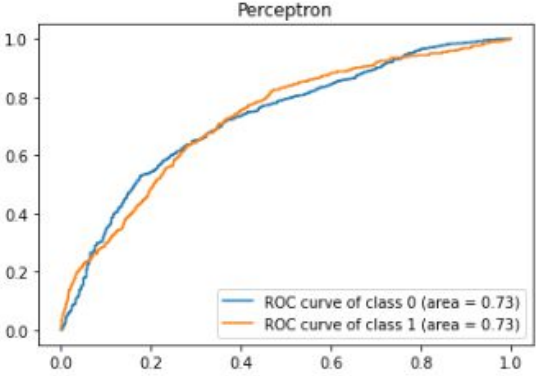


Resultado de las curvas ROC y PR

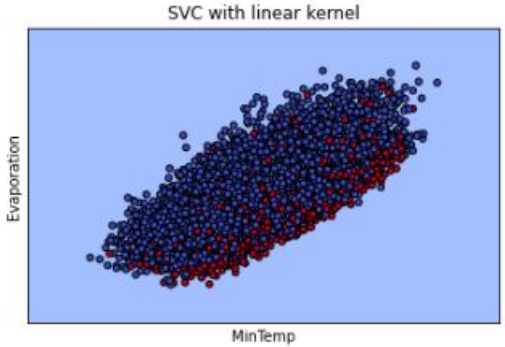


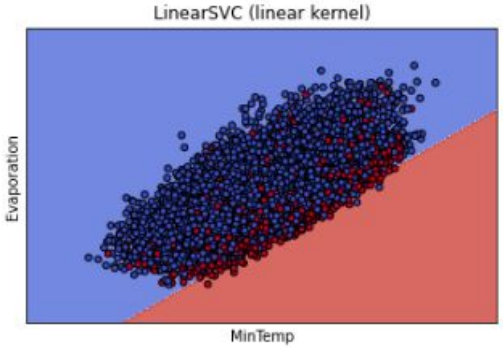


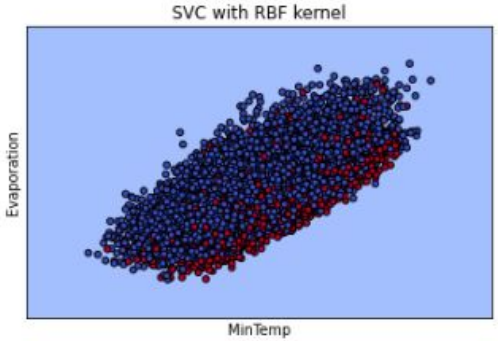


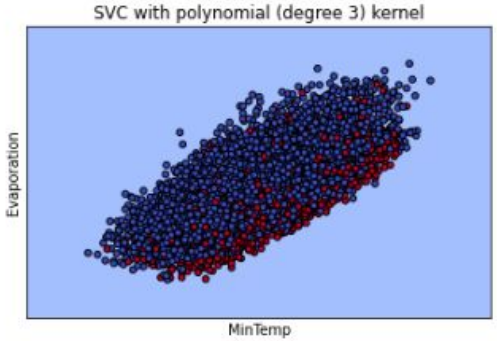


Últimos resultados de los modelos de clasificación









Apartado B

Clasificación numérica

En esta sección analizaremos nuestra base de datos para conocer el modelo más adecuado para clasificar nuestro atributo objetivo

Preprocesado, normalización y outliers

Normalización de los datos

- Valores de órdenes de magnitud parecidos
- **Escalar los datos** → dar la misma importancia a todos los atributos

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	Win
count	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	56420.000000	5
mean	13.464770	24.219206	2.130397	5.503135	7.735626	40.877366	15.667228	
std	6.416689	6.970676	7.014822	3.696282	3.758153	13.335232	8.317005	
min	-6.700000	4.100000	0.000000	0.000000	0.000000	9.000000	2.000000	
25%	8.600000	18.700000	0.000000	2.800000	5.000000	31.000000	9.000000	

Limpiado del dataset

- Borrar filas con algún NaN, al disponer de filas de sobra

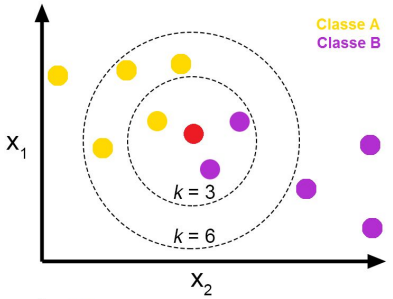
- Borrar / convertir atributos no numéricos, para poder clasificar

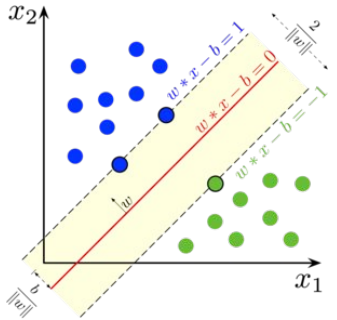
Date	0
Location	0
MinTemp	1485
MaxTemp	1261
Rainfall	3261
Evaporation	62790
Sunshine	69835
WindGustDir	10326
WindGustSpeed	10263
WindDir9am	10566
WindDir3pm	4228
WindSpeed9am	1767
WindSpeed3pm	3062
Humidity9am	2654
Humidity3pm	4507
Pressure9am	15065
Pressure3pm	15028
Cloud9am	55888
Cloud3pm	59358
Temp9am	1767
Temp3pm	3609
RainToday	3261
RainTomorrow	3267

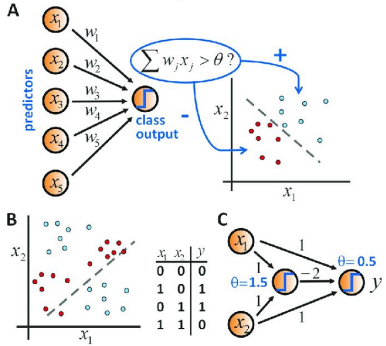
Modelo de Selección

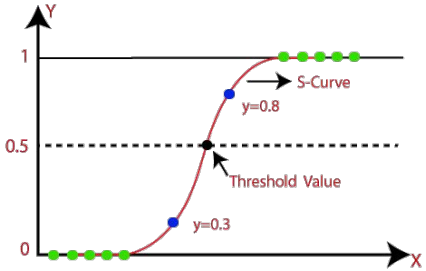
Modelo de selección

- Ejecución de diferentes conjuntos de **test - validation** en los datos
- Modelos utilizados:
  - KNN
  - Perceptrón
  - SVM
  - Linear Regression









Crossvalidation

Crossvalidation

Regressió Logística	Precisió	C	fit_intercept	penalty	tolerance
Bàsica					
50% test 50%val	85,38	2.0	True	None	0.001
80% test 20% val	85,31	2.0	True	None	0.001
70% test 30% val	85,31	2.0	True	None	0.001
K-Fold					
K = 2	0,8507	2.0	True	None	0.001
K = 3	0,8517	2.0	True	None	0.001
K = 4	0,8530	2.0	True	None	0.001
K = 5	0,8532	2.0	True	None	0.001
K = 6	0,8518	2.0	True	None	0.001
LOOVC	0,8532	2.0	True	None	0.001

Crossvalidation

KNN	Precisió	leaf_size	n_neighbors	metric	p
Bàsica					
50% test 50%val	83,91	30	5	minkowski	2
80% test 20% val	84,45	30	5	minkowski	2
70% test 30% val	84,45	30	5	minkowski	2
K-Fold					
K = 2	0.8270	30	5	minkowski	2
K = 3	0.8258	30	5	minkowski	2
K = 4	0.8325	30	5	minkowski	2
K = 5	0.8305	30	5	minkowski	2
K = 6	0.8315	30	5	minkowski	2
LOOVC	0,8445	2.0	True	None	0.001

Crossvalidation

SVM	Precisió	C	fit_intercept	penalty	tolerance
Bàsica					
50% test 50%val	85,35	2.0	True	None	1e-5
80% test 20% val	85,32	2.0	True	None	1e-5
70% test 30% val	**85,59**	2.0	True	None	1e-5
K-Fold					
K = 2	0,8509	1.0	True	None	1e-5
K = 3	0,8520	1.0	True	None	1e-5
K = 4	0,8530	1.0	True	None	1e-5
K = 5	0,8529	1.0	True	None	1e-5
K = 6	0,8522	1.0	True	None	1e-5
LOOVC	0,8530	1.0	True	None	1e-5

Crossvalidation

Perceptró	Precisió	**α**	fit_intercept	penalty	tolerance
Bàsica					
50% test 50%val	80,50	0.0001	True	None	0.001
80% test 20% val	78,33	0.0001	True	None	0.001
70% test 30% val	80,74	0.0001	True	None	0.001
K-Fold					
K = 2	0,8522	0.0001	True	None	0.001
K = 3	0,7886	0.0001	True	None	0.001
K = 4	0,7789	0.0001	True	None	0.001
K = 5	0,7981	0.0001	True	None	0.001
K = 6	0,8049	0.0001	True	None	0.001
LOOVC	0,8522	0.0001	True	None	0.001

Crossvalidation

- Validación del entrenamiento de los modelos, verificar si este ha sido efectivo
- Utiliza **accuracy**

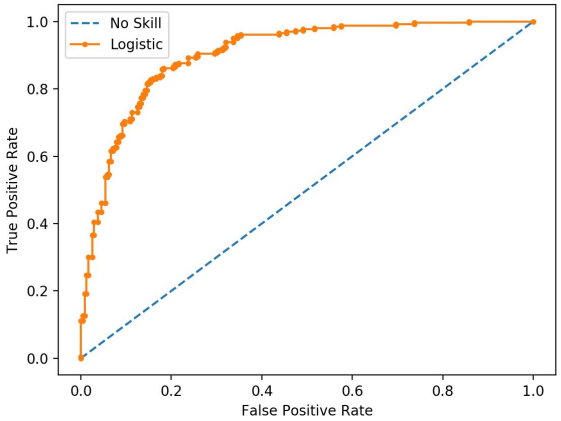
Test / Validació (%)	**Regressió logística**	**KNN**	**SVM**	**Perceptró**
50 / 50	85,38	83,91	85,35	80,50
80 / 20	85,31	84,45	85,32	78,33
70 / 30	85,31	84,45	**85,59**	80,74

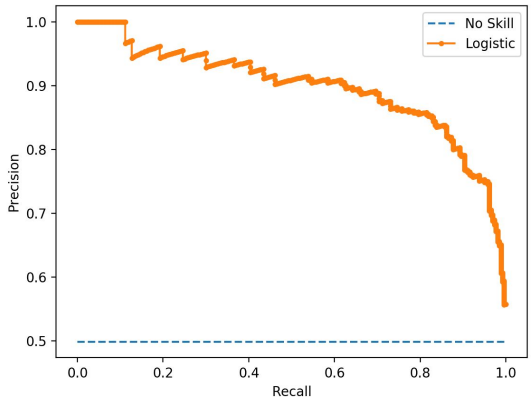
70 % Test
30 % Validació

Metric analysis

Metric analysis

- Visualización de diferentes métricas: accuracy, f1_score, average_precision, recall …
- ROC / PR curves

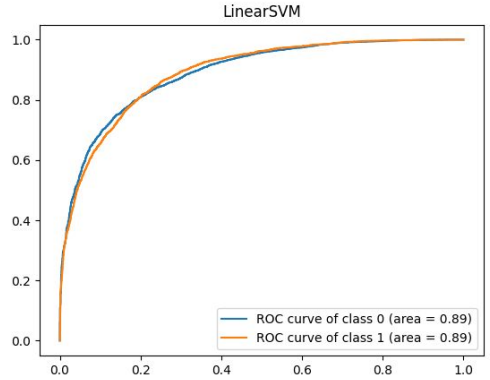


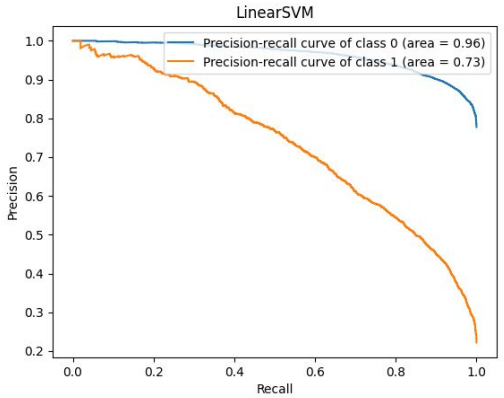


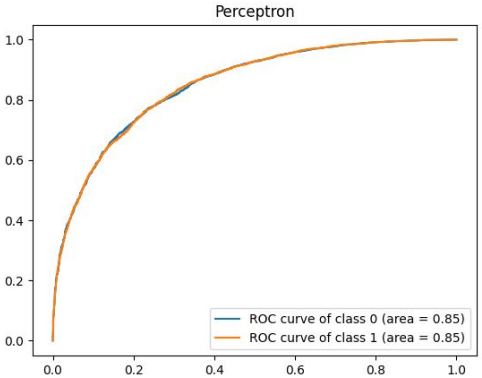
Metric analysis

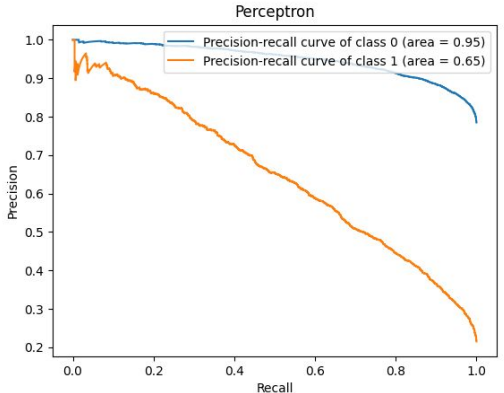
Model	Accuracy score	F1 score	Average Precision Score
SVM	0,8336	0,7931	0,3715
Perceptró	0,7132	0,7373	0,4069
KNN	0,8423	0,8347	0,4673
Regressió Logística	0,8521	0,8436	0,4891

ROC / PR curves

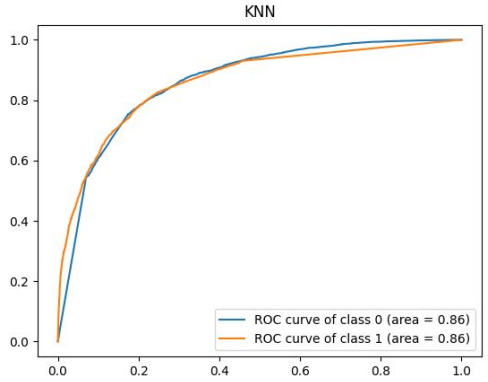


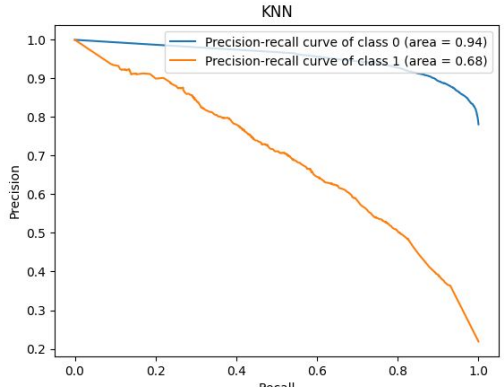


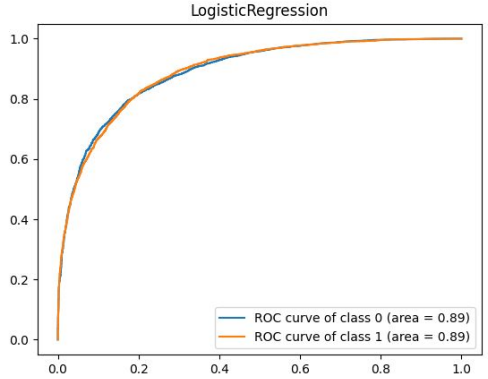


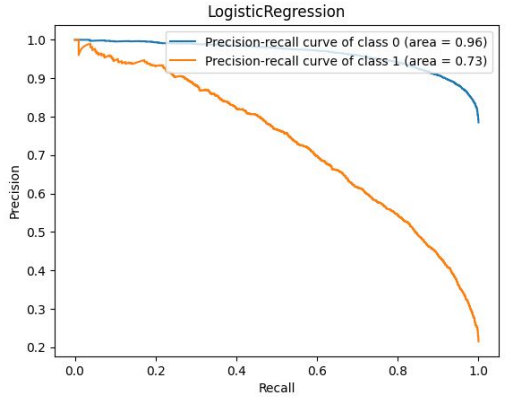


ROC / PR curves









ROC / PR curves resultados

Model	ROC	PR
SVM	0.89	0.96 (class 0)
Perceptron	0.85	0.95 (class 0)
KNN	0.86	0.94 (class 0)
Logistic Regression	0.89	0.96 (class 0)

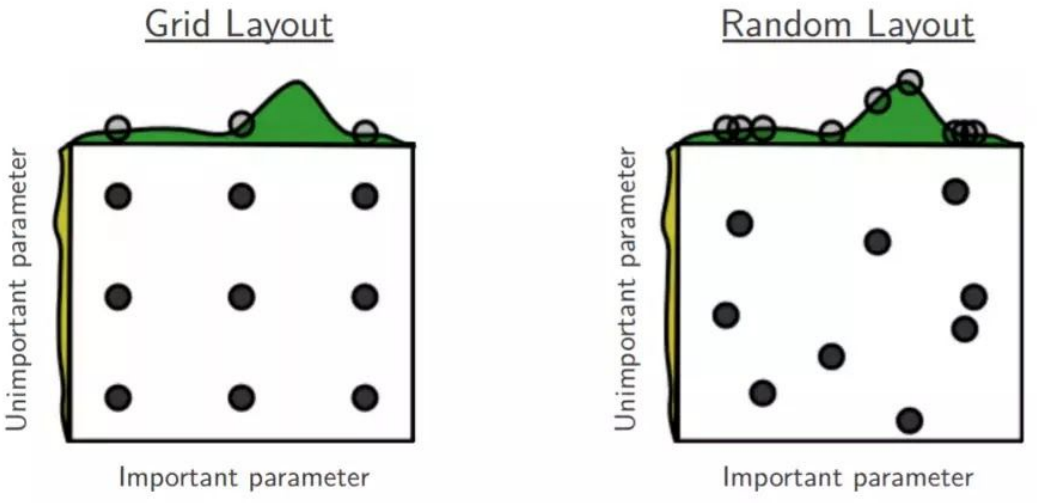
- Todos los modelos dan resultados muy parecidos
- Podemos destacar **ROC** de **SVM** y **Logistic Regression**

Hyperparameter search

Hyperparameter search

- Prueba conjunto de hiperparametros en un modelo de forma exhaustiva para encontrar los mejores

- Métodos más utilizados **GridSearchCV** y **RandomizedSearchCV**



Hyperparameter search

- Busqueda de los mejores hiperparametros en nuestro conjunto de datos

- Utilizaremos **Exhaustive Grid Search (GridSearchCV)**

Hyperparameter search

Perceptron	{ 'penalty': ['l2','l1'], 'alpha': [0.0001, 0.001, 0.01, 0.1, 1], 'fit_intercept': [True, False], 'shuffle': [True, False] }
KNN	{ 'n_neighbors' : [5,7,9,11,13], 'weights' : ['uniform','distance'], 'metric' ['minkowski','euclidean','manhattan'] }
Logistic Regression	{'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000] }
SVM	{'C': [0.1,1, 10, 100, 1000]}

Hyperparameter search

El algoritmo **GridSearchCV** ha escogido los siguientes como mejores hiperparámetros:

Perceptron	{
	'penalty': **'l1',** 'alpha': **0.0001**,
	'fit_intercept': **True**, 'shuffle': **True**
	}

KNN	{ 'n_neighbors' : **13**, 'weights' : **'uniform'**, 'metric': **'manhattan'** }
Logistic Regression	{'C': **1000** }
SVM	{'C': **10** }

Hyperparameter search

Mejora de los modelos con los mejores hiperparámetros

Model	Accuracy default parameters	Accuracy with best parameters (GridSearchCV)	Improvement
Perceptron	0.7132	0.8224	15.31%
KNN	0.8423	0.8413	-0.11%
Logistic Regression	0.8521	0.8506	-0.17%
SVM	0.8336	0.8333	-0.03%

Conclusiones

Modelos más destacados: **SVM** y **Logistic Regression**

Búsqueda de hiperparámetros mediante **GridSearchCV** → no mejoran todos los métodos (Perceptron)

Los atributos que mas ayudan a predecir la precipitación (RainTomorrow) son aquellos **relacionados con la lluvia.**