Startup Success Prediction Model

David Adrián Rodríguez García & Víctor Caínzos López

Preprocesado: Preparación de los datos

En primer lugar se realizará un análisis del dataset **startup_data.csv** para el cual se realizará la limpieza de los datos espurios o nulos y se procederá al filtrado de las columnas representativas y la recodificación de variables cualitativas a cuantitativas.

Data Missing dataframe: Contiene los datos eliminados del dataset original.

	feature	missing	(%) of total
0	closed_at	587	63.67
1	age_first_milestone_year	152	16.49
2	age_last_milestone_year	152	16.49
3	state_code.1	1	0.11

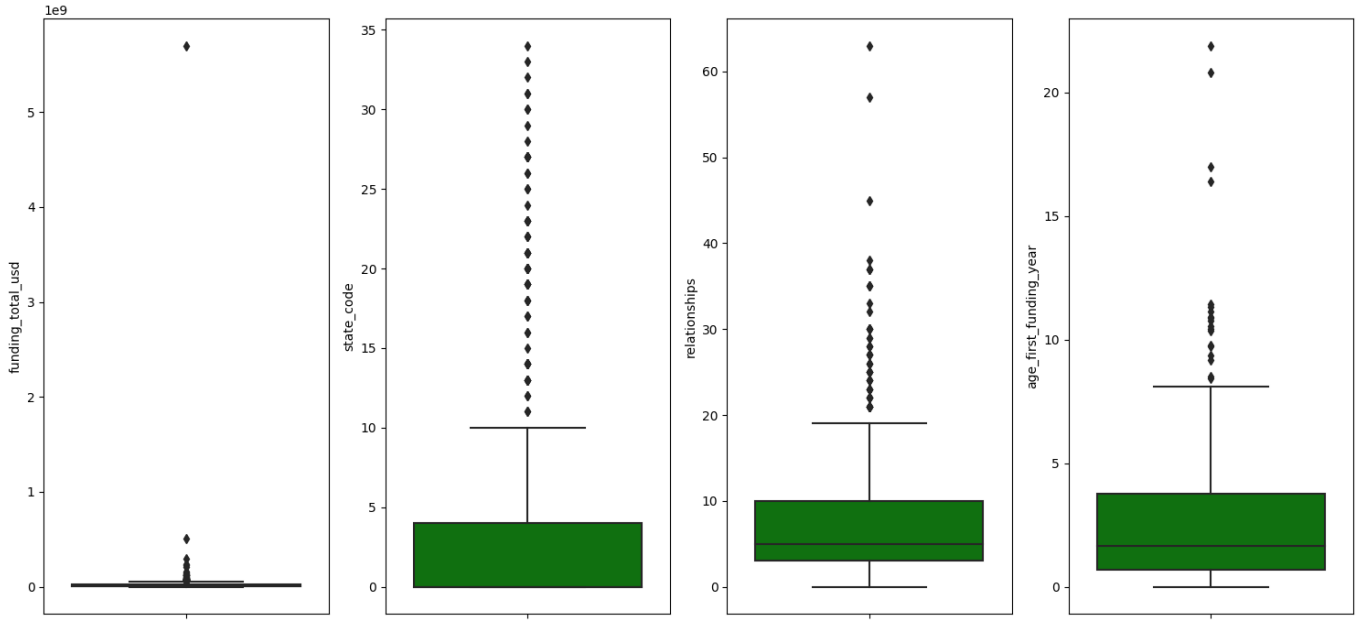
Data Spurious dataframe: Contiene los datos sin sentido del dataset original.

	age_first_funding_year	age_last_funding_year	age_first_milestone_year	age_last_milestone_year	age
88	0.8822	0.8822	0	0	-8
558	-9.0466	-9.0466	-6.0466	-3.8822	-4
73	1.6685	9.337	7.3808	10.474	-2
350	0.3288	0.3288	-0.4192	-0.4192	0
690	0	0.6904	0	0.6904	0

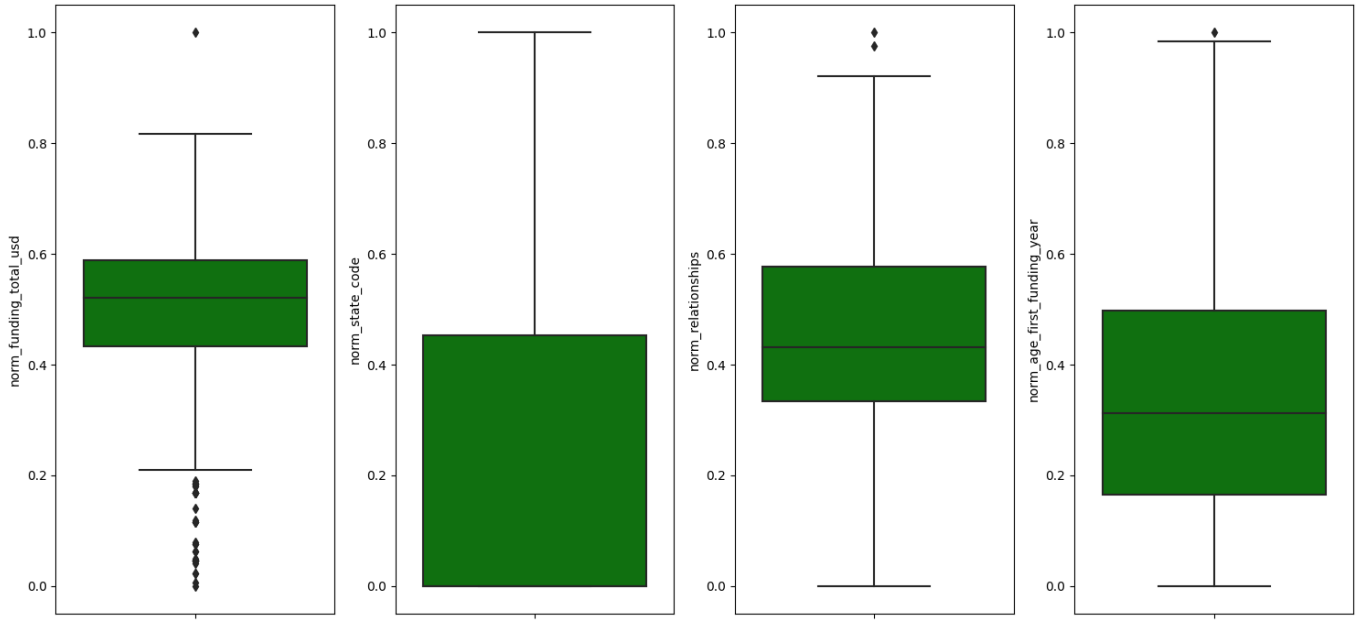
Data Skewness dataframe: Contiene los datos con alta dispersión del dataset original.

	feature	skewness
0	funding_total_usd	27.7868
1	state_code	2.54287
2	relationships	2.33271
3	age_first_funding_year	2.32396
4	avg_participants	1.75802

Boxplot Feature Skewness > 2: Muestra la dispersión de los datos para las características con asimetría mayor que 2.



Boxplot Norm Features: Muestra la dispersión de los datos para las características normalizadas.



X dataframe: Contiene la matriz de características.

---

	state_code	age_last_funding_year	age_first_milestone_year	age_last_milestone_year	funding_rounds
0	0	3.0027	4.6685	6.7041	3
1	0	9.9973	7.0055	7.0055	4
2	0	1.0329	1.4575	2.2055	1
3	0	5.3151	6.0027	6.0027	3
4	0	1.6685	0.0384	0.0384	2

	milestones	is_CA	is_NY	is_MA	is_TX
0	3	1	0	0	0

	milestones	is_CA	is_NY	is_MA	is_TX
1	1	1	0	0	0
2	2	1	0	0	0
3	1	1	0	0	0
4	1	1	0	0	0

	is_otherstate	is_software	is_web	is_mobile	is_enterprise
0	0	0	0	0	0
1	0	0	0	0	1
2	0	0	1	0	0
3	0	1	0	0	0
4	0	0	0	0	0

	is_advertising	is_gamesvideo	is_ecommerce	is_biotech	is_consulting
0	0	0	0	0	0
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	1	0	0	0

	is_othercategory	has_VC	has_angel	has_roundA	has_roundB
0	1	0	1	0	0
1	0	1	0	0	1
2	0	0	0	1	0
3	0	0	0	0	1
4	0	1	1	0	0

	has_roundC	has_roundD	avg_participants	is_top500	age
0	0	0	1	0	7
1	1	1	4.75	1	14
2	0	0	4	1	5
3	1	1	3.3333	1	12
4	0	0	1	1	2

	norm_funding_total_usd	norm_age_first_funding_year	norm_relationships
0	0.268198	0.376383	0.333333
1	0.623283	0.57891	0.553655
2	0.415358	0.226596	0.430827
3	0.623093	0.453101	0.430827
4	0.36268	0	0.26416

t dataframe: Contiene el vector de etiquetas.

	labels
0	1
1	1
2	1
3	1
4	0

Entrenamiento: Comparativa de modelos de aprendizaje automático

Se procederá a comparar los resultados obtenidos de diferentes modelos de aprendizaje automático variando tanto el tipo de modelo como los hiperparámetros de los que depende con el objetivo de obtener el mejor modelo que prediga el éxito o fracaso de las diferentes startups

Results dataframe: Muestra los resultados de los mejores modelos obtenidos

Folds	KNN_train_accuracy	KNN_val_accuracy	SVC_train_accuracy	SVC_val_accuracy	DNN_train_accuracy
1	0.760748	0.791045	0.656075	0.664179	0.741308
2	0.773832	0.731343	0.654206	0.649254	0.809252
3	0.764486	0.723881	0.654206	0.656716	0.782523
4	0.783178	0.80597	0.656075	0.649254	0.786355
5	0.779851	0.744361	0.658582	0.654135	0.805131

Folds	DNN_val_accuracy	KNN_train_recall	KNN_val_recall	SVC_train_recall	SVC_val_recall
1	0.685448	0.968391	0.943182	1	1
2	0.723134	0.948424	0.954023	1	1
3	0.834701	0.948424	0.988506	1	1
4	0.779104	0.962751	0.977011	1	1
5	0.809774	0.962751	0.942529	1	1

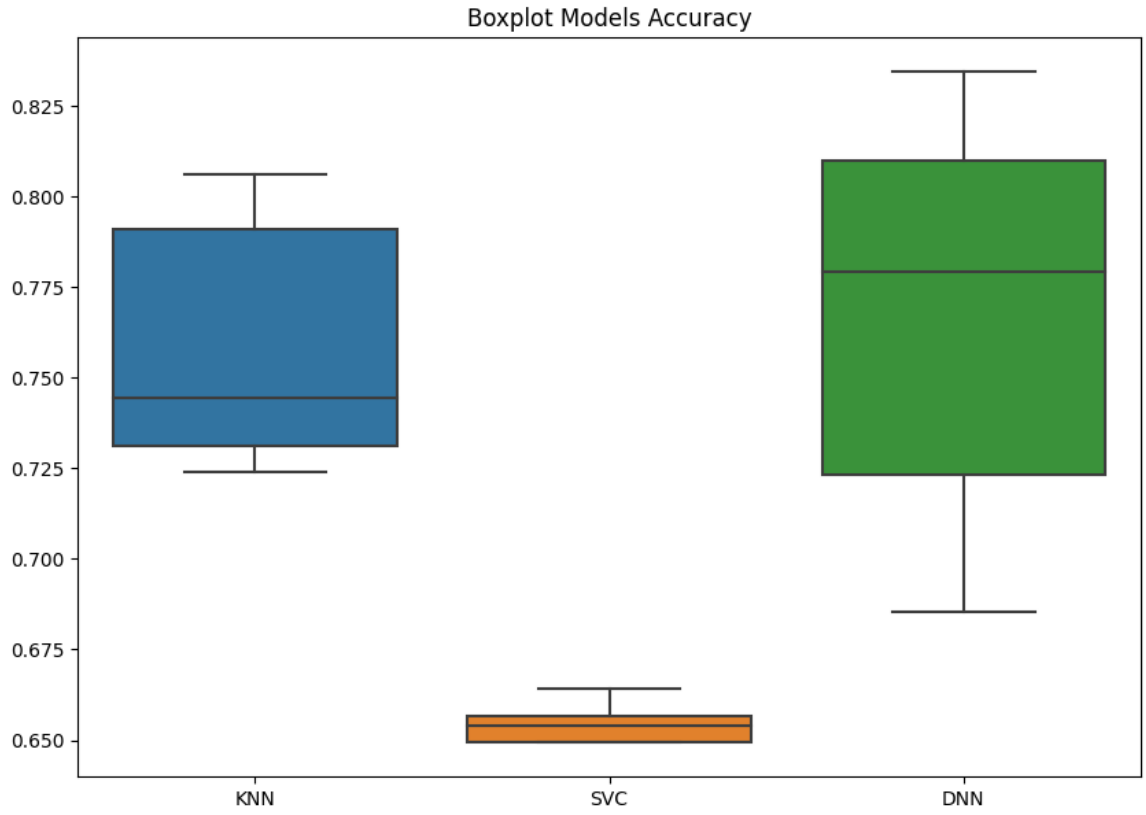
Folds	DNN_train_recall	DNN_val_recall	KNN_train_specificity	KNN_val_specificity	SVC_train_specificity
1	0.741308	0.685448	0.374332	0.5	0.0160428
2	0.809252	0.723134	0.446237	0.319149	0.00537634
3	0.782523	0.834701	0.419355	0.234043	0.00537634
4	0.786355	0.779104	0.446237	0.489362	0.0107527
5	0.805131	0.809774	0.438503	0.369565	0.0213904

Folds	SVC_val_specificity	DNN_train_specificity	DNN_val_specificity	KNN_train_precision	KNN_val_precision
1	0.0217391	0.739167	0.675625	0.742291	0.783019
2	0	0.810327	0.72808	0.762673	0.721739
3	0.0212766	0.784196	0.830804	0.753986	0.704918

Folds	SVC_val_specificity	DNN_train_specificity	DNN_val_specificity	KNN_train_precision	KNN_val_precision
4	0	0.787113	0.757545	0.765376	0.779817
5	0	0.804375	0.839928	0.761905	0.738739

Folds	SVC_train_precision	SVC_val_precision	DNN_train_precision	DNN_val_precision
1	0.654135	0.661654	0.741308	0.685448
2	0.653558	0.649254	0.809252	0.723134
3	0.653558	0.654135	0.782523	0.834701
4	0.654784	0.649254	0.786355	0.779104
5	0.656015	0.654135	0.805131	0.809774

Boxplot models: Muestra los valores de exactitud de los diferentes modelos



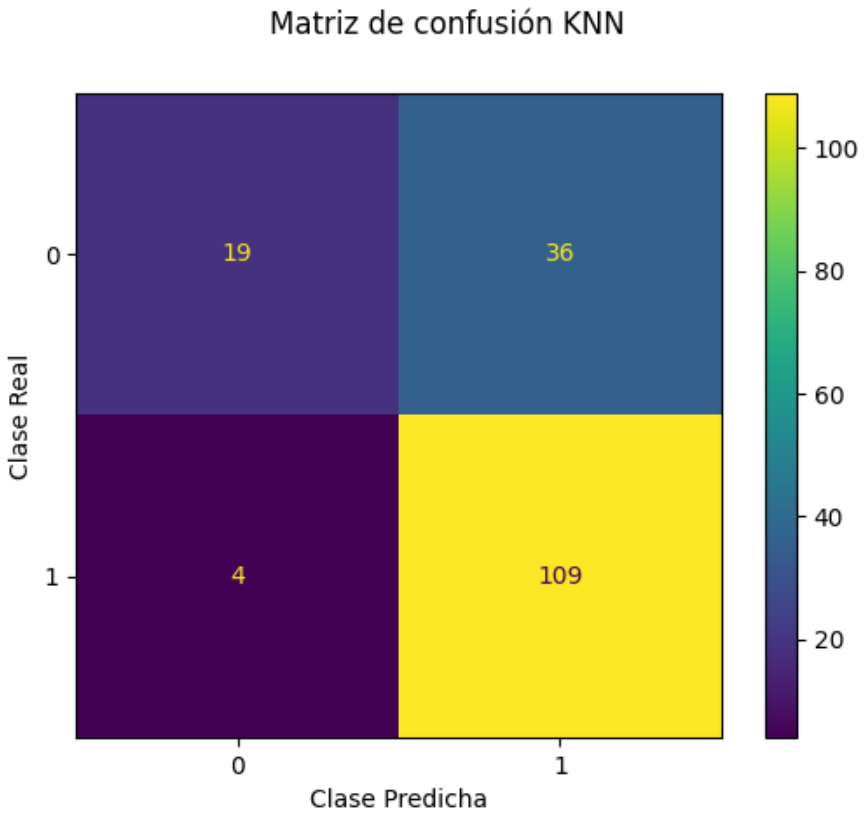
Contraste de hipótesis: Comparación de modelos mediante el test de Kruskal-Wallis

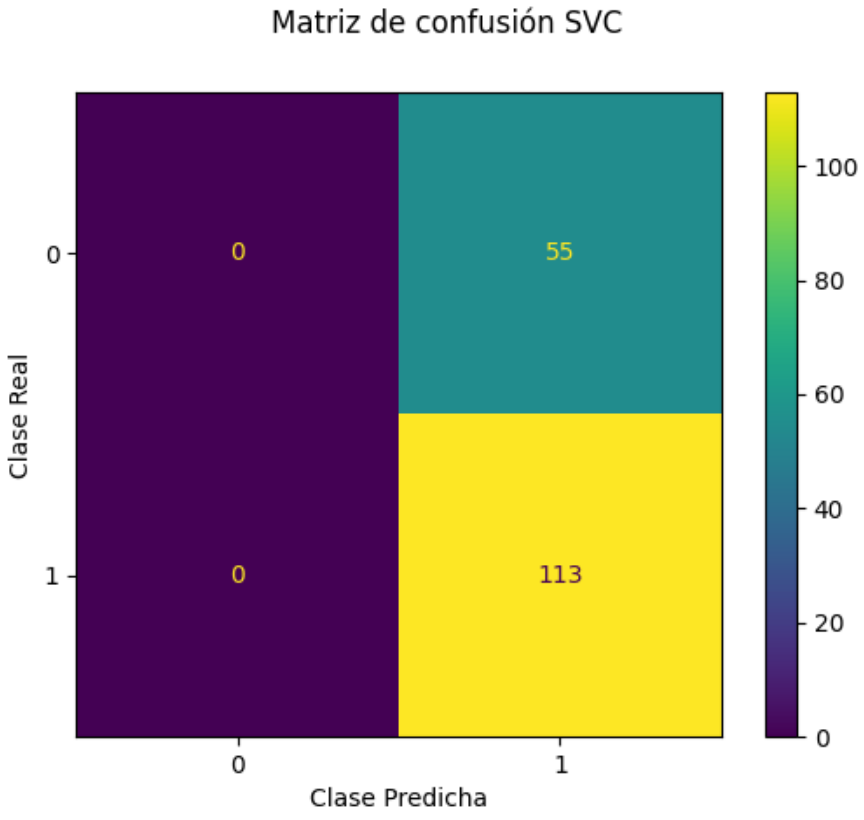
```
p-valor KrusW:0.009109932455060572
Hypotheses are being rejected: the models are different
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
 group1   group2  meandiff p-adj   lower   upper  reject
-----------------------------------------------------
modelDNN modelKNN  -0.0071    0.9 -0.0772   0.063  False
```

```
modelDNN modelSVC  -0.1117  0.003 -0.1818 -0.0416   True
modelKNN modelSVC  -0.1046 0.0048 -0.1747 -0.0345   True
--------------------------------------------------------
```

Matrices de confusión: Compara los valores reales con los valores predichos para cada modelo

---



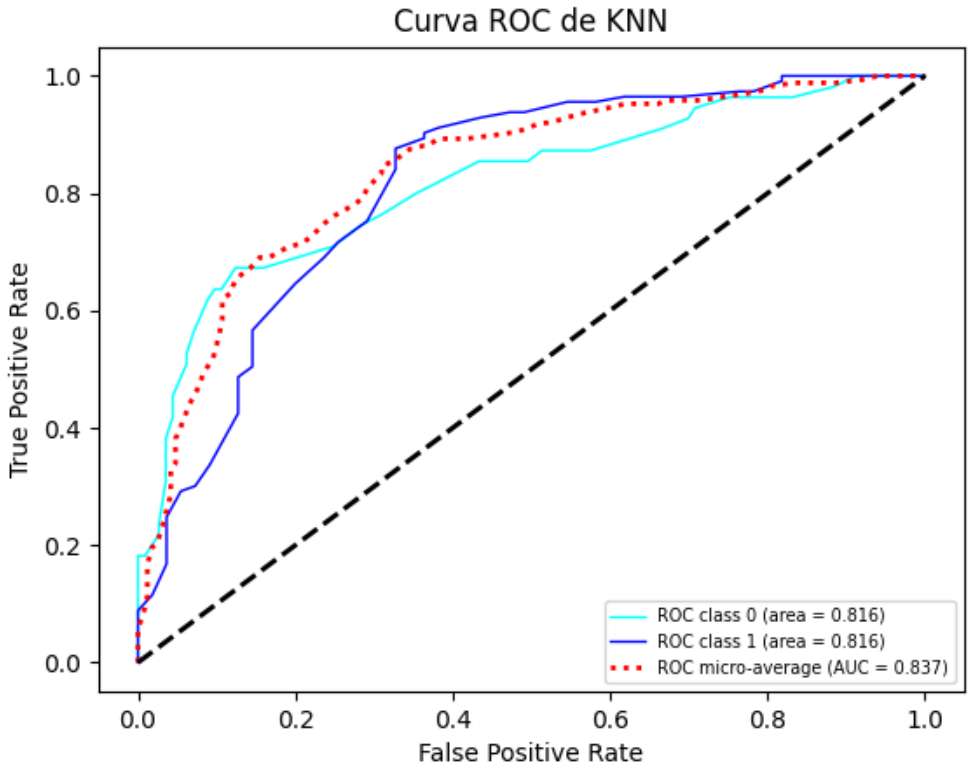


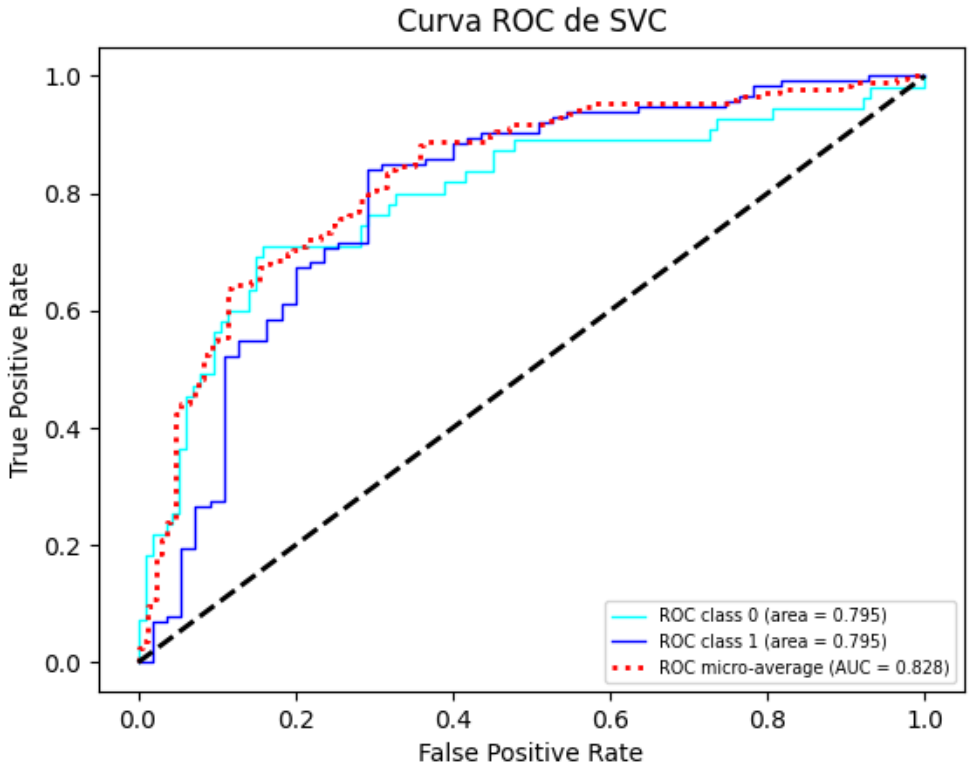
Matriz de confusión DNN

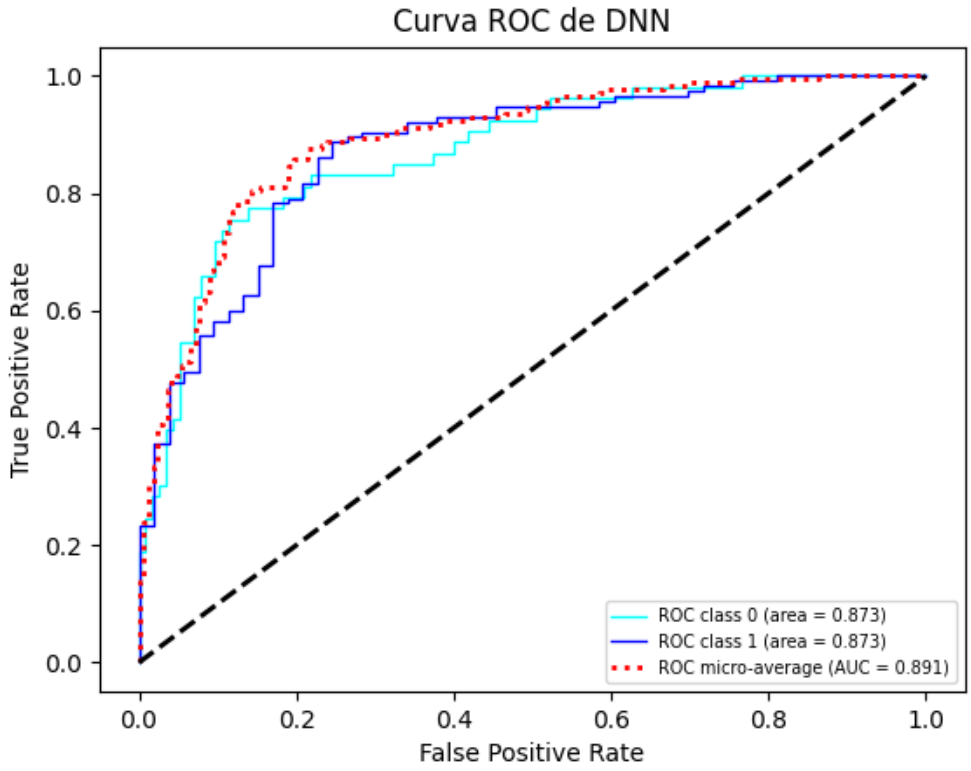
	Clase Real 0	Clase Real 1
Clase Predicha 0	29	24
Clase Predicha 1	7	108

Curva ROC: Compara el ajuste entre la especificidad y la sensibilidad para cada modelo

---







Informe de clasificación: Compara los resultados de cada modelo de clasificación

---

```
Classification report for model KNN:

              precision    recall  f1-score   support

           0       0.83      0.35      0.49        55
           1       0.75      0.96      0.84       113

    accuracy                           0.76       168
   macro avg       0.79      0.66      0.67       168
weighted avg       0.78      0.76      0.73       168
```

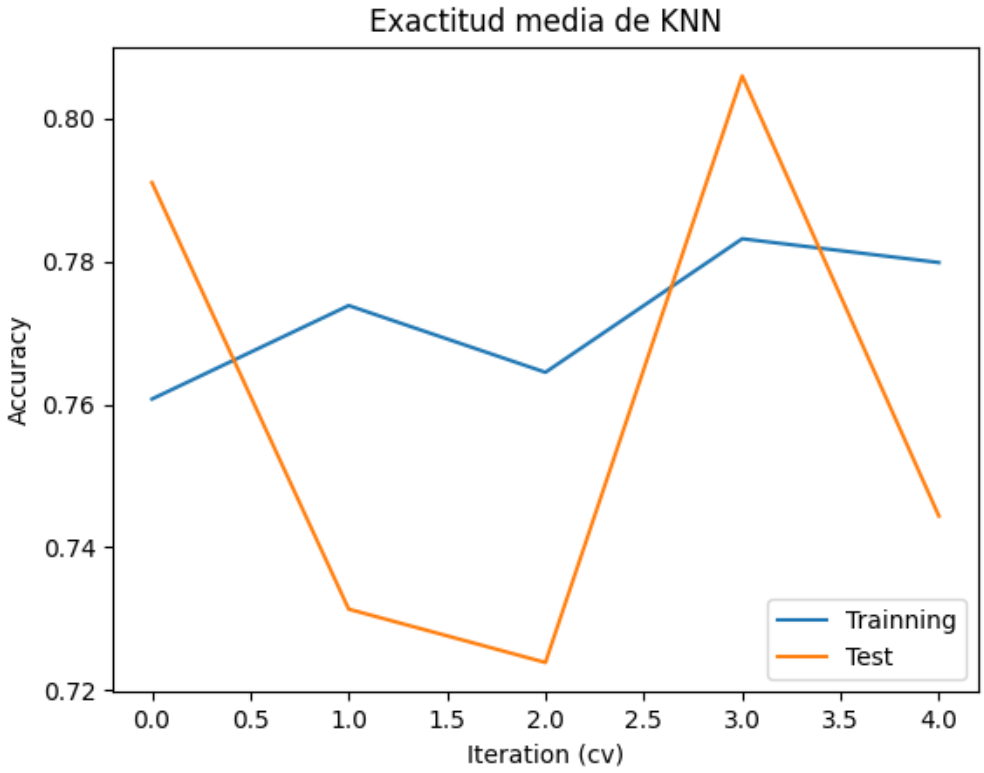
```
Classification report for model SVC:

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        55
           1       0.67      1.00      0.80       113

    accuracy                           0.67       168
   macro avg       0.34      0.50      0.40       168
weighted avg       0.45      0.67      0.54       168
```

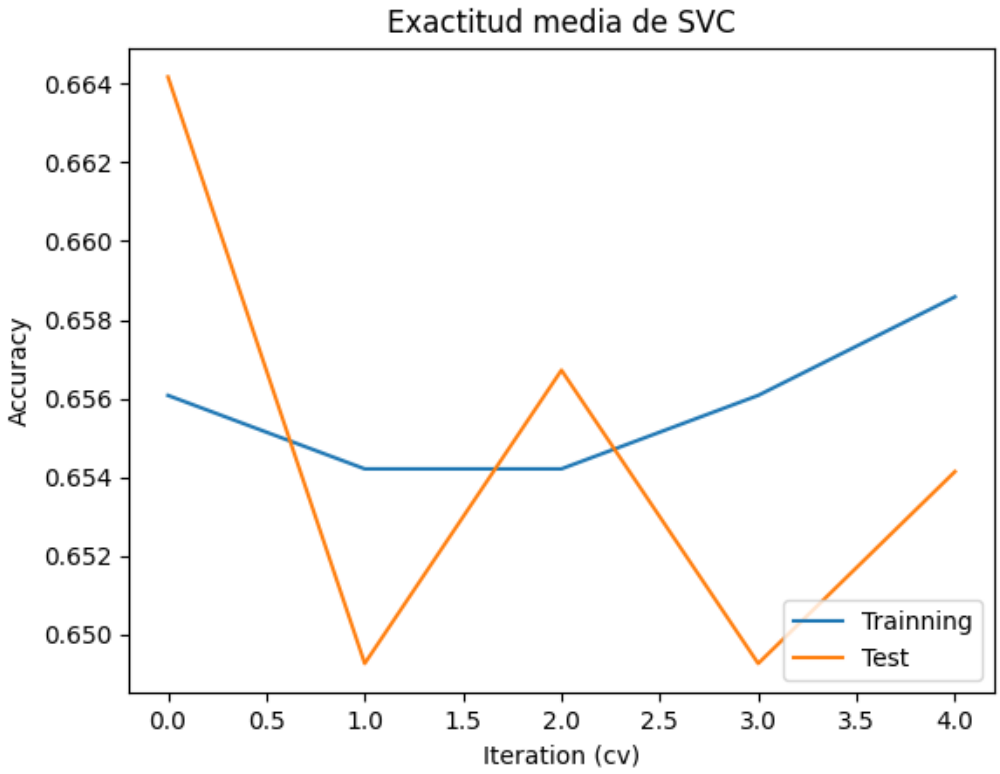
```
Classification report for model DNN:
```

```
              precision    recall  f1-score   support

           0       0.81      0.55      0.65        53
           1       0.82      0.94      0.87       115

    accuracy                           0.82       168
   macro avg       0.81      0.74      0.76       168
weighted avg       0.81      0.82      0.80       168
```

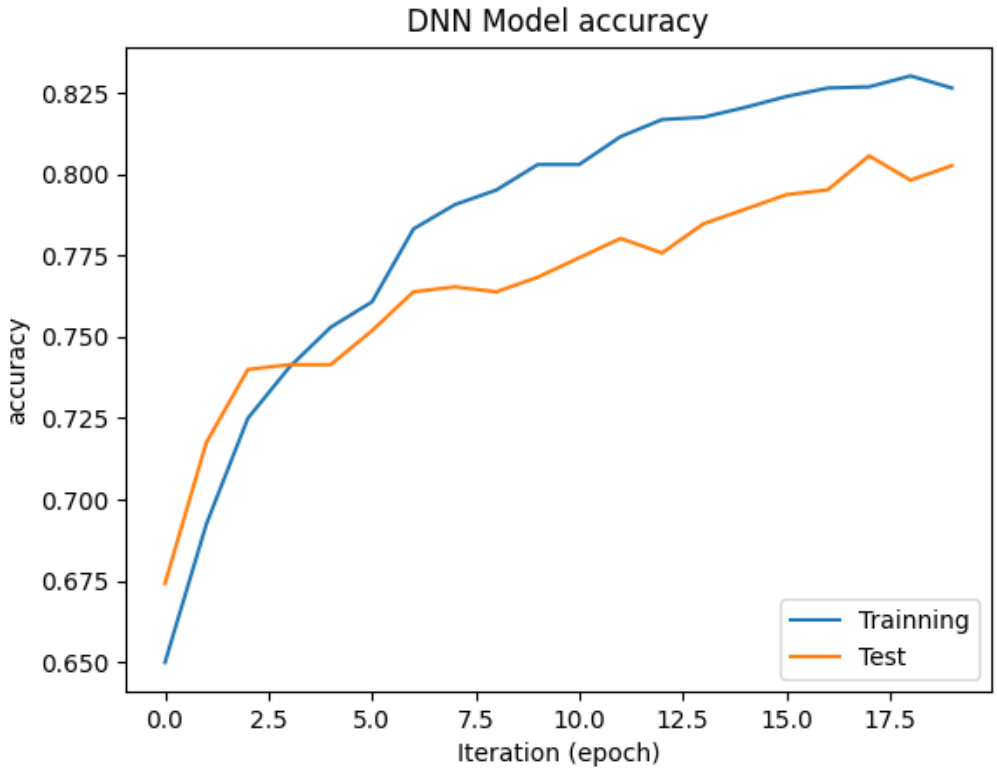
Exactitud media: Compara la exactitud de cada modelo en función de sus hiperparámetros

---





Curva de validación: Compara los resultados del modelo en función de sus hiperparámetros



Hiperparámetros: Muestra el dataframe con los hiperparámetros usados en el entrenamiento

---

Hiperparámetros del modelo KNN

	n_neighbors	weights
hyperparams	92	uniform

Hiperparámetros del modelo SVC

	C	decision_function_shape	gamma	kernel	probability
hyperparams	0.01811	ovo	scale	poly	True

Hiperparámetros de la red neuronal

	neurons	activation
layer 0	16	sigmoid
layer 1	15	relu
layer 2	17	relu
layer 3	2	softmax

	optimizer	lr
compiler	Adam	0.00081