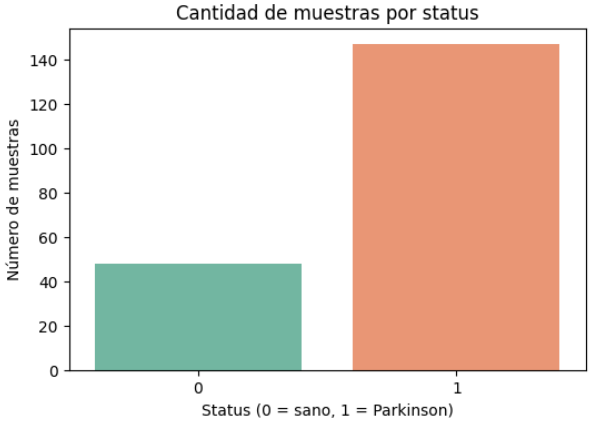
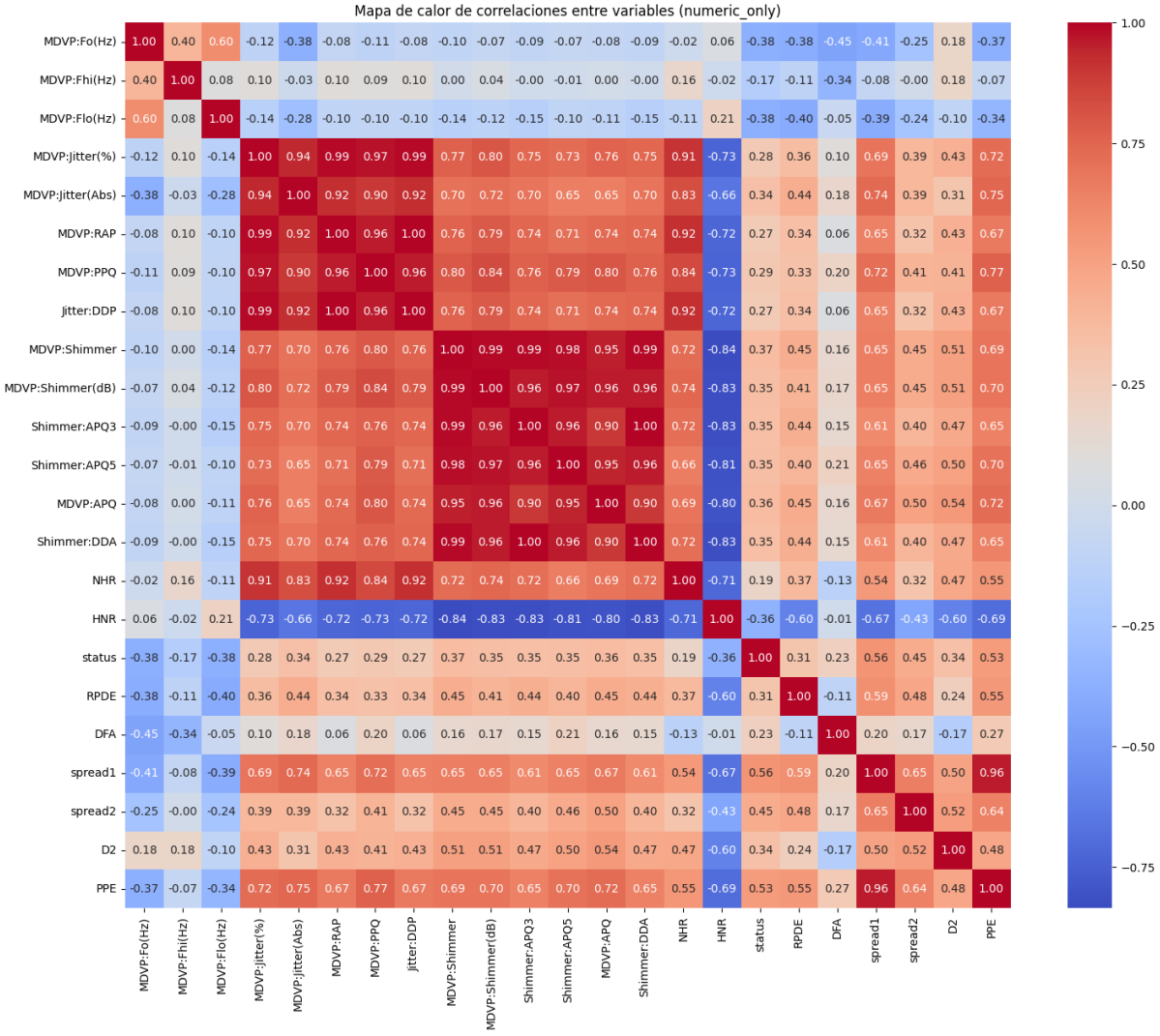
帕金森病预测

软件技术与统计报告

目标状态分布：



变量之间的Pearson相关性图：



假设检验与统计方法选择

验证经典参数统计方法的前提假设至关重要。当这些假设不满足时，我们倾向于选择对分布偏离更稳健的非参数方法。

正态性检验（Kolmogorov-Smirnov 检验）：
对于每个变量，我们根据样本的状态（0 = 健康组，1 = **帕金森病**组）将观察值分成两组，然后对每个子组分别进行Kolmogorov-Smirnov正态性检验。

- **零假设（H_0）**：数据来自正态分布。
- **判定标准：**
  - 若p值大于或等于0.05（$p \geq 0.05$），我们不拒绝零假设，认为样本近似服从正态分布；
  - 若p值小于0.05（$p < 0.05$），我们认为数据明显偏离正态分布，应使用非参数方法。

	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

方差齐性检验（Levene检验）：

即使**两个**样本都服从正态分布，类似ANOVA的参数检验也要求各组之间的方差是齐性的。我们使用Levene检验验证该假设：

- **零假设（H_0）**：各组方差相等。
- **判定标准：**
  - 若p值大于或等于0.05（$p \geq 0.05$），则认为方差齐性；
  - 若p值小于0.05（$p < 0.05$），则认为方差存在显著差异。

最终统计方法选择

根据正态性（Kolmogorov-Smirnov）和方差齐性（Levene）检验的结果：

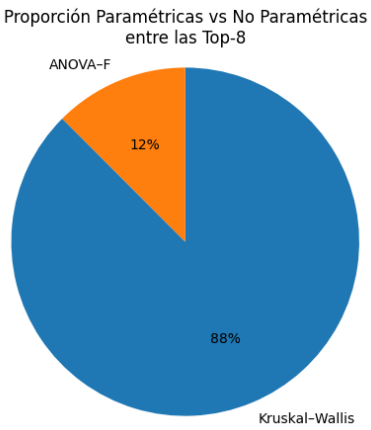
- 如果**两**组数据同时满足正态性（Kolmogorov-Smirnov：$p \geq 0.05$）和方差齐性（Levene：$p \geq 0.05$）条件，我们使用单因素方差分析（ANOVA-F检验）。
  F检验测量组间变异与组内变异的比值，该值越高，变量的区分能力越强。
- 如果上述条件不同时满足，我们则采用非参数的Kruskal–Wallis检验，它不要求正态性和方差齐性。此时的p值表明**两**组数据中位数差异的显著程度。

	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

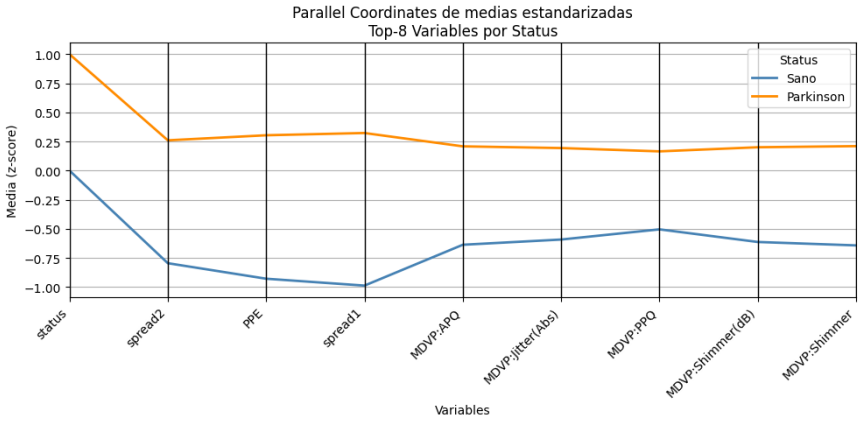
11	MDVP:RAP	Kruskal–Wallis	33.13	<0.001
12	NHR	Kruskal–Wallis	32.24	<0.001
13	Shimmer:APQ5	Kruskal–Wallis	31.47	<0.001
14	Shimmer:APQ3	Kruskal–Wallis	28.05	<0.001
15	Shimmer:DDA	Kruskal–Wallis	28.02	<0.001
16	HNR	Kruskal–Wallis	24.46	<0.001
17	D2	Kruskal–Wallis	21.85	<0.001
18	RPDE	Kruskal–Wallis	18.55	<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

判别统计量解释

- 对于参数变量（同时满足Kolmogorov-Smirnov正态性检验和Levene方差齐性检验），使用ANOVA的F统计量：
  - 数学表**达式**：F = 组间变异 / 组内变**异**
  - 解释：F值越高，说明健康组和帕金森组均值的相对差异越大，该变量的区分能力越强。
- 对于非参数变量（未满足参数检验条件），我们使用Kruskal–Wallis的H统计量：
  - 数学意义：H统计量衡量两组观测数据秩次分布的差**异程度**；
  - 解释：H值越高，表示**两**组分布差异越明显，变量的区分能力越强。



基于变量的判别能力排序后，我们选择了前8个最具判别力的特征，进行Z分数标准化以统一尺度，然后分别计算两组（健康组 vs. 帕金森病组）的平均值。接着，我们绘制了平行坐标图，在图中帕金森病组（橙色）的折线在这8个变量中始终位于健康组（蓝色）的上方，尤其在spread2与spread1变量上，两组的区分最为明显。



模型训练

我们采用「Hold-out」数据分割方法，80%的数据用于模型训练，剩余20%作为测试集。

交叉验证表（CV）

该表展示每个模型的平均训练时间（TrainTime）以及5折分层交叉验证（5-fold Stratified Cross Validation）下的平均性能指标（Accuracy、Precision、Recall、F1和MCC）。
模型根据MCC指标由高到低排序，以强调哪个模型在类别间分类能力的平衡性表现最佳。

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

系统配置表：

详细描述了进行实验的软硬件环境：操作系统、处理器和内存。这提供了训练时间的上下文，并有助于结果的复现。

Recurso	Especificación
OS	Linux 6.1.123+
CPU	x86_64
RAM	12.67 GB

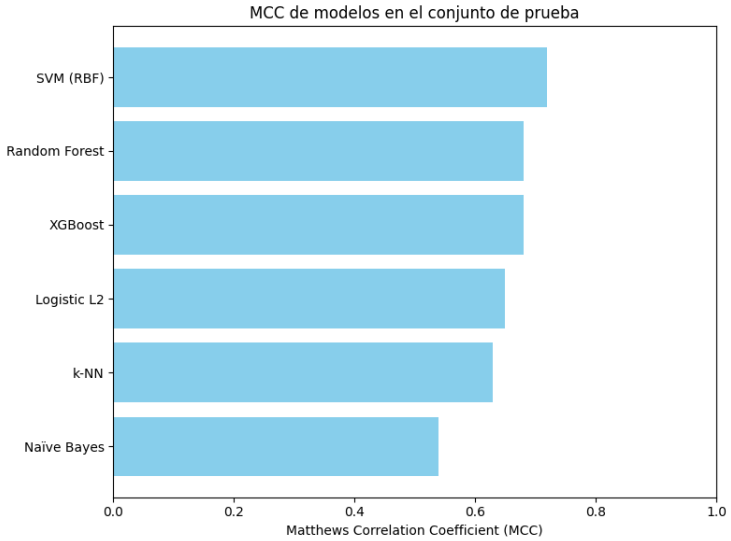
测试阶段：

在最后阶段，每个完整模型管道（pipeline）使用80%的数据进行训练，并序列化保存到磁盘以确保结果可复现。随后，我们在剩余20%的测试集中评估了模型性能。
对于每个模型，我们计算了Accuracy、Precision、Recall、F1和MCC，并绘制了混淆矩阵的热力图（标注真/假**阳性、真**/假**阴性**）。**最后**，**所有指**标汇总到一张表中，并按照MCC排序，以反映整体类别的平衡性能。

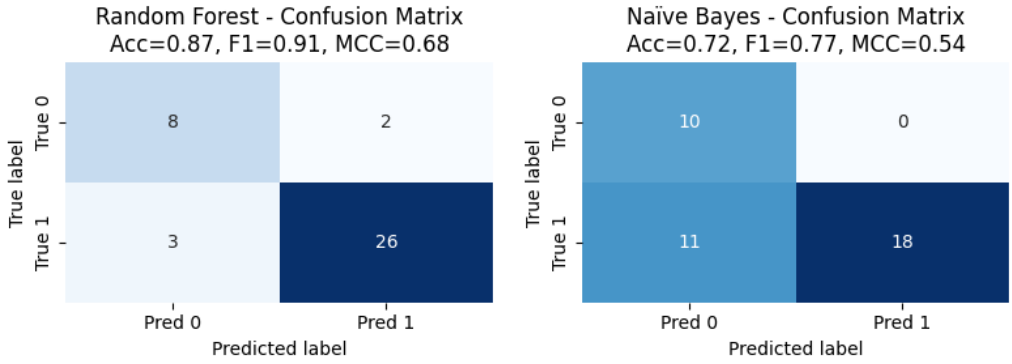
Modelo	Nombre archivo	Accuracy	Precision	Recall	F1	MCC
SVM (RBF)	SVM_(RBF).h	0.9	0.9	0.97	0.93	0.72
Random Forest	Random_Forest.h	0.87	0.93	0.9	0.91	0.68
XGBoost	XGBoost.h	0.87	0.93	0.9	0.91	0.68
Logistic L2	Logistic_L2.h	0.87	0.9	0.93	0.92	0.65
k-NN	k-NN.h	0.85	0.93	0.86	0.89	0.63
Naïve Bayes	Naïve_Bayes.h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

Matthews相**关系数**（MCC）：

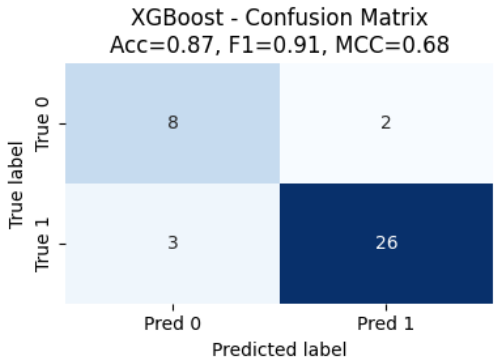
MCC是一**种适用于二分**类问题的相关性指标，综合考虑TP、TN、FP和FN，提供单一的解释性数值（范围-1到+1）。我们选择它作为主要衡量标准，因为它能在类别可能不平衡的数据集中有效衡量模型的预测能力。

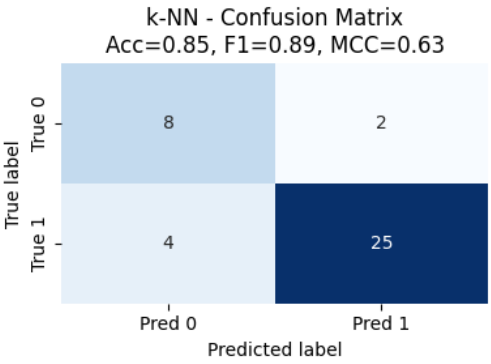

MCC de modelos en el conjunto de prueba
SVM (RBF)
Random Forest
XGBoost
Logistic L2
k-NN
Naïve Bayes
0.0
0.2
0.4
0.6
0.8
1.0
Matthews Correlation Coefficient (MCC)

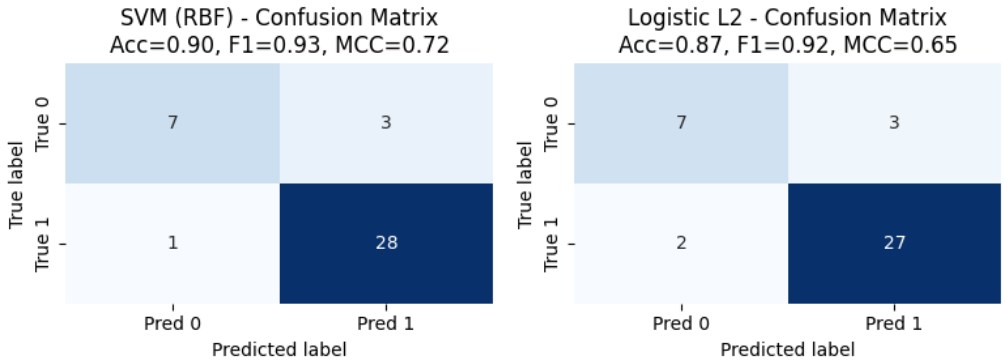
所有模型的混淆矩阵：


Random Forest - Confusion Matrix
Acc=0.87, F1=0.91, MCC=0.68
True label
True 0
True 1
8
2
3
26
Pred 0
Pred 1
Predicted label

Naïve Bayes - Confusion Matrix
Acc=0.72, F1=0.77, MCC=0.54
True label
True 0
True 1
10
0
11
18
Pred 0
Pred 1
Predicted label


XGBoost - Confusion Matrix
Acc=0.87, F1=0.91, MCC=0.68
True label
True 0
True 1
8
2
3
26
Pred 0
Pred 1
Predicted label


k-NN - Confusion Matrix
Acc=0.85, F1=0.89, MCC=0.63
True label
True 0
True 1
8
2
4
25
Pred 0
Pred 1
Predicted label



在最后阶段，我们根据测试集上MCC的表现选择最佳模型（本研究中为SVM带RBF核，MCC=0.720），并保存了该模型管道。为了验证该模型的有效性，我们重新载入模型并在测试样本上进行了预测，得到了预测标签和对应的概率值。随后，为确保该模型的优越性并非偶然，我们使用McNemar检验对最佳模型与其他每个模型进行比较。我们在同一测试集的预测结果上构建列联表，计算p值，以确定准确预测次数之间差异的显著性（显著性水平α=0.05）。

McNemar检验（最佳 vs. 其他模型）：

- SVM (RBF) vs Random Forest：p-value = 0.250 → 显著差异：否
- SVM (RBF) vs XGBoost：p-value = 0.375 → 显著差异：否
- SVM (RBF) vs Logistic L2：p-value = 1.000 → 显著差异：否
- SVM (RBF) vs k-NN：p-value = 0.219 → 显著差异：否
- SVM (RBF) vs Naïve Bayes：p-value = 0.000 → 显著差异：是

超参数优化：

我们针对每个关键性能指标（AUC-ROC、Accuracy、Precision、Recall、F1和MCC），使用5折分层交叉验证的GridSearchCV方法进行了超参数优化。

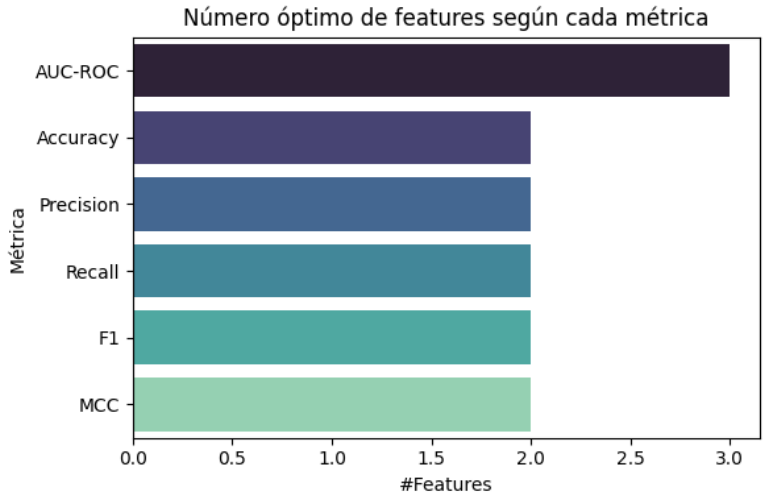
Métricas	**Parámetros**	**CV_score**
AUC-ROC	{'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel': 'linear'}	0.90
Accuracy	{'clf__C': 100, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.84
Precision	{'clf__C': 100, 'clf__gamma': 1, 'clf__kernel': 'rbf'}	0.88
Recall	{'clf__C': 0.1, 'clf__gamma': 'scale', 'clf__kernel':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

基于MCC的优化与RFECV特征选择：

为了进一步优化SVM模型，我们连续进行了两个步骤：

1. 使用GridSearchCV方法优化MCC指标，以平衡真/假**阳性和阴性**；
2. 使用线性SVM进行递归特征消除与交叉验证（RFECV），为每个指标（AUC-ROC、Accuracy、Precision、Recall、F1、MCC）**确定最**优的特征数量。

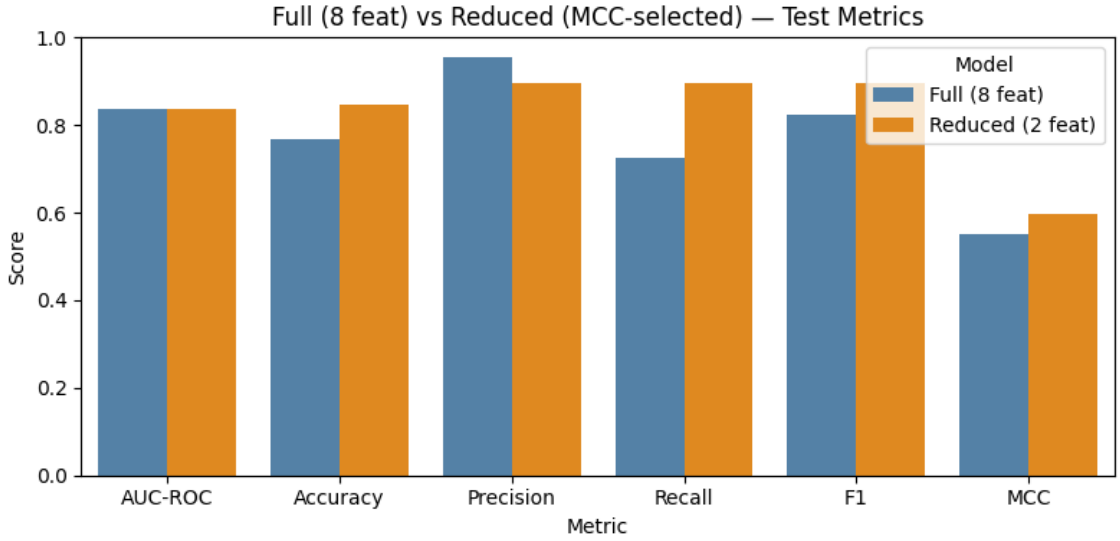
Métrica	**#Features**	**CV_Score**	**Test_Score**	**Features**
AUC-ROC	3	0.912	0.921	spread1, MDVP:APQ, MDVP:Shimmer
Accuracy	2	0.847	0.846	MDVP:APQ, MDVP:Shimmer
Precision	2	0.876	0.897	MDVP:APQ, MDVP:Shimmer
Recall	2	0.933	0.897	MDVP:APQ, MDVP:Shimmer
F1	2	0.902	0.897	MDVP:APQ, MDVP:Shimmer
MCC	2	0.573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



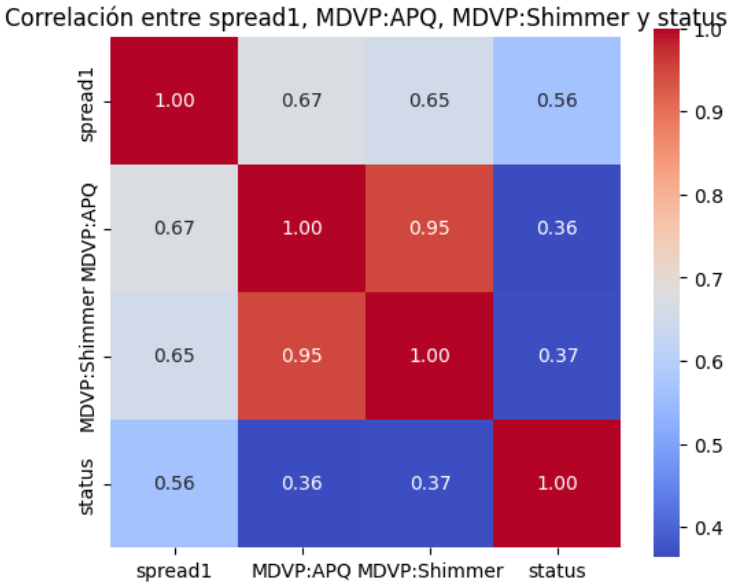
Comparativa de rendimiento en Test: Full (8 features) vs Reduced (2 features)

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

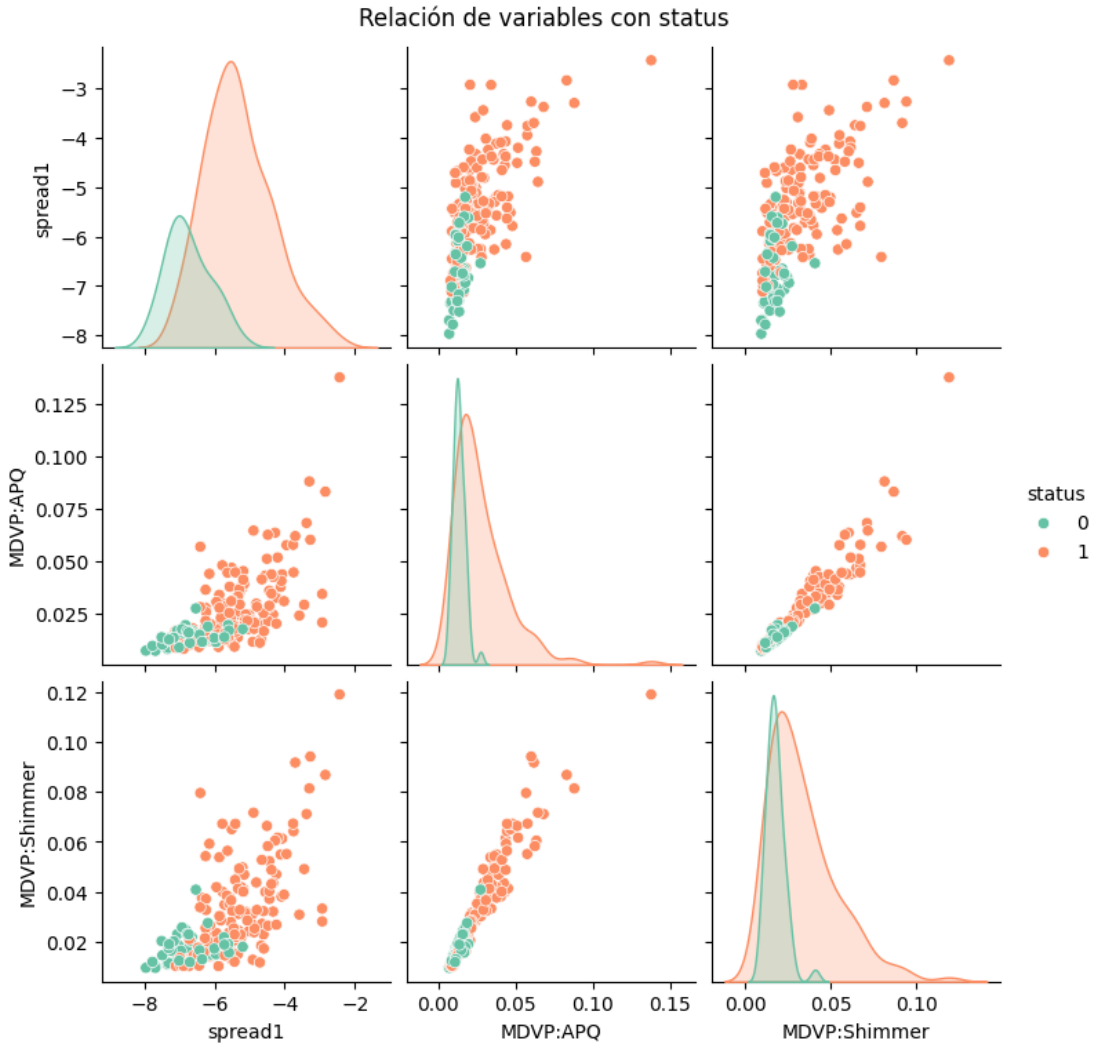
将特征减少至仅包含MDVP:APQ和MDVP:Shimmer，不仅将模型变量减少了75%，同时也提升了模型的类别均衡分类能力（MCC显著提高）。这证实了这两个特征捕获了大部分判别信息，并未损失模型整体的可靠性。



最终变量的状态相关性图：



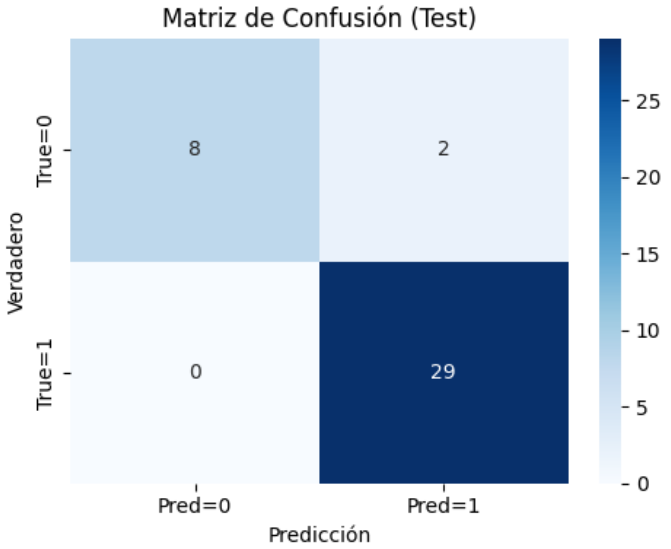
患有**帕金森病与健康患者之**间变量的判别能力：



采用最终优化的SVM-RBF**模型（基于**MCC优化），在最终选定的三个变量（spread1、MDVP:APQ**与**MDVP:Shimmer）**上**进行训练：

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

最终混淆矩阵：



混合模型选择：

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