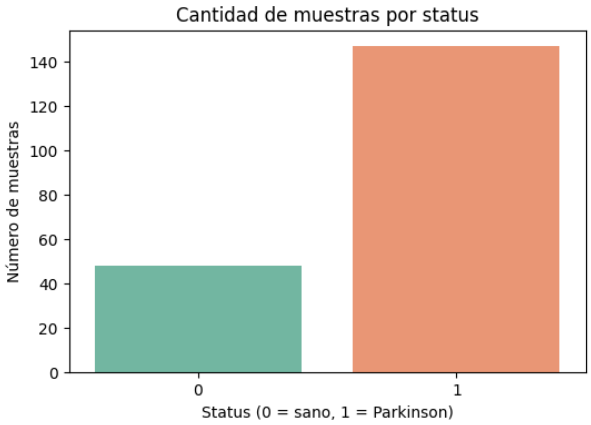
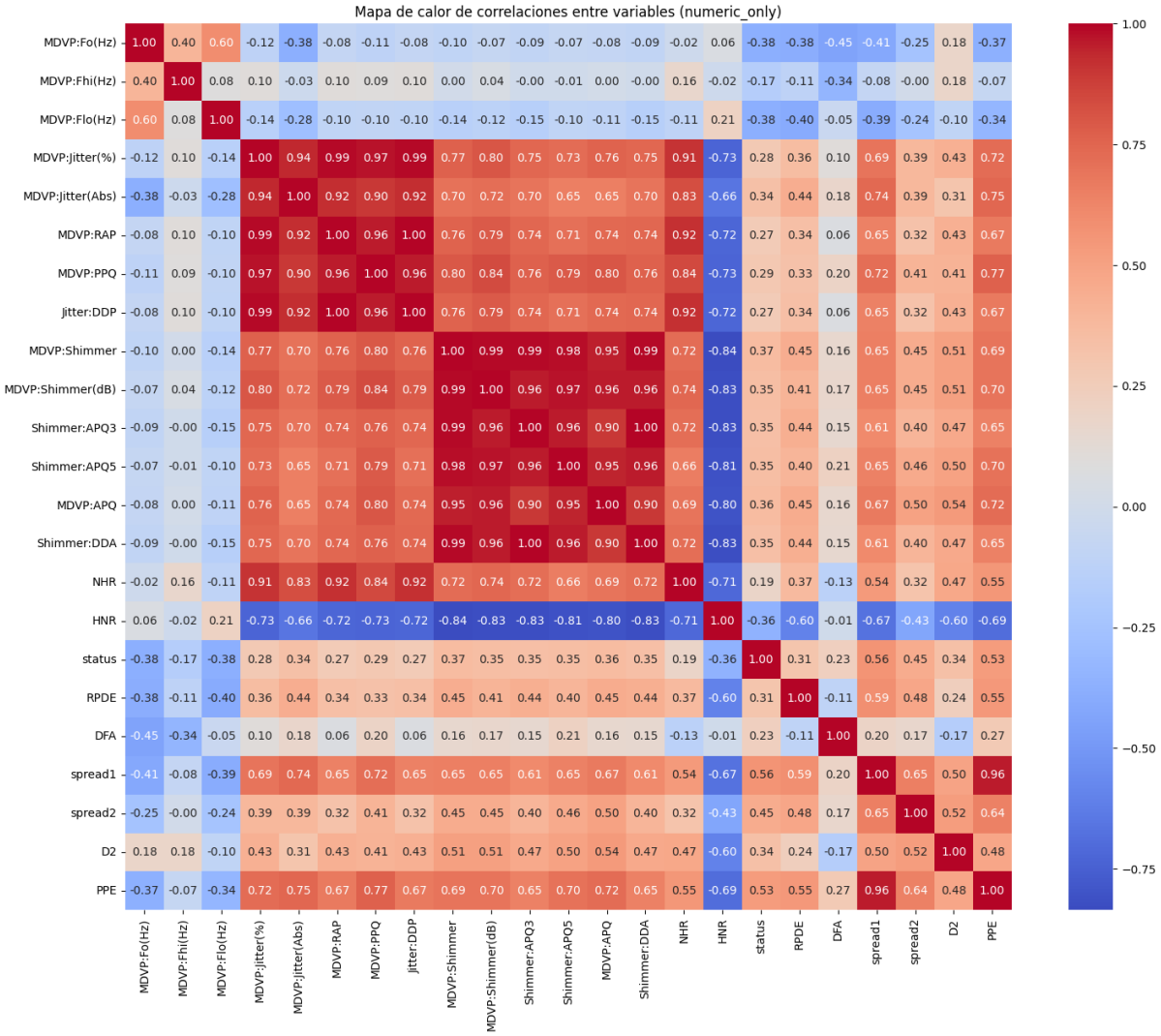
PRÉDICTION DE LA MALADIE DE PARKINSON

Rapport technique et statistique du logiciel

Répartition par statut cible:



Carte de corrélation de Pearson entre les variables :



Validation des hypothèses et choix du test statistique

Il est essentiel de vérifier les hypothèses sous-jacentes aux tests paramétriques classiques. Lorsque ces hypothèses ne sont pas satisfaites, nous privilégions l'utilisation des méthodes non paramétriques, qui sont plus robustes face aux écarts par rapport aux distributions théoriques attendues.

Vérification de la normalité (Kolmogorov-Smirnov)

Pour chaque variable étudiée, nous divisons les observations en deux groupes selon leur statut (0 = sujet sain, 1 = sujet atteint de Parkinson). Ensuite, nous appliquons le test de Kolmogorov-Smirnov indépendamment à chaque sous-groupe.

- **Hypothèse nulle (H_0)** : les données proviennent d'une distribution normale.
- **Décision** :
    - Si la valeur p est supérieure ou égale à 0,05 ($p \geq 0,05$), nous ne rejetons pas H_0 et considérons que l'échantillon suit une distribution suffisamment proche de la normale.
    - Si la valeur p est inférieure à 0,05 ($p < 0,05$), nous concluons à un écart significatif par rapport à la normalité, indiquant que l'utilisation de tests non paramétriques serait alors appropriée.

	Variable	**0**	**Normalidad**	**1**	**Normalidad**
0	MDVP:Fo(Hz)	<0.001	No normal	<0.001	No normal
1	MDVP:Fhi(Hz)	<0.001	No normal	<0.001	No normal
2	MDVP:Flo(Hz)	<0.001	No normal	<0.001	No normal
3	MDVP:Jitter(%)	<0.001	No normal	<0.001	No normal
4	MDVP:Jitter(Abs)	<0.001	No normal	<0.001	No normal
5	MDVP:RAP	<0.001	No normal	<0.001	No normal
6	MDVP:PPQ	<0.001	No normal	<0.001	No normal
7	Jitter:DDP	<0.001	No normal	<0.001	No normal
8	MDVP:Shimmer	<0.001	No normal	<0.001	No normal
9	MDVP:Shimmer(dB)	<0.001	No normal	<0.001	No normal
10	Shimmer:APQ3	<0.001	No normal	<0.001	No normal
11	Shimmer:APQ5	<0.001	No normal	<0.001	No normal
12	MDVP:APQ	0.015	No normal	<0.001	No normal
13	Shimmer:DDA	<0.001	No normal	<0.001	No normal
14	NHR	<0.001	No normal	<0.001	No normal
15	HNR	0.013	No normal	0.002	No normal

16	RPDE	0.630	Normal	<0.001	No normal
17	DFA	<0.001	No normal	0.090	Normal
18	spread1	0.371	Normal	0.009	No normal
19	spread2	0.640	Normal	0.520	Normal
20	D2	0.917	Normal	0.025	No normal
21	PPE	0.030	No normal	<0.001	No normal

2. Homogénéité des variances (Test de Levene)

Même si les deux échantillons suivent une distribution normale, les tests paramétriques tels que l'ANOVA exigent que les variances des deux groupes soient homogènes. Pour vérifier cette hypothèse, nous utilisons le test de Levene:

- **Hypothèse nulle (H_0)** : les variances des deux groupes sont égales.
- **Décision** :
  - Si la valeur p est supérieure ou égale à 0,05 ($p \geq 0,05$), nous considérons que les variances sont homogènes.
  - Si la valeur p est inférieure à 0,05 ($p < 0,05$), nous concluons à une différence significative entre les variances.

3. Choix du test statistique final

Sur la base des résultats des tests de normalité (Kolmogorov-Smirnov) et d'homogénéité des variances (Levene) :

- Si les deux groupes satisfont simultanément les conditions de normalité (Kolmogorov-Smirnov : $p \geq 0,05$) et d'homogénéité des variances (Levene : $p \geq 0,05$), nous utilisons l'ANOVA à un facteur (ANOVA-F).
  Le test F mesure le rapport entre la variabilité inter-groupes et la variabilité intra-groupe : plus ce rapport est élevé, plus la variable est discriminante.
- Dans le cas contraire, nous utilisons le test non paramétrique de Kruskal–Wallis, qui ne suppose ni la normalité ni l'égalité des variances. La valeur p obtenue indique alors si les médianes des deux groupes diffèrent de manière statistiquement significative.

	Variable	**Prueba**	**Estadístico**	**p-valor**
1	spread2	ANOVA–F	50.34	<0.001
2	PPE	Kruskal–Wallis	68.08	<0.001
3	spread1	Kruskal–Wallis	68.08	<0.001
4	MDVP:APQ	Kruskal–Wallis	45.88	<0.001
5	MDVP:Jitter(Abs)	Kruskal–Wallis	36.87	<0.001

6	MDVP:PPQ	Kruskal–Wallis	35.63	<0.001
7	MDVP:Shimmer(dB)	Kruskal–Wallis	35.11	<0.001
8	MDVP:Shimmer	Kruskal–Wallis	34.53	<0.001
9	MDVP:Jitter(%)	Kruskal–Wallis	33.32	<0.001
10	Jitter:DDP	Kruskal–Wallis	33.25	<0.001
11	MDVP:RAP	Kruskal–Wallis	33.13	<0.001
12	NHR	Kruskal–Wallis	32.24	<0.001
13	Shimmer:APQ5	Kruskal–Wallis	31.47	<0.001
14	Shimmer:APQ3	Kruskal–Wallis	28.05	<0.001
15	Shimmer:DDA	Kruskal–Wallis	28.02	<0.001
16	HNR	Kruskal–Wallis	24.46	<0.001
17	D2	Kruskal–Wallis	21.85	<0.001
18	RPDE	Kruskal–Wallis	18.55	<0.001
19	MDVP:Fo(Hz)	Kruskal–Wallis	17.40	<0.001
20	MDVP:Flo(Hz)	Kruskal–Wallis	16.81	<0.001
21	MDVP:Fhi(Hz)	Kruskal–Wallis	13.21	<0.001
22	DFA	Kruskal–Wallis	9.69	0.002

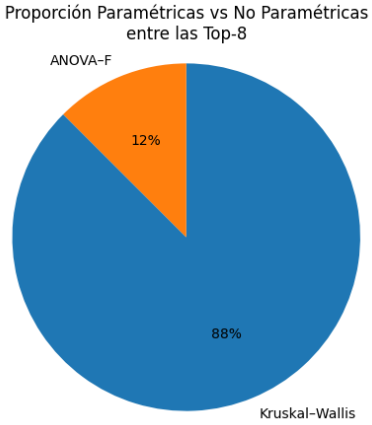
Interprétation des statistiques discriminantes

Pour les variables **paramétriques** (celles qui respectent les hypothèses de normalité selon le test de Kolmogorov–Smirnov et d'homogénéité des variances selon Levene), nous utilisons la statistique **F** issue de l'ANOVA :

- **Mathématiquement :**
  F=Variabilite´ inter-groupesVariabilite´ intra-groupeF = \frac{\text{Variabilité inter-groupes}}{\text{Variabilité intra-groupe}}F=Variabilite´ intra-groupeVariabilite´ inter-groupes
- **Interprétation :**
  Plus la valeur de **F** est élevée, plus grande est la différence relative entre les moyennes des sujets sains et ceux atteints de Parkinson. Ainsi, une valeur élevée du F indique un fort pouvoir discriminant de la variable analysée.

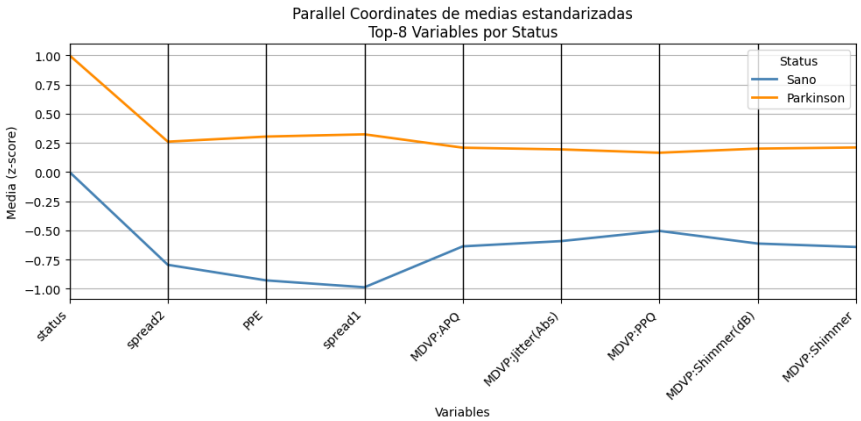
Pour les variables **non paramétriques** (celles qui ne respectent pas les conditions nécessaires à l'utilisation de tests paramétriques), nous utilisons la statistique **H** issue du test de Kruskal–Wallis :

- **Mathématiquement:**
  La statistique **H** mesure la dispersion des rangs des observations entre les deux groupes.
- **Interprétation:**
  Une valeur élevée de **H** indique une différence marquée entre les distributions des deux groupes analysés, témoignant ainsi d'un pouvoir discriminant important de la variable concernée.



Après avoir établi le classement (ranking) des variables selon leur pouvoir discriminant, nous avons retenu les huit caractéristiques les plus discriminantes. Nous les avons ensuite standardisées en scores-Z afin d'uniformiser leurs échelles respectives, puis calculé leur moyenne par groupe (sujets sains vs. sujets atteints de Parkinson).

Ensuite, nous avons tracé un graphique en coordonnées parallèles, où la ligne représentant les sujets atteints de Parkinson (en orange) se situe systématiquement au-dessus de celle des sujets sains (en bleu) à travers ces huit variables. Ce graphique montre notamment, au niveau des variables **spread2** et **spread1**, la séparation la plus nette entre les deux groupes.



Entraînement du modèle

Nous avons divisé les données selon une approche « hold-out » : 80 % des données ont servi à l'entraînement, tandis que 20 % ont été réservées pour le test.

Tableau de validation croisée (CV)

Le tableau présente, pour chaque modèle, le temps moyen d'entraînement (**TrainTime**) ainsi que les mesures de performance (**Accuracy**, **Precision**, **Recall**, **F1** et **MCC**) moyennées selon une validation croisée stratifiée à 5 plis (« 5-fold Stratified Cross Validation).
Les modèles sont classés par ordre décroissant selon le critère MCC afin de mettre en évidence celui qui offre la meilleure capacité de classification équilibrée entre les deux classes.

Modelo	**Tiempo (s)**	**Accuracy**	**Precision**	**Recall**	**F1**	**MCC**
Random Forest	0.58	0.85	0.89	0.92	0.9	0.59
Naïve Bayes	0.01	0.76	0.97	0.71	0.82	0.54
XGBoost	0.12	0.83	0.89	0.88	0.88	0.54
k-NN	0.01	0.81	0.87	0.88	0.87	0.47
SVM (RBF)	0.01	0.82	0.85	0.93	0.89	0.46
Logistic L2	0.02	0.81	0.85	0.91	0.88	0.44

Tableau des caractéristiques du système :

Décrit l'environnement dans lequel les expériences ont été exécutées : système d'exploitation, processeur et mémoire RAM. Cela fournit un contexte relatif aux temps d'entraînement et permet la reproductibilité des résultats.

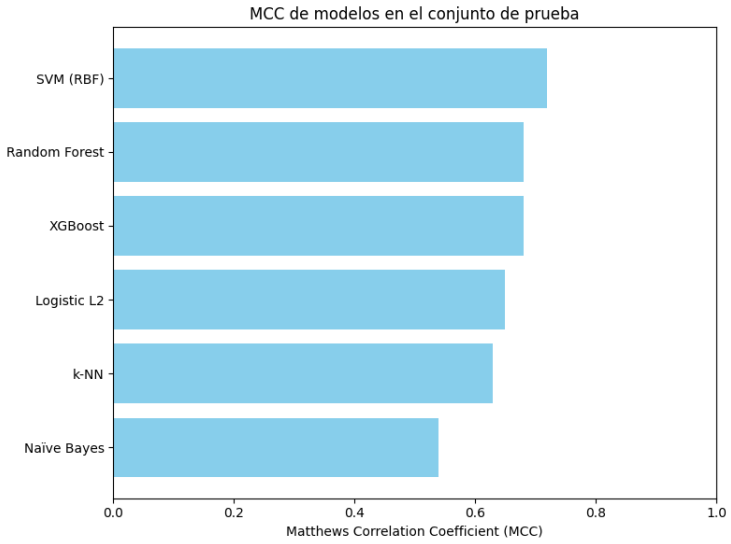
Recurso	**Especificación**
OS	Linux 6.1.123+
CPU	x86_64
RAM	12.67 GB

Tests :

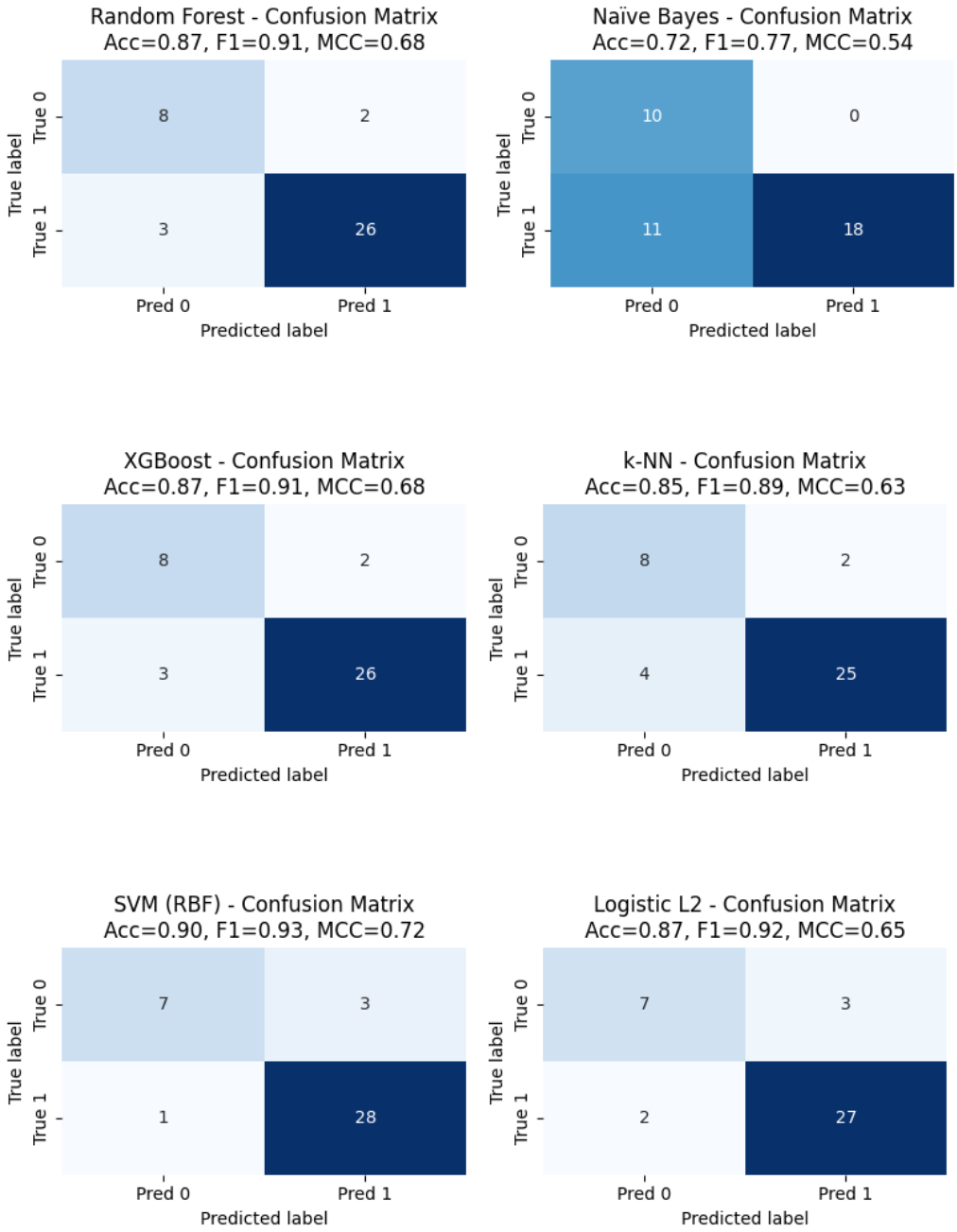
Lors de la dernière phase, chaque pipeline complet a été entraîné avec 80 % des données, puis sérialisé sur disque afin d'assurer la reproductibilité. Nous avons ensuite évalué ses performances sur les 20 % restants (« ensemble de test »). Pour chaque modèle, nous avons calculé les métriques suivantes : Accuracy, Precision, Recall, F1 et MCC, puis tracé sa matrice de confusion sous forme d'une carte de chaleur annotée avec les valeurs de vrais/faux positifs/négatifs. Enfin, nous avons rassemblé l'ensemble de ces métriques dans un tableau ordonné selon la valeur du MCC, reflétant l'équilibre global entre les classes.

Modelo	Nombre archivo	Accuracy	Precision	Recall	F1	MCC
SVM (RBF)	SVM_(RBF).h	0.9	0.9	0.97	0.93	0.72
Random Forest	Random_Forest.h	0.87	0.93	0.9	0.91	0.68
XGBoost	XGBoost.h	0.87	0.93	0.9	0.91	0.68
Logistic L2	Logistic_L2.h	0.87	0.9	0.93	0.92	0.65
k-NN	k-NN.h	0.85	0.93	0.86	0.89	0.63
Naïve Bayes	Naïve_Bayes.h	0.72	1.0	0.62	0.77	0.54

MCC (Coefficient de corrélation de Matthews) : métrique de corrélation pour la classification binaire qui prend en compte les TP, TN, FP et FN, et fournit une valeur interprétable unique (entre –1 et +1). Ce critère a été choisi comme principal indicateur, car il équilibre l'efficacité des prédictions sur des jeux de données potentiellement déséquilibrés.



Matrice de confusion de tous les modèles:



Dans la dernière étape, nous avons sélectionné le meilleur modèle selon le MCC obtenu sur l'ensemble de test (dans ce cas, SVM avec noyau RBF, MCC = 0.720), et sauvegardé son pipeline. Pour vérifier le bon fonctionnement, nous avons chargé ce pipeline et réalisé une inférence exemple sur un échantillon du test, obtenant à la fois l'étiquette prédite et la probabilité associée.

Ensuite, afin de confirmer que cette supériorité n'était pas due au hasard, nous avons appliqué le test de McNemar entre le meilleur modèle et chacun des autres modèles. Nous avons construit des tables de contingence à partir des prédictions réalisées sur le même ensemble de test, calculé les valeurs p et déterminé si les différences dans le nombre de réussites/échecs étaient statistiquement significatives ($\alpha = 0.05$).

Test de McNemar (meilleur vs. autres) :

- SVM (RBF) vs Random Forest : p-value = 0.250 → différence significative : non
- SVM (RBF) vs XGBoost : p-value = 0.375 → différence significative : non
- SVM (RBF) vs Logistic L2 : p-value = 1.000 → différence significative : non
- SVM (RBF) vs k-NN : p-value = 0.219 → différence significative : non
- SVM (RBF) vs Naïve Bayes : p-value = 0.000 → différence significative : oui

Optimisation des hyperparamètres

Nous avons utilisé une procédure GridSearchCV avec une validation croisée stratifiée à 5 plis (identique à celle utilisée lors de la validation précédente) pour chacune des métriques clés suivantes : AUC-ROC, Accuracy, Precision, Recall, F1 et MCC.

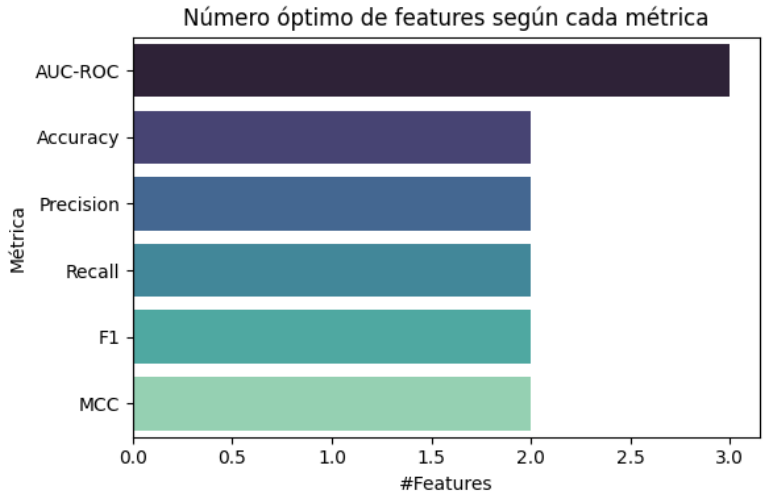
Métricas	**Parámetros**	**CV_score**
AUC-ROC	{'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'linear'}	0.90
Accuracy	{'clf__C': 100, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.84
Precision	{'clf__C': 100, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.88
Recall	{'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'rbf'}	1.0
F1	{'clf__C': 1, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.89
MCC	{'clf__C': 100, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.56

Optimización por MCC y Selección de Variables con RFECV

Para refinar nuestro SVM, realizamos dos pasos encadenados:

1. **GridSearchCV** optimizando la métrica MCC (Matthews Correlation Coefficient), que equilibra verdaderos/falsos positivos y negativos.
2. **RFECV** con SVM lineal, para determinar el número óptimo de variables según cada métrica (AUC-ROC, Accuracy, Precision, Recall, F1 y MCC).

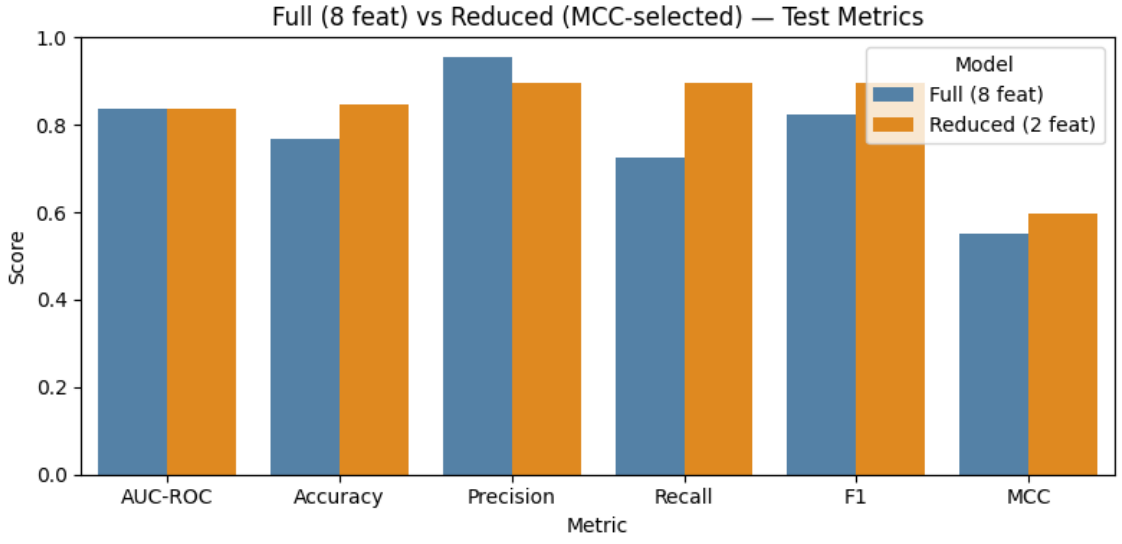
Métrica	**#Features**	**CV_Score**	**Test_Score**	**Features**
AUC-ROC	3	0.912	0.921	spread1, MDVP:APQ, MDVP:Shimmer
Accuracy	2	0.847	0.846	MDVP:APQ, MDVP:Shimmer
Precision	2	0.876	0.897	MDVP:APQ, MDVP:Shimmer
Recall	2	0.933	0.897	MDVP:APQ, MDVP:Shimmer
F1	2	0.902	0.897	MDVP:APQ, MDVP:Shimmer
MCC	2	0.573	0.597	MDVP:APQ, MDVP:Shimmer



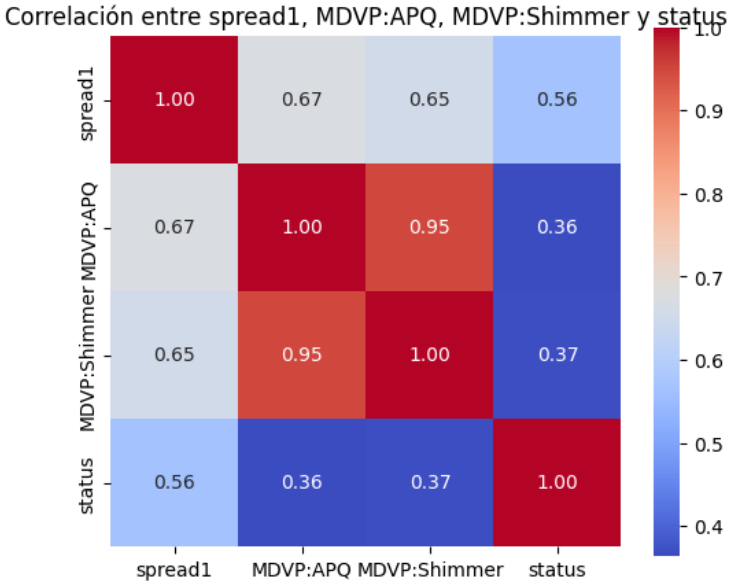
Comparativa de rendimiento en Test: Full (8 features) vs Reduced (2 features)

Modelo	**AUC-ROC**	**Accuracy**	**Precision**	**Recall**	**F1**	**MCC**
Full (8 feat)	0.838	0.769	0.955	0.724	0.824	0.55
Reduced (2 feat)	0.838	0.846	0.897	0.897	0.897	0.597

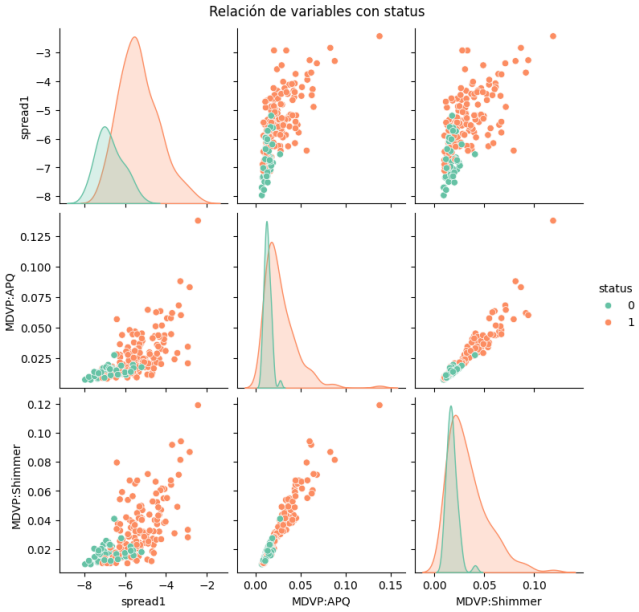
Reducir a solo MDVP:APQ y MDVP:Shimmer no solo simplifica el modelo en un 75 % menos de variables, sino que también **mejora su capacidad de clasificación balanceada**, reflejado en el aumento de MCC. Esto confirma que estas dos características capturan la mayor parte de la información discriminativa sin sacrificar la fiabilidad global del clasificador.



Mapa de correlación de las variables finales por estatus:



Discriminación de variables entre pacientes que tiene vs con lo que no tiene Parkinson:

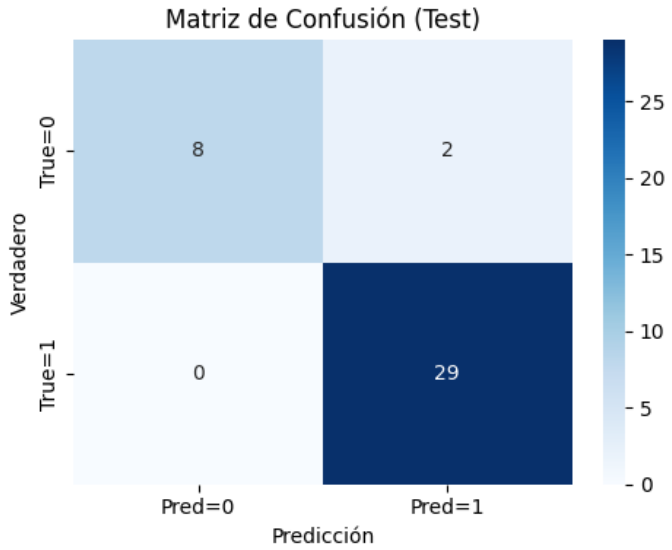


Con el modelo final (SVM-RBF optimizado por MCC) entrenado sobre las **tres variables finales** (spread1, MDVP:APQ y MDVP:Shimmer):

Métrica	**Train**	**Test**
AUC-ROC	0.951	0.921
Accuracy	0.885	0.949
Precision	0.879	0.935

Recall	0.983	1.0
F1	0.928	0.967
MCC	0.669	0.865

Matrix de confusión final:



Selección de modelos híbridos:

Model	**Auc**	Accuracy	Precision	Recall	F1	Mcc
Soft_voting_parkinson	0.917	0.846	0.897	0.897	0.897	0.597
Stacking_parkinson	0.914	0.821	0.867	0.897	0.881	0.515