



# Modelos de regresión

“Dementia patient health and prescriptions dataset.”

# Info del dataset

```
Index: 485 entries, 1 to 998
Data columns (total 24 columns):
#   Column                               Non-Null Count  Dtype
---  -
0   diabetic                             485 non-null    object
1   alcohol_level                        485 non-null    float64
2   heart_rate                           485 non-null    int64
3   blood_oxygen_level                   485 non-null    float64
4   body_temperature                     485 non-null    float64
5   weight                               485 non-null    float64
6   mri_delay                            485 non-null    float64
7   prescription                         485 non-null    object
8   dosage_in_mg                        485 non-null    float64
9   age                                  485 non-null    int64
10  education_level                      485 non-null    object
11  dominant_hand                        485 non-null    object
12  gender                               485 non-null    object
13  family_history                      485 non-null    object
14  smoking_status                      485 non-null    object
15  apoe_4                              485 non-null    object
16  physical_activity                    485 non-null    object
17  depression_status                    485 non-null    object
18  cognitive_test_scores                485 non-null    int64
19  medication_history                   485 non-null    object
20  nutrition_diet                       485 non-null    object
21  sleep_quality                        485 non-null    object
22  chronic_health_conditions            397 non-null    object
23  dementia                             485 non-null    object
dtypes: float64(6), int64(3), object(15)
```

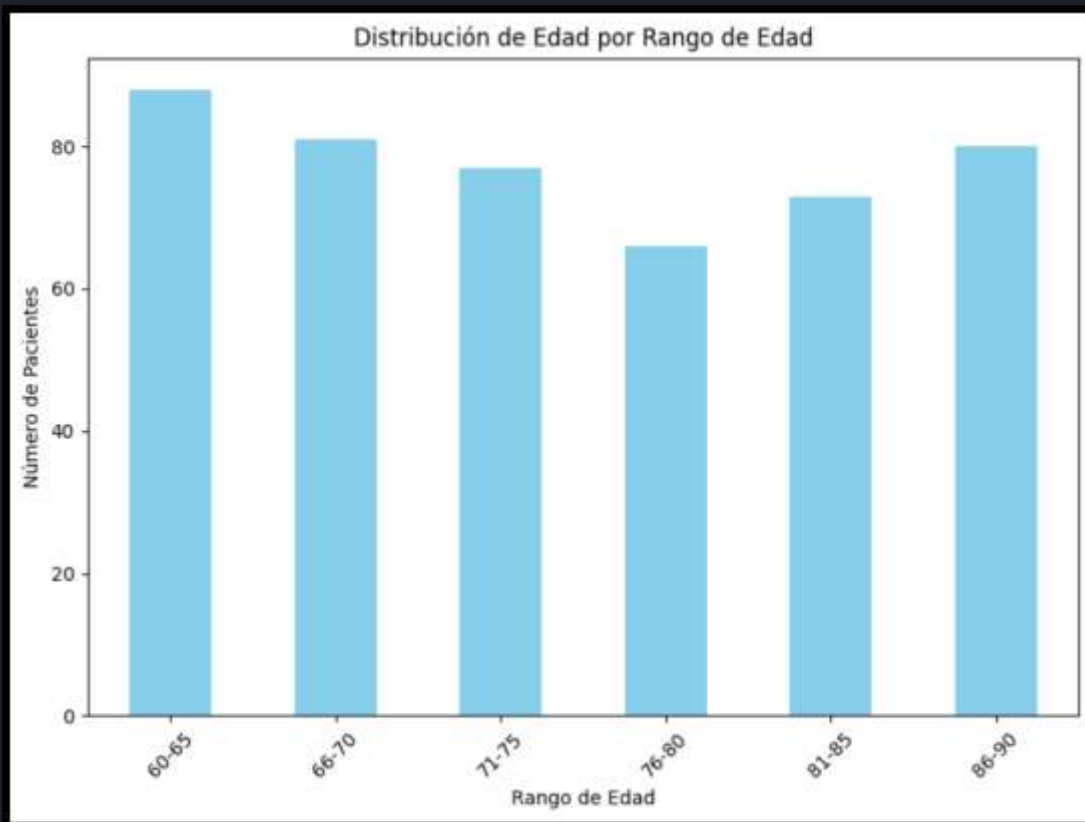
El dataset utilizado para el entrenamiento de los modelos es un subconjunto del dataset original, donde solo se utilizarán aquellas filas con caso de demencia positiva.



# Elección del Atributo para la Regresión

Dado que ninguno de los atributos numéricos presentes en el dataset tiene una relevancia clara y directa que justifique una regresión sobre ellos en el contexto de los datos, hemos decidido suponer que el atributo “age” representa la edad en la cual el paciente fue diagnosticado con demencia.

Esta suposición le otorga un contexto clínico relevante, haciendo que la predicción de esta variable tenga un sentido más lógico dentro del análisis de los datos.



```
count    485.0000
mean      74.3258
std        9.3398
min       60.0000
25%       66.0000
50%       73.0000
75%       82.0000
max       90.0000
Name: age, dtype: float64
```



# Datos de entrenamiento y test

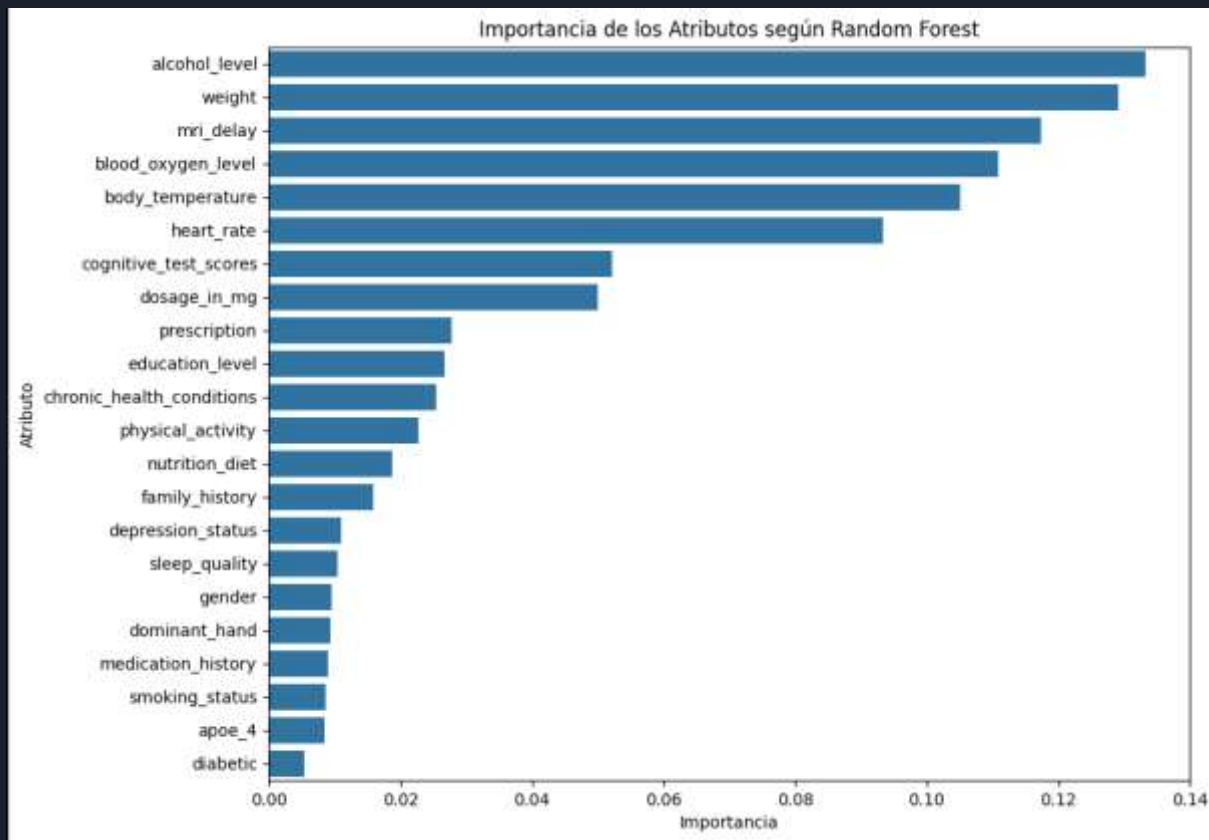
Estos datos fueron divididos, en datos de entrenamiento y test, por una única vez utilizando **train\_test\_split**.

A partir de estos datos de entrenamiento y test, se han generado dos subconjuntos de datos de entrenamiento para posterior comparación:

- Totalidad de los atributos.
- Selección de atributos más importantes en base al modelo RandomForest (feature importance).

**GridSearchCV** fue utilizado en todos los modelos para la obtención de los mejores hiperparámetros.

**neg\_mean\_squared\_error** fue utilizado como scoring para minimizar el impacto de los errores grandes.





# Modelos utilizados

## Modelos Individuales:

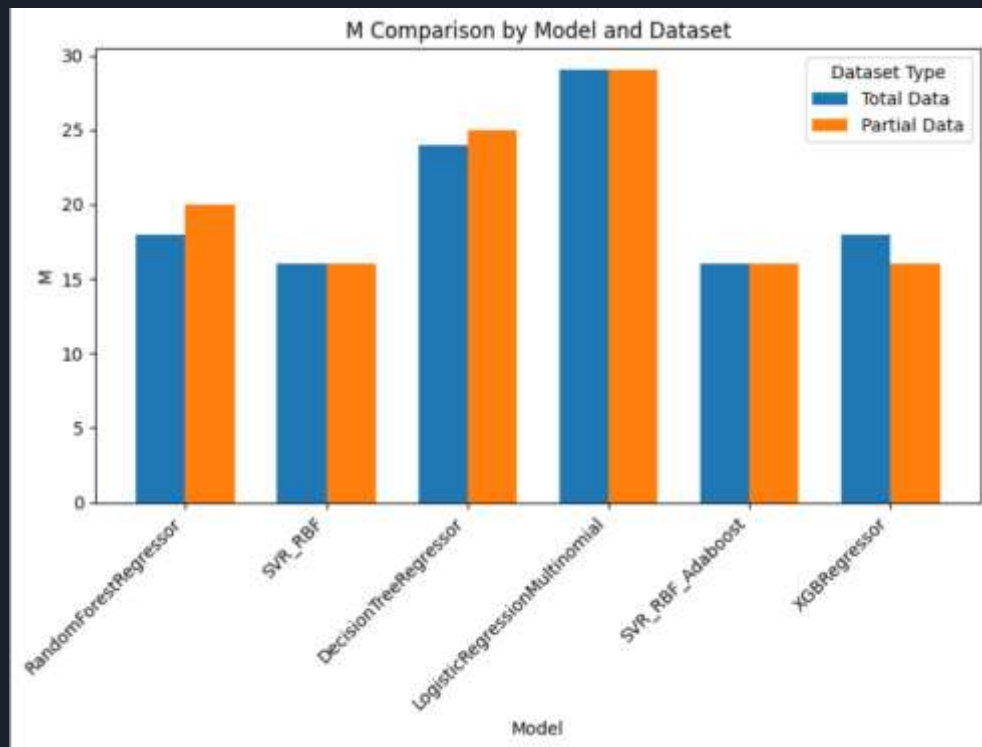
- SVR
- Decision Tree Regressor
- Logistic Regression Multinomial

## Modelos de Ensemble:

- Random Forest Regressor
- SVR + Adaboost
- XGB Regressor

# Comparación de métricas: Error absoluto máximo (M)

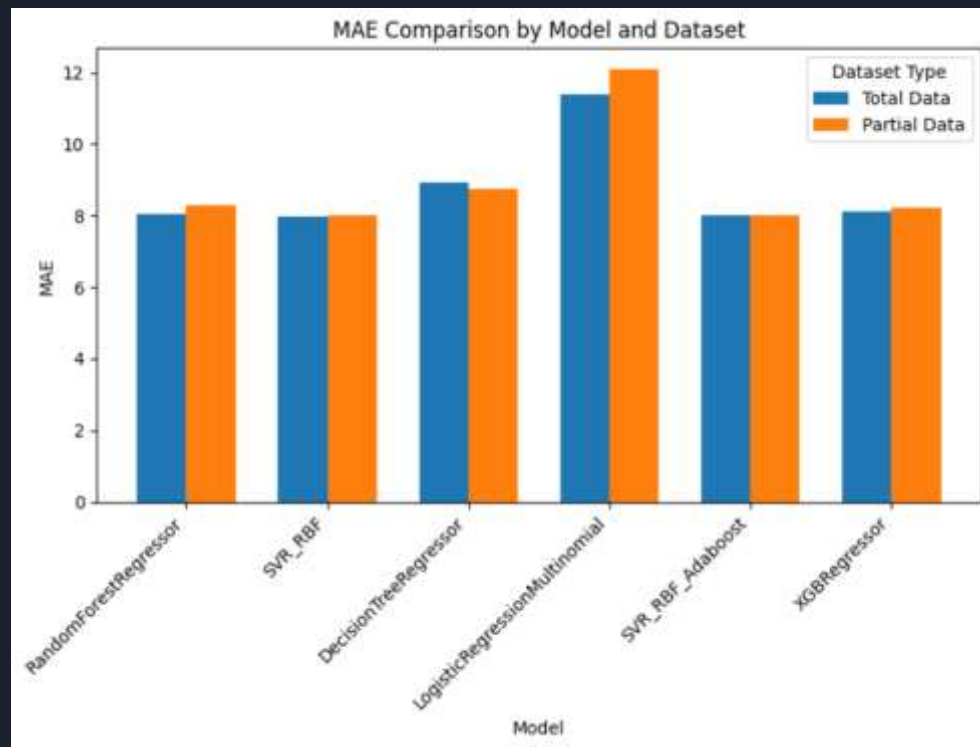
	Model	Dataset	M
	RandomForestRegressor	total_data	18.0000
	RandomForestRegressor	partial_data	20.0000
	SVR_RBF	total_data	16.0000
	SVR_RBF	partial_data	16.0000
	DecisionTreeRegressor	total_data	24.0000
	DecisionTreeRegressor	partial_data	25.0000
	LogisticRegressionMultinomial	total_data	29.0000
	LogisticRegressionMultinomial	partial_data	29.0000
	SVR_RBF_Adaboost	total_data	16.0000
	SVR_RBF_Adaboost	partial_data	16.0000
	XGBRegressor	total_data	18.0000
	XGBRegressor	partial_data	16.0000





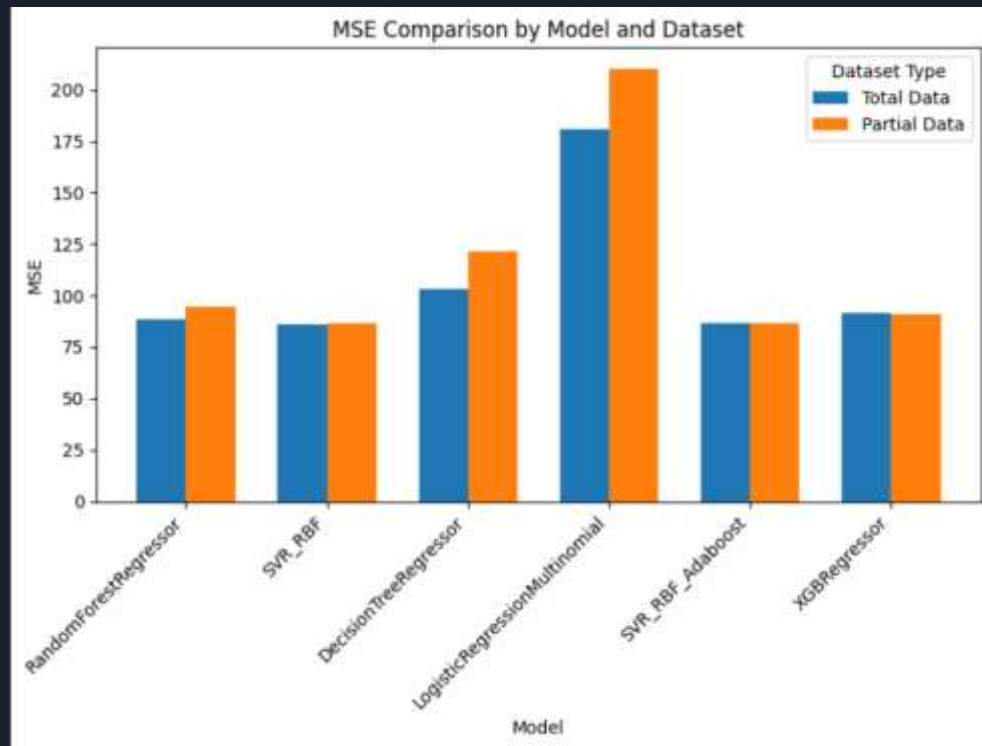
# Comparación de métricas: Error absoluto medio (MAE)

Model	Dataset	MAE
RandomForestRegressor	total_data	8.0412
RandomForestRegressor	partial_data	8.2784
SVR_RBF	total_data	7.9794
SVR_RBF	partial_data	8.0000
DecisionTreeRegressor	total_data	8.9278
DecisionTreeRegressor	partial_data	8.7629
LogisticRegressionMultinomial	total_data	11.3814
LogisticRegressionMultinomial	partial_data	12.0722
SVR_RBF_Adaboost	total_data	8.0103
SVR_RBF_Adaboost	partial_data	8.0000
XGBRegressor	total_data	8.1031
XGBRegressor	partial_data	8.2062



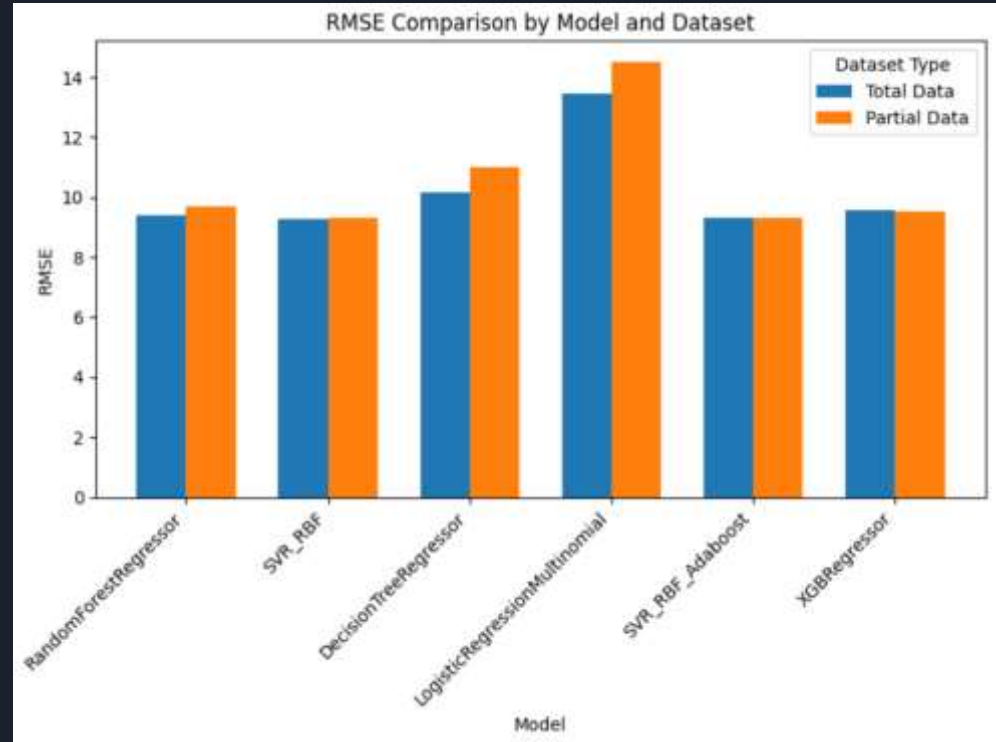
# Comparación de métricas: Error cuadrático medio (MSE)

	Model	Dataset	MSE
	RandomForestRegressor	total_data	88.3093
	RandomForestRegressor	partial_data	94.3814
	SVR_RBF	total_data	86.1649
	SVR_RBF	partial_data	86.4124
	DecisionTreeRegressor	total_data	103.1753
	DecisionTreeRegressor	partial_data	121.4021
	LogisticRegressionMultinomial	total_data	180.6392
	LogisticRegressionMultinomial	partial_data	209.9691
	SVR_RBF_Adaboost	total_data	86.4433
	SVR_RBF_Adaboost	partial_data	86.4124
	XGBRegressor	total_data	91.4845
	XGBRegressor	partial_data	90.6186



# Comparación de métricas: Raíz cuadrada del error cuadrático medio (RMSE)

Model	Dataset	RMSE
RandomForestRegressor	total_data	9.3973
RandomForestRegressor	partial_data	9.7150
SVR_RBF	total_data	9.2825
SVR_RBF	partial_data	9.2958
DecisionTreeRegressor	total_data	10.1575
DecisionTreeRegressor	partial_data	11.0183
LogisticRegressionMultinomial	total_data	13.4402
LogisticRegressionMultinomial	partial_data	14.4903
SVR_RBF_Adaboost	total_data	9.2975
SVR_RBF_Adaboost	partial_data	9.2958
XGBRegressor	total_data	9.5648
XGBRegressor	partial_data	9.5194



# Comparación de métricas

Model	Dataset	M	MAE	MSE	RMSE
RandomForestRegressor	total_data	18.00	8.04	88.31	9.40
RandomForestRegressor	partial_data	20.00	8.28	94.38	9.72
<u>SVR_RBF</u>	total_data	<u>16.00</u>	<u>7.98</u>	<u>86.16</u>	<u>9.28</u>
SVR_RBF	partial_data	<u>16.00</u>	8.00	86.41	9.30
DecisionTreeRegressor	total_data	24.00	8.93	103.18	10.16
DecisionTreeRegressor	partial_data	25.00	8.76	121.40	11.02
LogisticRegressionMultinomial	total_data	29.00	11.38	180.64	13.44
LogisticRegressionMultinomial	partial_data	29.00	12.07	209.97	14.49
SVR_RBF_Adaboost	total_data	<u>16.00</u>	8.01	86.44	9.30
SVR_RBF_Adaboost	partial_data	<u>16.00</u>	8.00	86.41	9.30
XGBRegressor	total_data	18.00	8.10	91.48	9.56
XGBRegressor	partial_data	<u>16.00</u>	8.21	90.62	9.52

# Predicción con instancias de prueba

DecisionTreeRegressor, total data.

Prediccion: 74 68 74

Valor esperado: [74 75 73]

SVR\_RBF, total data.

Prediccion: 74 74 74

Valor esperado: [74 75 73]

LogisticRegressionMultinomial, total data.

Prediccion: 89 75 75

Valor esperado: [74 75 73]

RandomForestRegressor, total data.

Prediccion: 73 73 75

Valor esperado: [74 75 73]

SVR\_RBF\_Adaboost, total data.

Prediccion: 74 74 74

Valor esperado: [74 75 73]

XGBRegressor, total data.

Prediccion: 74 72 73

Valor esperado: [74 75 73]

DecisionTreeRegressor, partial data.

Prediccion: 72 72 72

Valor esperado: [74 75 73]

SVR\_RBF, partial data.

Prediccion: 74 74 74

Valor esperado: [74 75 73]

LogisticRegressionMultinomial, partial data.

Prediccion: 89 75 75

Valor esperado: [74 75 73]

RandomForestRegressor, partial data.

Prediccion: 74 75 73

Valor esperado: [74 75 73]

SVR\_RBF\_Adaboost, partial data.

Prediccion: 74 74 74

Valor esperado: [74 75 73]

XGBRegressor, partial data.

Prediccion: 74 75 73

Valor esperado: [74 75 73]



# Algoritmos de Clasificación

Conjunto de Datos de Salud y Prescripciones de  
Pacientes con Demencia



# Contenido

## Objetivo

Comparar algoritmos y seleccionar el modelo que mejor clasifique, en base a parámetros relacionados con el estilo de vida y resultados médicos, a pacientes con y sin demencia.

## Algoritmos utilizados

Bosques Aleatorios  
Máquinas de Soporte Vectorial (SVM)  
Regresión Logística

## Métricas

Accuracy  
Precision  
Recall  
F1 score  
Curva ROC



## Modelos

Total de Características  
PCA  
Importancia de Características



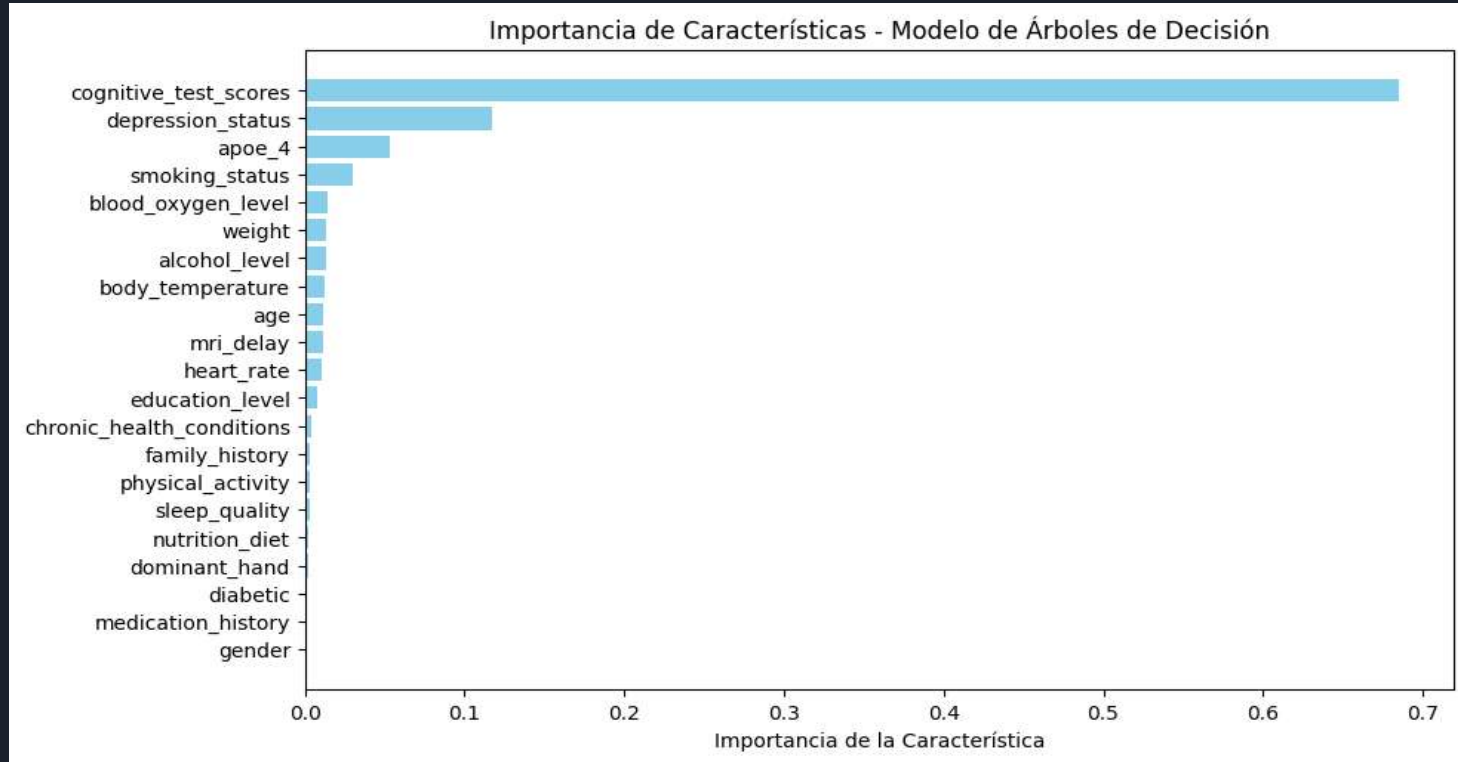
# Variables utilizadas para la clasificación

- Diabético
- Nivel de alcohol
- Ritmo cardíaco
- Nivel de oxígeno en sangre
- Temperatura corporal
- Peso
- Retraso de MRI
- Edad
- Nivel de Educación
- Mano Dominante
- Sexo
- Historial Familiar
- Hábitos de Tabaquismo
- Apoe4
- Actividad Física
- Estado de Depresión
- Puntaje en Test Cognitivo
- Historial de Medicación
- Dieta Nutricional
- Calidad de Sueño
- Condiciones Crónicas de Salud

21 características en total



# Importancia de Características



Puntajes en tests cognitivos - Estado de depresión - Apoe4 - Hábitos de tabaquismo - Nivel de oxígeno en sangre - Peso - Nivel de alcohol - Temperatura corporal - Edad - Retraso de MRI - Ritmo cardíaco - Nivel de educación (12)



# Transformación de Datos

## Codificación de variables categóricas

Se le asignó un valor numérico discreto a cada posible valor de las variables categóricas a utilizar

## Escalado de los datos

Método de **corrección** para evitar resultados de análisis incorrectos

Búsqueda de **efectividad** en los algoritmos que lo requieran (Ej: Máquinas de Soporte Vectorial)



# Justificación de los Algoritmos

## Bosques Aleatorios

- Capturan fácilmente patrones no lineales

## Máquinas de Soporte vectorial

- Efectivo en espacios de alta dimensionalidad
- Adecuado para conjuntos de datos pequeños o medianos
- Permite especificar funciones de kernel para datos no linealmente separables

## Regresión Logística

- Clasificador binario

## k-NN

- Pierde rendimiento con muchas dimensiones

## Regresión Softmax

- Admite múltiples clases

## Árboles de Decisión

- Sobreajuste con muchas dimensiones

## Bayes Ingenuo

- No captura patrones no lineales

# Bosques Aleatorios

Separación de conjuntos  
Escalado de características  
Búsqueda de hiperparámetros con Validación Cruzada

	0	1
0	92	2
1	22	84

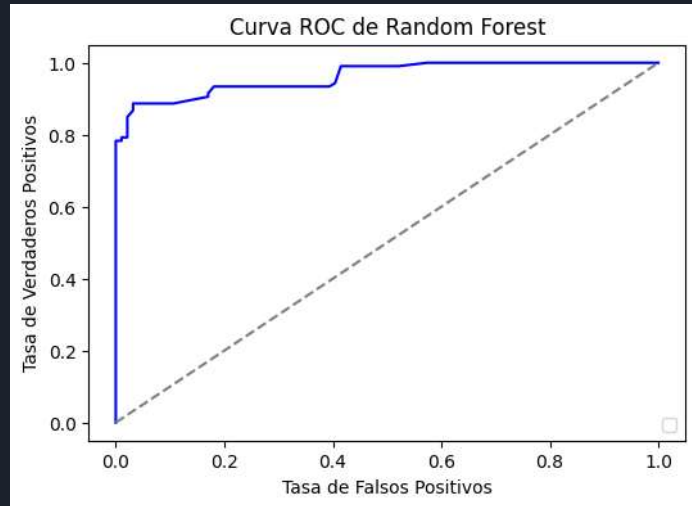
## Reporte de clasificación:

	precision	recall	f1-score	support
0	0.81	0.98	0.88	94
1	0.98	0.79	0.88	106
accuracy			0.88	200
macro avg	0.89	0.89	0.88	200
weighted avg	0.90	0.88	0.88	200

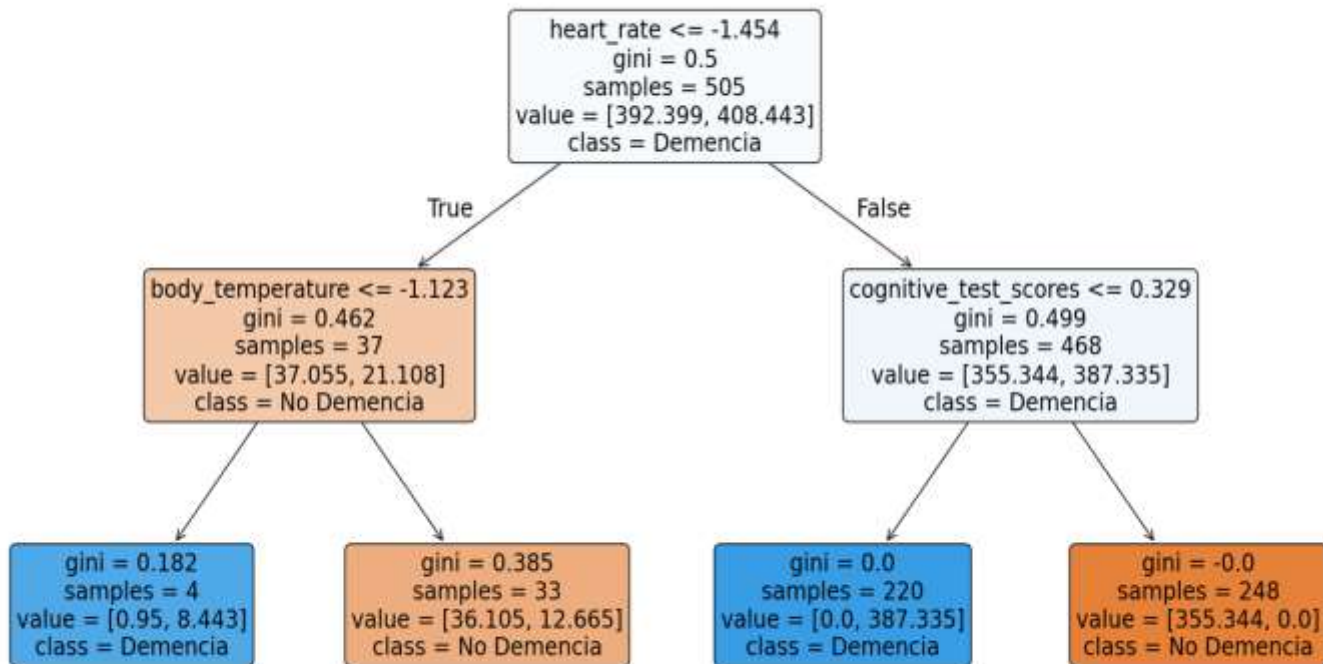
Exactitud del modelo: 0.88

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]



Árbol de Decisión 1 en el Bosque Aleatorio



# Bosques Aleatorios aplicando PCA

Separación de conjuntos

Escalado de características

Aplicación de PCA a las características de los conjuntos

Búsqueda de hiperparámetros con Validación Cruzada

	0	1
0	90	4
1	10	96

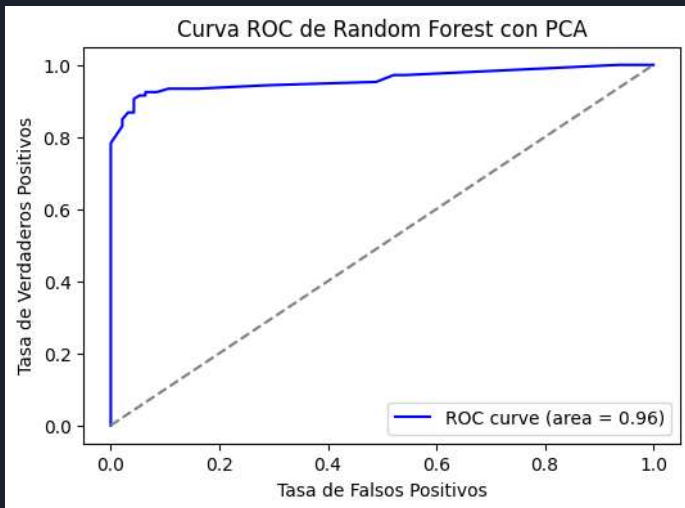
Reporte de clasificación:

	precision	recall	f1-score	support
0	0.90	0.96	0.93	94
1	0.96	0.91	0.93	106
accuracy			0.93	200
macro avg	0.93	0.93	0.93	200
weighted avg	0.93	0.93	0.93	200

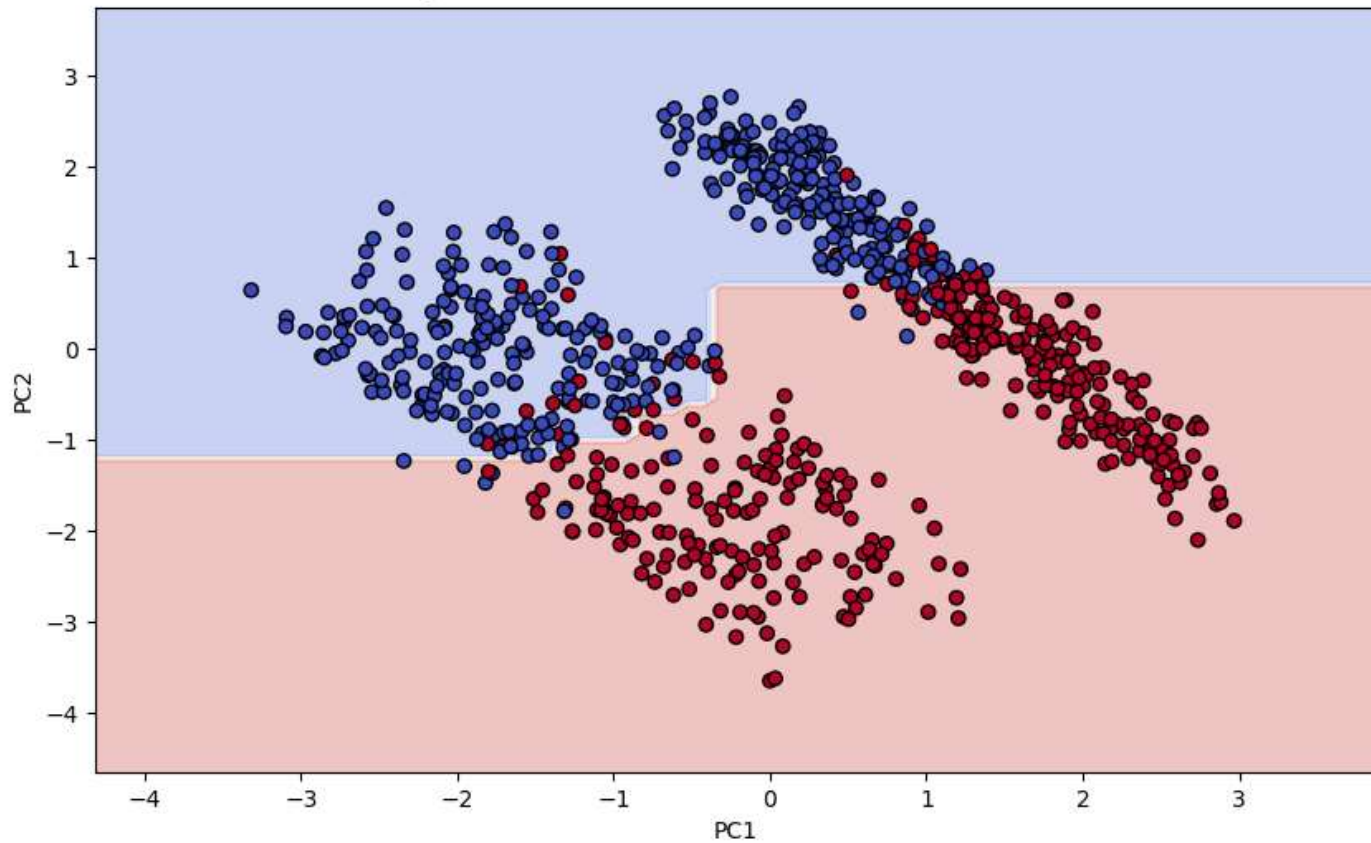
Exactitud del modelo: 0.93

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]



Superficie de Decisión de Random Forest con PCA



Azul: 0  
Rojo: 1

## Bosques Aleatorios aplicando importancia de características

Separación de conjuntos

Escalado de características

Búsqueda de hiperparámetros con Validación Cruzada

	0	1
0	94	0
1	0	106

### Reporte de clasificación:

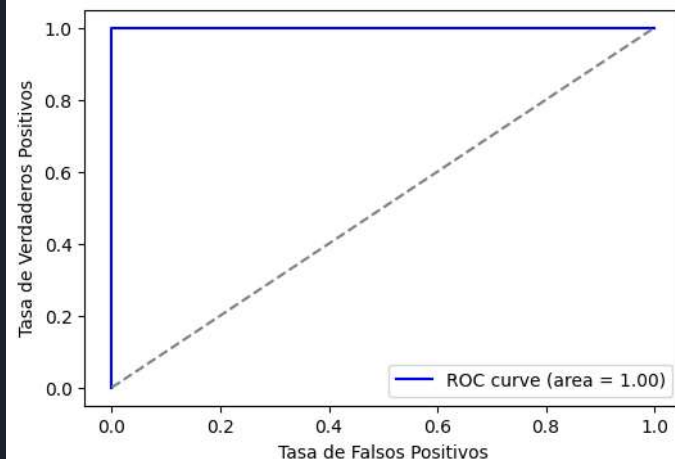
	precision	recall	f1-score	support
0	1.00	1.00	1.00	94
1	1.00	1.00	1.00	106
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Exactitud del modelo: 1.0

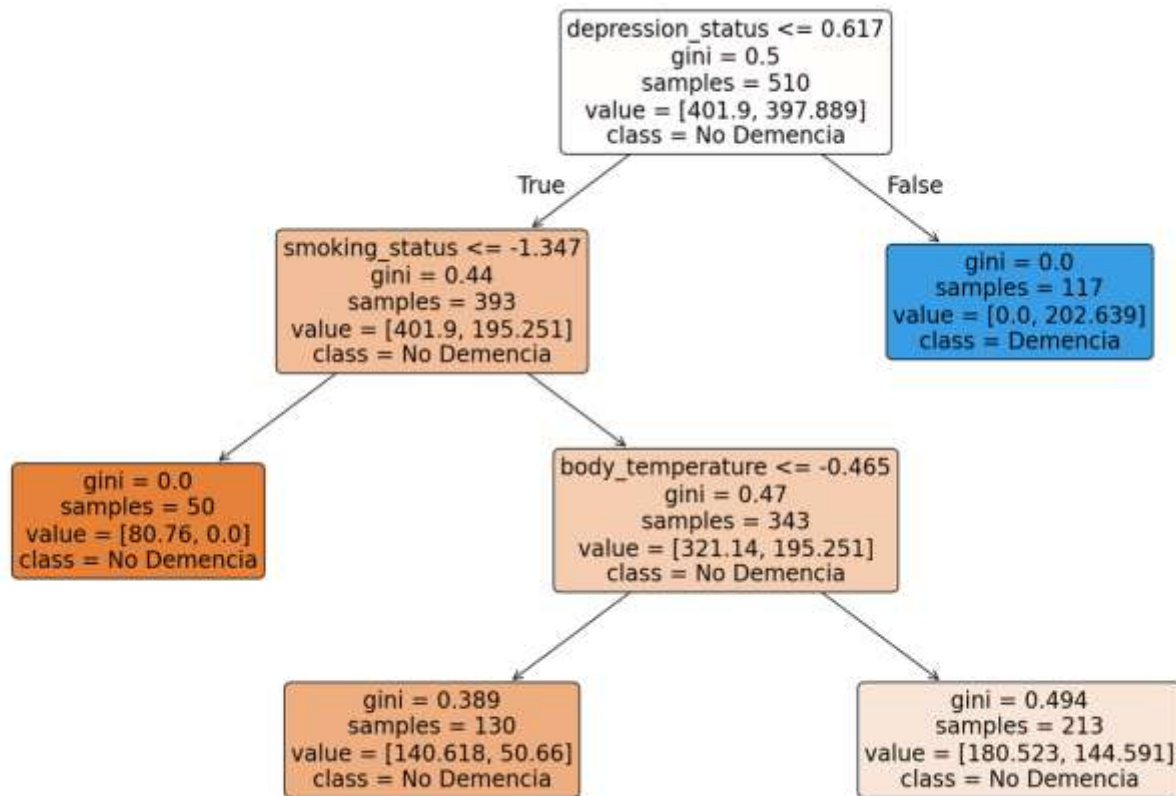
Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]

Curva ROC de Random Forest con PCA









# Comparación de Modelos Random Forest

	Accuracy	Precision	Recall	F1 score	Curva ROC	Clasificación
Random Forest	0.88	0.976744	0.792453	0.875000	0.885588	0
Random Forest PCA	0.93	0.960000	0.905660	0.932039	0.931554	0
Random Forest Importancia de Características	1.00	1.000000	1.000000	1.000000	1.000000	0

Mejor modelo: Random Forest aplicando Importancia de Características

# SVM

Separación de conjuntos

Escalado de características

Implementación de kernel RBF porque los datos no son linealmente separables

Búsqueda de hiperparámetros con Validación Cruzada

Reporte de clasificación:

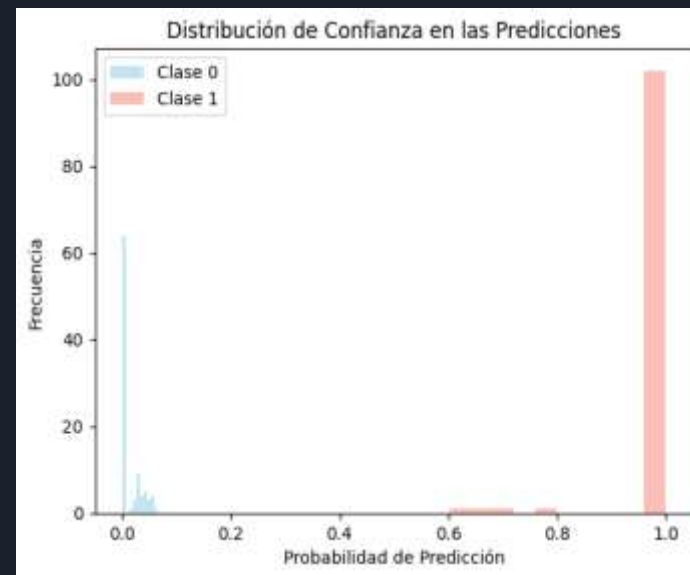
	precision	recall	f1-score	support
0	0.97	1.00	0.98	94
1	1.00	0.97	0.99	106
accuracy			0.98	200
macro avg	0.98	0.99	0.98	200
weighted avg	0.99	0.98	0.99	200

Exactitud del modelo: 0.985

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]

	0	1
0	94	0
1	3	103



# SVM aplicando PCA

Separación de conjuntos

Escalado de características

Aplicación de PCA a las características de los conjuntos

Búsqueda de hiperparámetros con Validación Cruzada

	0	1
0	92	2
1	12	94

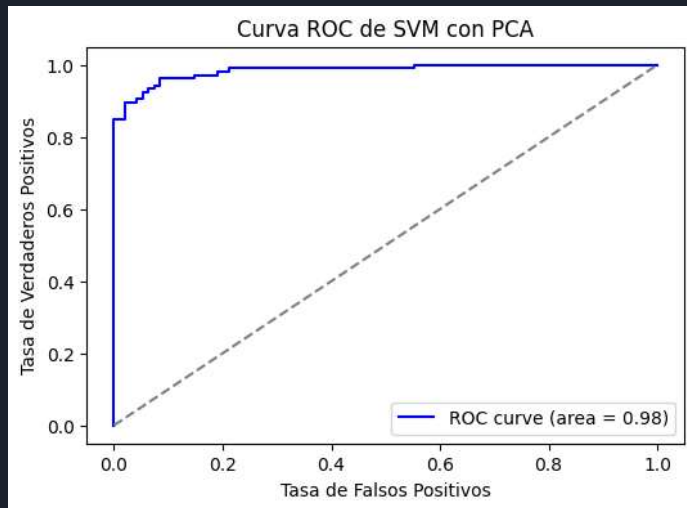
Reporte de clasificación:

	precision	recall	f1-score	support
0	0.88	0.98	0.93	94
1	0.98	0.89	0.93	106
accuracy			0.93	200
macro avg	0.93	0.93	0.93	200
weighted avg	0.93	0.93	0.93	200

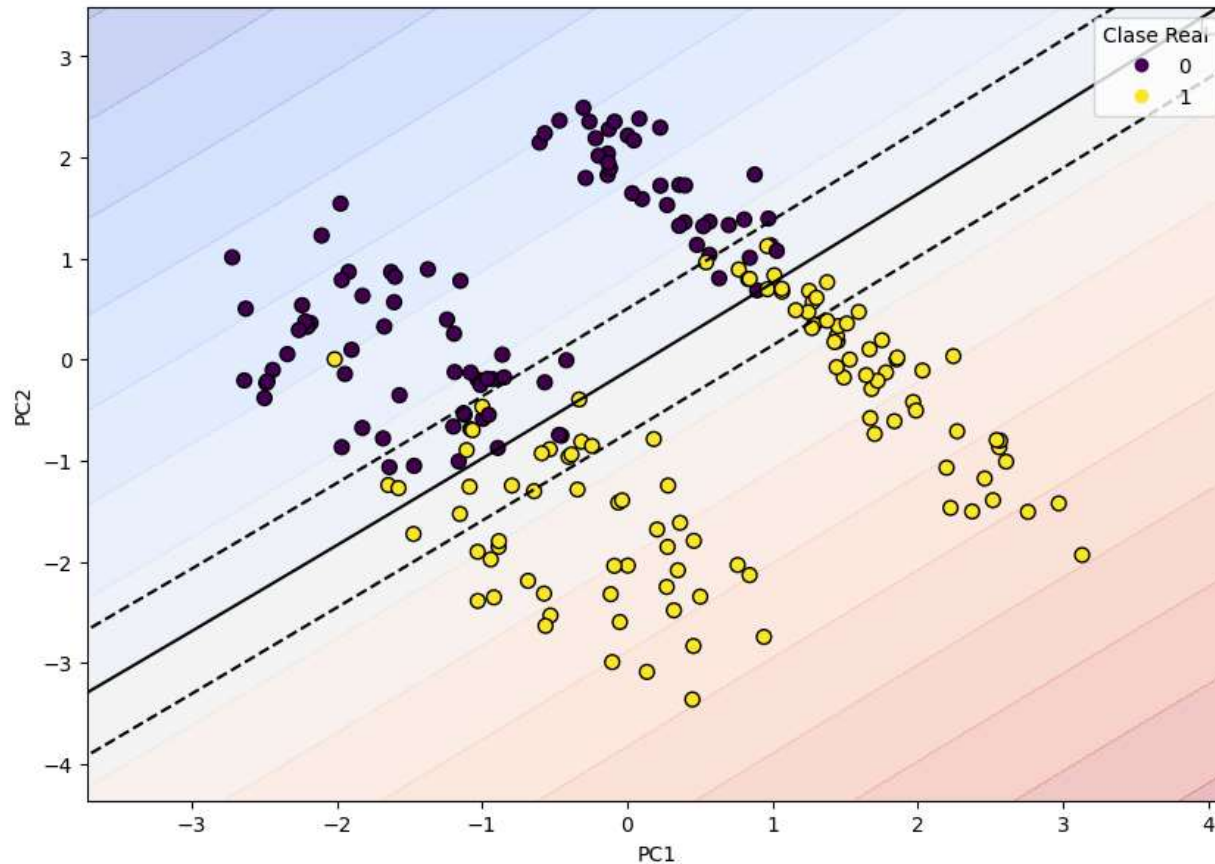
Exactitud del modelo: 0.93

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]



Clasificación de Demencia usando SVM con PCA



# SVM aplicando importancia de características

Separación de conjuntos

Escalado de características

Búsqueda de hiperparámetros con Validación

Cruzada

	0	1
0	94	0
1	0	106

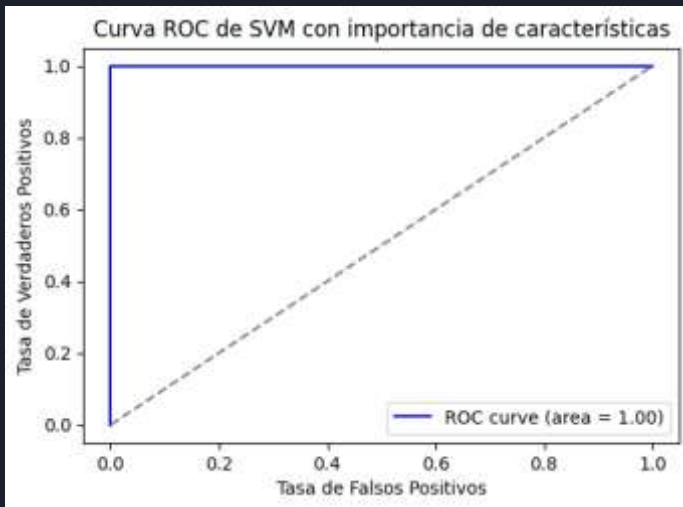
Reporte de clasificación:


	precision	recall	f1-score	support
0	1.00	1.00	1.00	94
1	1.00	1.00	1.00	106
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

Exactitud del modelo: 1.0

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]





# Comparación de Modelos de Máquinas de Soporte Vectorial

	Accuracy	Precision	Recall	F1 score	Curva ROC	Clasificación
SVM	0.985	1.000000	0.971698	0.985646	0.985849	0
SVM con PCA	0.930	0.979167	0.886792	0.930693	0.932758	0
SVM Importancia de Características	1.000	1.000000	1.000000	1.000000	1.000000	0

Mejor modelo: SVM aplicando Importancia de Características



# Mejor método de Regresión Logística

1. LogisticRegression y GrdSearchCV
2. LogisticRegressionCV
3. LogisticRegression

	Accuracy	Precision	Recall	F1 score	Curva ROC
Regresión Logística 1	0.980	1.0	0.962264	0.980769	0.981132
Regresión Logística 2	0.985	1.0	0.971698	0.985646	0.985849
Regresión Logística 3	0.980	1.0	0.962264	0.980769	0.981132

Mejor método: 'LogisticRegressionCV'



# Regresión Logística

Separación de conjuntos

Escalado de características

Penalidad 'Lasso' porque sólo una parte de las características son relevantes

	0	1
0	94	0
1	3	103

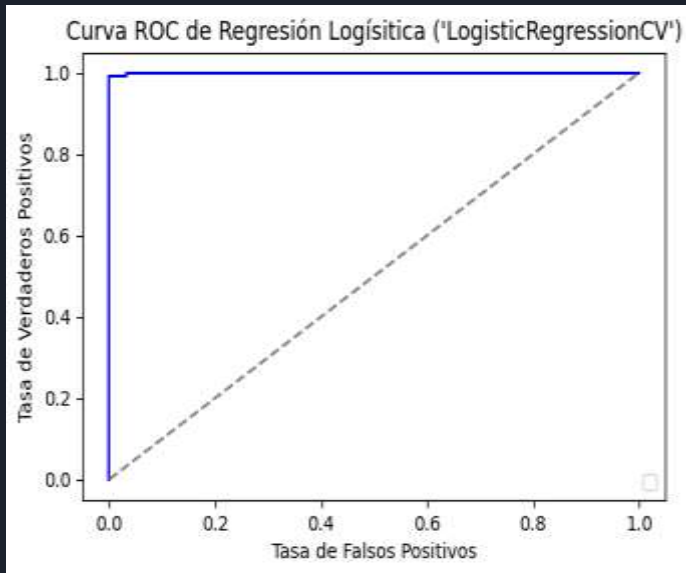
## Reporte de clasificación:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	94
1	1.00	0.97	0.99	106
accuracy			0.98	200
macro avg	0.98	0.99	0.98	200
weighted avg	0.99	0.98	0.99	200

Exactitud del modelo: 0.985

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]



# Regresión Logística aplicando PCA

Separación de conjuntos  
Escalado de características  
Aplicación de PCA a las características de los conjuntos  
Búsqueda de hiperparámetros

	0	1
0	89	5
1	11	95

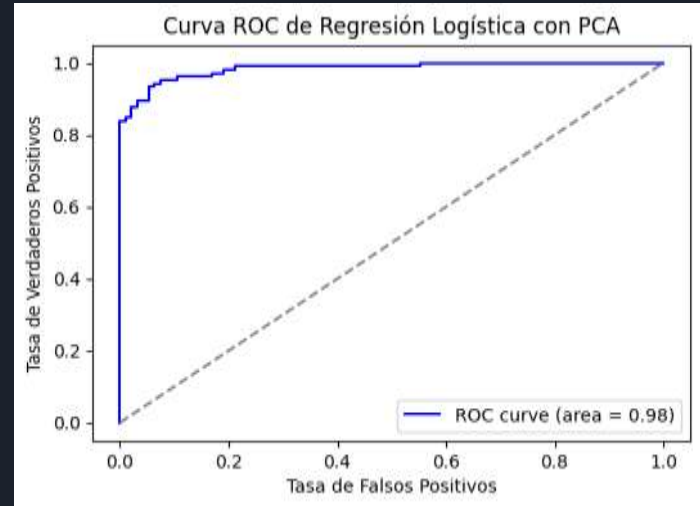
Reporte de clasificación:

	precision	recall	f1-score	support
0	0.89	0.95	0.92	94
1	0.95	0.90	0.92	106
accuracy			0.92	200
macro avg	0.92	0.92	0.92	200
weighted avg	0.92	0.92	0.92	200

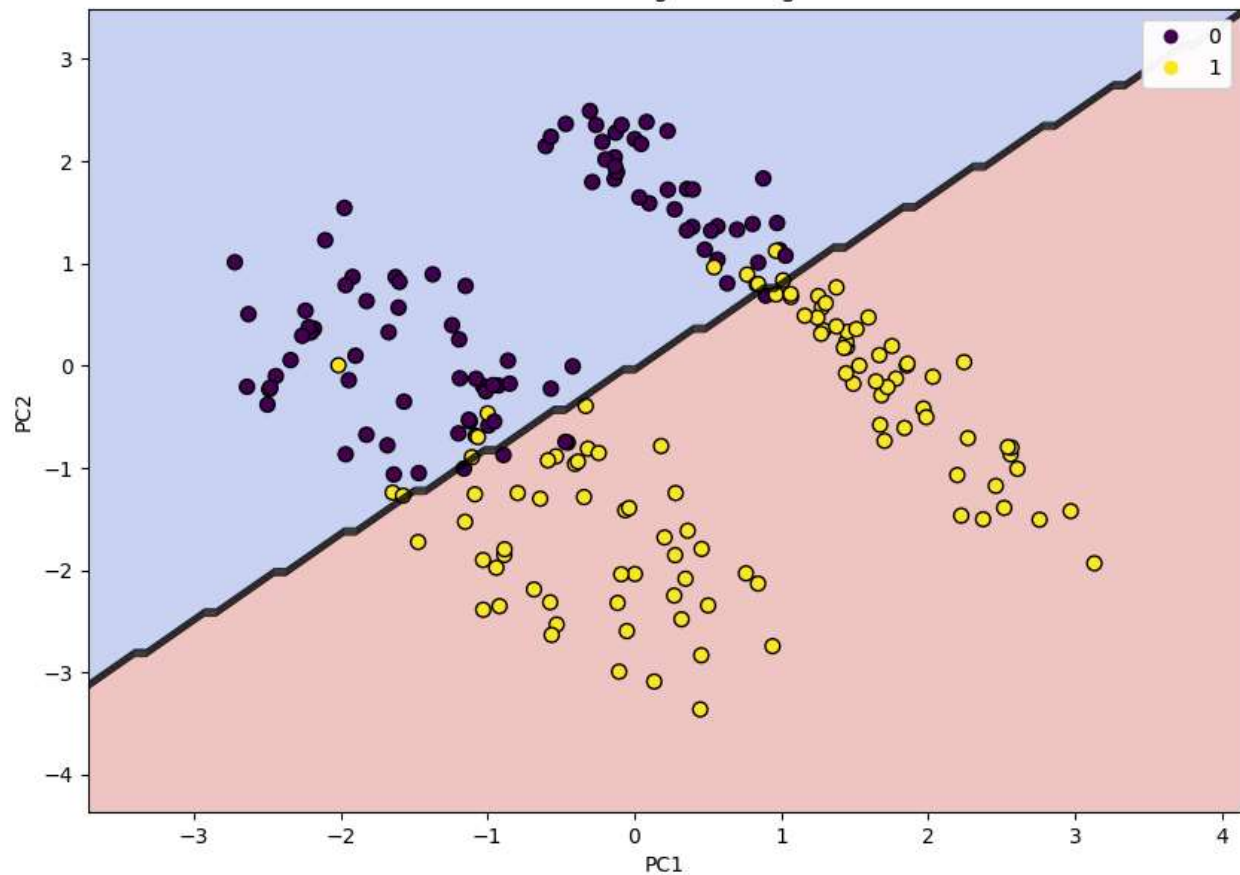
Exactitud del modelo: 0.92

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]



Plano de decisión de Regresión Logística con PCA



# Regresión Logística aplicando importancia de características

Separación de conjuntos

Escalado de características

Búsqueda de hiperparámetros

	0	1
0	94	0
1	2	104

## Reporte de clasificación:

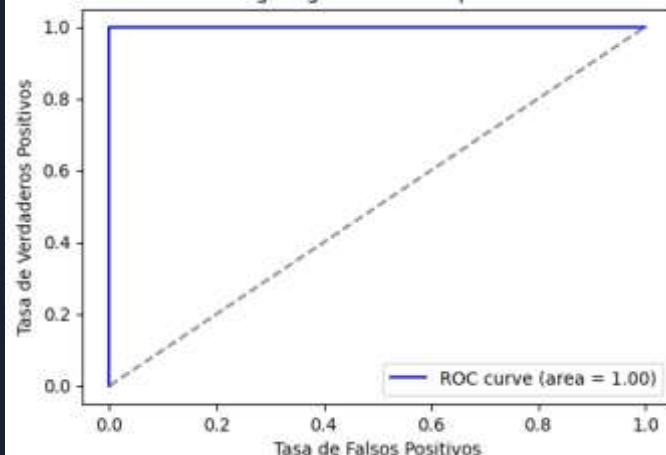
	precision	recall	f1-score	support
0	0.98	1.00	0.99	94
1	1.00	0.98	0.99	106
accuracy			0.99	200
macro avg	0.99	0.99	0.99	200
weighted avg	0.99	0.99	0.99	200

Exactitud del modelo: 0.99

Predicción de instancia 'sin demencia': [0]

Predicción de instancia 'con demencia': [1]

Curva ROC de Reg. Logística con importancia de caracs.





# Comparación de Modelos de Regresión Logística

	Accuracy	Precision	Recall	F1 score	Curva ROC	Clasificación
Regresión Logística	0.985	1.00	0.971698	0.985646	0.985849	0
Regresión Logística PCA	0.920	0.95	0.896226	0.922330	0.921517	0
Regresión Logística Importancia de Características	0.990	1.00	0.981132	0.990476	0.990566	0

A pesar de que el algoritmo requiere de un análisis de componentes principales, el modelo con 12 características presentó un mejor rendimiento.



# Comparación de Métricas

	Accuracy	Precision	Recall	F1 score	Curva ROC	Clasificación
Random Forest	0.880	0.976744	0.792453	0.875000	0.885588	0
SVC	0.985	1.000000	0.971698	0.985646	0.985849	0
Regresión Logística	0.985	1.000000	0.971698	0.985646	0.985849	0

Al comparar las métricas de los modelos con el total de características de cada algoritmo, concluimos tanto Máquinas de Soporte Vectorial como Regresión Logística tienen un mejor rendimiento para clasificar a pacientes con y sin demencia.