CIENCIA DE DATOS

Material suplementario

Fernanda G. Brenda S. Alejandra S. Elizabeth S.
Clase de Ciencia de Datos 2020
Módulo: Aprendizaje de Máquina

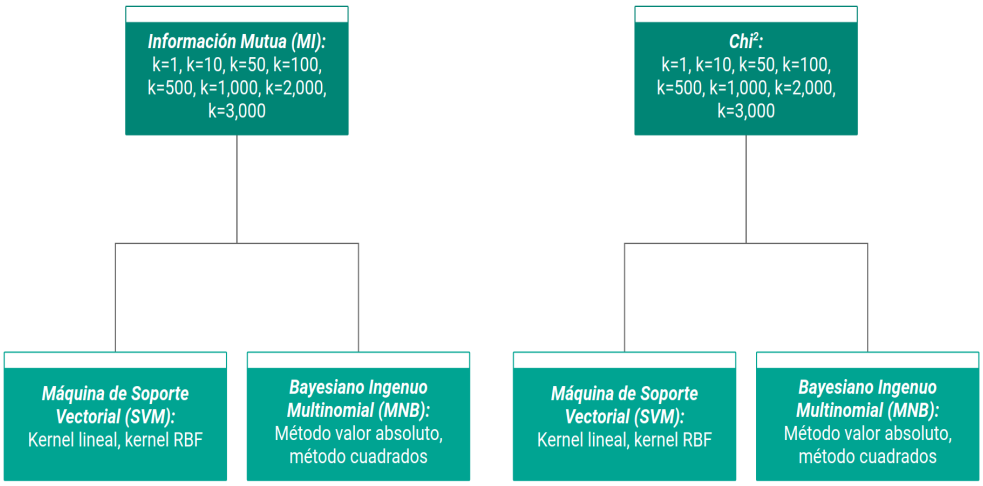

Figura suplementaria S1: Visualización gráfica de las combinaciones de clasificadores generadas para los datos de dominio estructural. Se aprecian los métodos de selección de características Información Mutua (MI) y Chi2. Cada uno utiliza el clasificador de Maquina de Soporte Vectorial (SVM) y Bayesiano Ingenuo Multinomial (MNB), con sus respectivos parámetros.

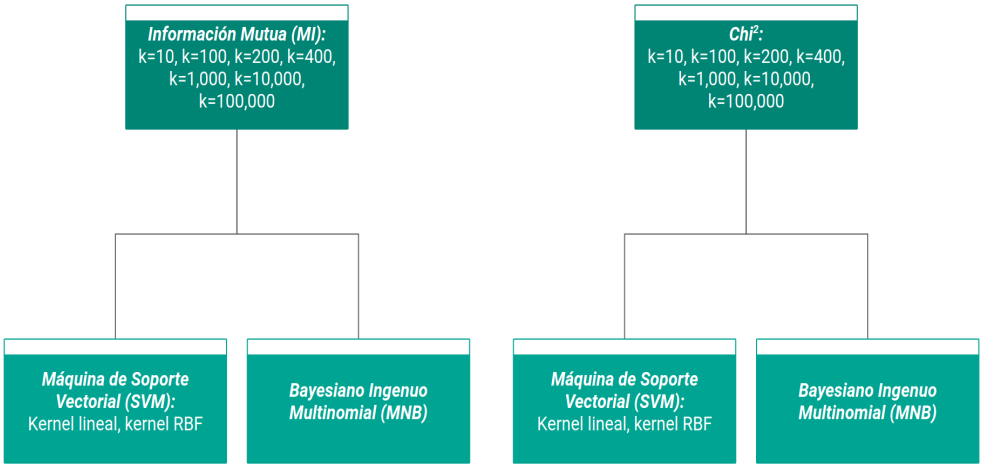

Figura suplementaria S2: Visualización gráfica de las combinaciones de clasificadores generadas para los datos de union a trombina. Se aprecian los métodos de selección de características Información Mutua (MI) y Chi2. Cada uno utiliza el clasificador de Maquina de Soporte Vectorial (SVM) y Bayesiano Ingenuo Multinomial (MNB), con sus respectivos parámetros.

Figura suplementaria S3: Visualización gráfica de las combinaciones de clasificadores generadas para los datos de unión a trombina. Se aprecian los métodos de reducción de dimensionalidad Análisis de Componentes Principales (PCA), Descomposición en Valores Singulares (SVD) e Incrustación Estocástica de Vecinos t-distribuida (tSNE). Cada uno utiliza el clasificador de Máquina de Soporte Vectorial (SVM) y Bayesiano Ingenuo Multinomial (MNB) con sus respectivos parámetros.

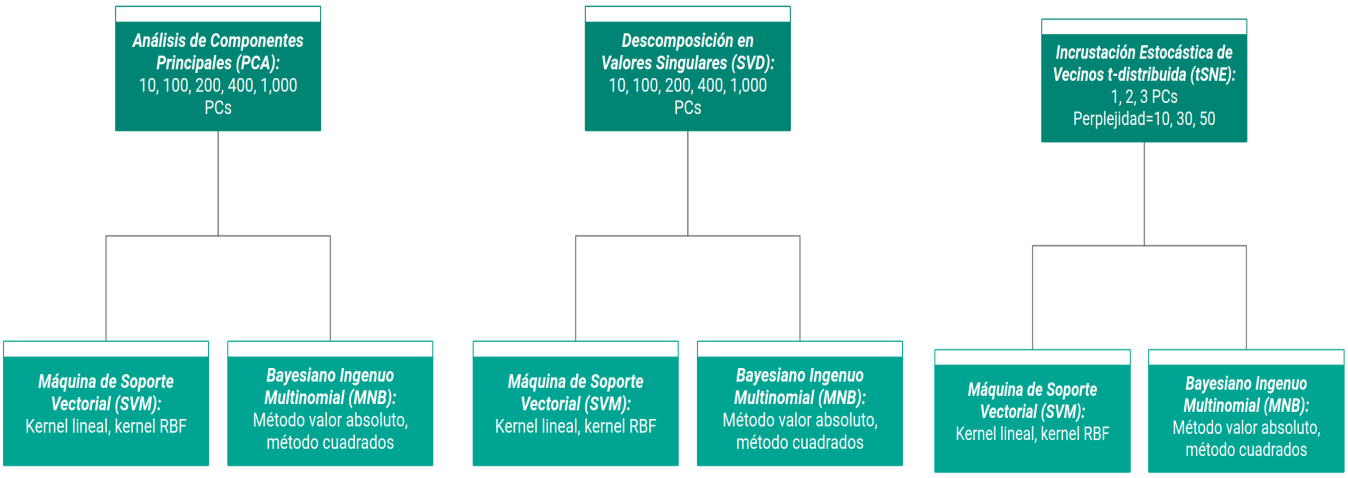

Tabla suplementaria S1: Todas las combinaciones que se corrieron para los datos de dominio estructural. La útlima columna cuenta con el nombre del archivo, mismos con los que se pueden encontrar en el repositorio.

Clasificador	Reducción/Selección	Elementos	Kernel/Positivo	F1-Score	Archivo
MNB	MI	1	absoluto	0	MNB_MI_1_absoluto.txt
MNB	MI	1	cuadrados	0	MNB_MI_1_cuadrados.txt
MNB	MI	10	absoluto	0.62	MNB_MI_10_absoluto.txt
MNB	MI	10	cuadrados	0.62	MNB_MI_10_cuadrados.txt
MNB	MI	50	absoluto	0.77	MNB_MI_50_absoluto.txt
MNB	MI	50	cuadrados	0.77	MNB_MI_50_cuadrados.txt
MNB	MI	100	absoluto	0.8	MNB_MI_100_absoluto.txt
MNB	MI	100	cuadrados	0.8	MNB_MI_100_cuadrados.txt
MNB	MI	500	absoluto	0.79	MNB_MI_500_absoluto.txt
MNB	MI	500	cuadrados	0.8	MNB_MI_500_cuadrados.txt
MNB	MI	1000	absoluto	0.78	MNB_MI_1000_absoluto.txt
MNB	MI	1000	cuadrados	0.78	MNB_MI_1000_cuadrados.txt
MNB	MI	2000	absoluto	0.78	MNB_MI_2000_absoluto.txt
MNB	MI	2000	cuadrados	0.78	MNB_MI_2000_cuadrados.txt

MNB	MI	3000	absoluto	0.78	MNB_MI_3000_absoluto.txt
MNB	MI	3000	cuadrados	0.78	MNB_MI_3000_cuadrados.txt
MNB	CHI2	1	absoluto	0	MNB_CHI2_1_absolute.txt
MNB	CHI2	1	cuadrados	0	MNB_CHI2_1_squares.txt
MNB	CHI2	10	absoluto	0.61	MNB_CHI2_10_absolute.txt
MNB	CHI2	10	cuadrados	0.61	MNB_CHI2_10_squares.txt
MNB	CHI2	50	absoluto	0.76	MNB_CHI2_50_absolute.txt
MNB	CHI2	50	cuadrados	0.76	MNB_CHI2_50_squares.txt
MNB	CHI2	100	absoluto	0.78	MNB_CHI2_100_absolute.txt
MNB	CHI2	100	cuadrados	0.78	MNB_CHI2_100_squares.txt
MNB	CHI2	500	absoluto	0.8	MNB_CHI2_500_absolute.txt
MNB	CHI2	500	cuadrados	0.8	MNB_CHI2_500_squares.txt
MNB	CHI2	1000	absoluto	0.78	MNB_CHI2_1000_absolute.txt
MNB	CHI2	1000	cuadrados	0.78	MNB_CHI2_1000_squares.txt
MNB	CHI2	2000	absoluto	0.78	MNB_CHI2_2000_absolute.txt
MNB	CHI2	2000	cuadrados	0.78	MNB_CHI2_2000_squares.txt
MNB	CHI2	3000	absoluto	0.78	MNB_CHI2_3000_absolute.txt
MNB	CHI2	3000	cuadrados	0.78	MNB_CHI2_3000_squares.txt
SVM	MI	1	linear	0.61	SVM_MI_1_linear.txt
SVM	MI	1	RBF	0.61	SVM_MI_1_RBF.txt
SVM	MI	10	linear	0.71	SVM_MI_10_linear.txt
SVM	MI	10	RBF	0.74	SVM_MI_10_RBF.txt
SVM	MI	50	linear	0.76	SVM_MI_50_linear.txt
SVM	MI	50	RBF	0.75	SVM_MI_50_RBF.txt
SVM	MI	100	linear	0.78	SVM_MI_100_linear.txt
SVM	MI	100	RBF	0.75	SVM_MI_100_RBF.txt
SVM	MI	500	linear	0.75	SVM_MI_500_linear.txt
SVM	MI	500	RBF	0.73	SVM_MI_500_RBF.txt
SVM	MI	1000	linear	0.73	SVM_MI_1000_linear.txt
SVM	MI	1000	RBF	0.75	SVM_MI_1000_RBF.txt
SVM	MI	2000	linear	0.74	SVM_MI_2000_linear.txt
SVM	MI	2000	RBF	0.76	SVM_MI_2000_RBF.txt
SVM	MI	3000	linear	0.75	SVM_MI_3000_linear.txt
SVM	MI	3000	RBF	0.73	SVM_MI_3000_RBF.txt
SVM	CHI2	1	lineal	0.61	SVM_CHI2_1_linear.txt
SVM	CHI2	1	RBF	0.61	SVM_CHI2_1_RBF.txt

SVM	CHI2	10	lineal	0.71	SVM_CHI2_10_linear.txt
SVM	CHI2	10	RBF	0.75	SVM_CHI2_10_RBF.txt
SVM	CHI2	50	lineal	0.78	SVM_CHI2_50_linear.txt
SVM	CHI2	50	RBF	0.76	SVM_CHI2_50_RBF.txt
SVM	CHI2	100	lineal	0.75	SVM_CHI2_100_linear.txt
SVM	CHI2	100	RBF	0.73	SVM_CHI2_100_RBF.txt
SVM	CHI2	500	lineal	0.77	SVM_CHI2_500_linear.txt
SVM	CHI2	500	RBF	0.74	SVM_CHI2_500_RBF.txt
SVM	CHI2	1000	lineal	0.7	SVM_CHI2_1000_linear.txt
SVM	CHI2	1000	RBF	0.74	SVM_CHI2_1000_RBF.txt
SVM	CHI2	2000	lineal	0.75	SVM_CHI2_2000_linear.txt
SVM	CHI2	2000	RBF	0.75	SVM_CHI2_2000_RBF.txt
SVM	CHI2	3000	lineal	0.76	SVM_CHI2_3000_linear.txt
SVM	CHI2	3000	RBF	0.79	SVM_CHI2_3000_RBF.txt

Tabla suplementaria S2: Primeras 10 características con mejor puntuación seleccionadas por los 3 clasificadores con mejor rendimiento para el conjunto de datos de dominio estructural.

MNB, CHI2 500 características	Puntuación	MNB, MI 100 características	Puntuación	MNB, MI 500 características	Puntuación
Domain	554.5507	Domain	0.132413	Domain	0.132413
Terminal	499.8154	Terminal	0.115568	Terminal	0.115568
Helix	387.3758	Helix	0.083669	Helix	0.083669
Turn	316.6974	Turn	0.067167	Turn	0.067167
Domains	292.8823	Domains	0.062196	Domains	0.062196
Motif	229.5909	Motif	0.048325	Motif	0.048325
Contains	204.5237	Contains	0.043501	Contains	0.043501
Dimerization	111.485	Binding	0.033897	Binding	0.033897
Binding	103.3191	DNA	0.030428	DNA	0.030428
DNA	93.91634	Family	0.022996	Family	0.022996

Tabla suplementaria S3: Todas las combinaciones que se corrieron para los datos de unión a la trombina. La útlima columna cuenta con el nombre del archivo, mismos con los que se pueden encontrar en el repositorio.

Clasificador	Reducción/Selección	Elementos	Kernel/Positivo	Perplejidad	F1-Score	Archivo
MNB	MI	10	-	-	0	MNB_MI_10.txt
MNB	MI	100	-	-	0.34	MNB_MI_100.txt
MNB	MI	200	-	-	0.3	MNB_MI_200.txt
MNB	MI	400	-	-	0.3	MNB_MI_400.txt
MNB	MI	1000	-	-	0.28	MNB_MI_1000.txt
MNB	MI	10000	-	-	0.26	MNB_MI_10000.txt
MNB	MI	100000	-	-	0.21	MNB_MI_100000.txt
MNB	CHI2	10	-	-	0.5	MNB_CHI2_10.txt
MNB	CHI2	100	-	-	0.37	MNB_CHI2_100.txt
MNB	CHI2	200	-	-	0.39	MNB_CHI2_200.txt
MNB	CHI2	400	-	-	0.35	MNB_CHI2_400.txt
MNB	CHI2	1000	-	-	0.25	MNB_CHI2_1000.txt
MNB	CHI2	10000	-	-	0.23	MNB_CHI2_10000.txt
MNB	CHI2	100000	-	-	0.19	MNB_CHI2_100000.txt
MNB	PCA	10	absoluto	-	0.29	MNB_PCA_10_absolute.txt
MNB	PCA	10	cuadrados	-	0.37	MNB_PCA_10_squares.txt
MNB	PCA	100	absoluto	-	0.3	MNB_PCA_100_absolute.txt
MNB	PCA	100	cuadrados	-	0.28	MNB_PCA_100_squares.txt
MNB	PCA	200	absoluto	-	0.25	MNB_PCA_200_absolute.txt
MNB	PCA	200	cuadrados	-	0.31	MNB_PCA_200_squares.txt
MNB	PCA	400	absoluto	-	0.35	MNB_PCA_400_absolute.txt
MNB	PCA	400	cuadrados	-	0.33	MNB_PCA_400_squares.txt
MNB	PCA	1000	absoluto	-	0.38	MNB_PCA_1000_absolute.txt
MNB	PCA	1000	cuadrados	-	0.37	MNB_PCA_1000_squares.txt
MNB	SVD	10	absoluto	-	0.3	MNB_SVD_10_absolute.txt
MNB	SVD	10	cuadrados	-	0.27	MNB_SVD_10_squares.txt
MNB	SVD	100	absoluto	-	0.24	MNB_SVD_100_absolute.txt
MNB	SVD	100	cuadrados	-	0.21	MNB_SVD_100_squares.txt
MNB	SVD	200	absoluto	-	0.32	MNB_SVD_200_absolute.txt
MNB	SVD	200	cuadrados	-	0.29	MNB_SVD_200_squares.txt
MNB	SVD	400	absoluto	-	0.36	MNB_SVD_400_absolute.txt
MNB	SVD	400	cuadrados	-	0.35	MNB_SVD_400_squares.txt
MNB	SVD	1000	absoluto	-	0.37	MNB_SVD_1000_absolute.txt
MNB	SVD	1000	cuadrados	-	0.34	MNB_SVD_1000_squares.txt
MNB	tSNE	1	absoluto	10	0	MNB_tSNE_1_absoluto_10.txt
MNB	tSNE	1	absoluto	30	0	MNB_tSNE_1_absoluto_30.txt
MNB	tSNE	1	absoluto	50	0	MNB_tSNE_1_absoluto_50.txt
MNB	tSNE	1	cuadrados	10	0	MNB_tSNE_1_cuadrados_10.txt

MNB	tSNE	1	cuadrados	30	0	MNB_tSNE_1_cuadrados_30.txt
MNB	tSNE	1	cuadrados	50	0	MNB_tSNE_1_cuadrados_50.txt
MNB	tSNE	2	absoluto	10	0.15	MNB_tSNE_2_absoluto_10.txt
MNB	tSNE	2	absoluto	30	0.03	MNB_tSNE_2_absoluto_30.txt
MNB	tSNE	2	absoluto	50	0	MNB_tSNE_2_absoluto_50.txt
MNB	tSNE	2	cuadrados	10	0.14	MNB_tSNE_2_cuadrados_10.txt
MNB	tSNE	2	cuadrados	30	0.03	MNB_tSNE_2_cuadrados_30.txt
MNB	tSNE	2	cuadrados	50	0.13	MNB_tSNE_2_cuadrados_50.txt
MNB	tSNE	3	absoluto	10	0.25	MNB_tSNE_3_absoluto_10.txt
MNB	tSNE	3	absoluto	30	0.36	MNB_tSNE_3_absoluto_30.txt
MNB	tSNE	3	absoluto	50	0.31	MNB_tSNE_3_absoluto_50.txt
MNB	tSNE	3	cuadrados	10	0.27	MNB_tSNE_3_cuadrados_10.txt
MNB	tSNE	3	cuadrados	30	0.37	MNB_tSNE_3_cuadrados_30.txt
MNB	tSNE	3	cuadrados	50	0.34	MNB_tSNE_3_cuadrados_50.txt
SVM	MI	10	lineal	-	0.28	SVM_MI_10_lineal.txt
SVM	MI	10	RBF	-	0.29	SVM_MI_10_RBF.txt
SVM	MI	100	lineal	-	0.38	SVM_MI_100_lineal.txt
SVM	MI	100	RBF	-	0.33	SVM_MI_100_RBF.txt
SVM	MI	200	lineal	-	0.35	SVM_MI_200_lineal.txt
SVM	MI	200	RBF	-	0.29	SVM_MI_200_RBF.txt
SVM	MI	400	lineal	-	0.3	SVM_MI_400_lineal.txt
SVM	MI	400	RBF	-	0.28	SVM_MI_400_RBF.txt
SVM	MI	1000	lineal	-	0.25	SVM_MI_1000_linear.txt
SVM	MI	1000	RBF	-	0.28	SVM_MI_1000_RBF.txt
SVM	MI	10000	lineal	-	0.23	SVM_MI_10000_linear.txt
SVM	MI	10000	RBF	-	0.23	SVM_MI_10000_RBF.txt
SVM	MI	100000	lineal	-	0.22	SVM_MI_100000_linear.txt
SVM	MI	100000	RBF	-	0.22	SVM_MI_100000_RBF.txt
SVM	CHI2	10	lineal	-	0.45	SVM_CHI2_10_linear.txt
SVM	CHI2	10	RBF	-	0.43	SVM_CHI2_10_RBF.txt
SVM	CHI2	100	lineal	-	0.3	SVM_CHI2_100_linear.txt
SVM	CHI2	100	RBF	-	0	SVM_CHI2_100_RBF.txt
SVM	CHI2	200	lineal	-	0.34	SVM_CHI2_200_linear.txt
SVM	CHI2	200	RBF	-	0.3	SVM_CHI2_200_RBF.txt
SVM	CHI2	400	lineal	-	0.36	SVM_CHI2_400_linear.txt
SVM	CHI2	400	RBF	-	0.35	SVM_CHI2_400_RBF.txt
SVM	CHI2	1000	lineal	-	0.28	SVM_CHI2_1000_linear.txt
SVM	CHI2	1000	RBF	-	0	SVM_CHI2_1000_RBF.txt
SVM	CHI2	10000	lineal	-	0.25	SVM_CHI2_10000_linear.txt
SVM	CHI2	10000	RBF	-	0	SVM_CHI2_10000_RBF.txt
SVM	CHI2	100000	lineal	-	0.24	SVM_CHI2_100000_linear.txt
SVM	PCA	100	RBF	-	0	SVM_PCA_100_RBF.txt
SVM	PCA	200	lineal	-	0.28	SVM_PCA_200_linear.txt

SVM	PCA	200	RBF	-	0	SVM_PCA_200_RBF.txt
SVM	PCA	400	lineal	-	0.26	SVM_PCA_400_linear.txt
SVM	PCA	400	RBF	-	0	SVM_PCA_400_RBF.txt
SVM	PCA	1000	lineal	-	0.21	SVM_PCA_1000_linear.txt
SVM	PCA	1000	RBF	-	0	SVM_PCA_1000_RBF.txt
SVM	SVD	10	lineal	-	0.34	SVM_SVD_10_linear.txt
SVM	SVD	10	RBF	-	0.33	SVM_SVD_10_RBF.txt
SVM	SVD	100	lineal	-	0.3	SVM_SVD_100_lineal.txt
SVM	SVD	100	RBF	-	0	SVM_SVD_100_RBF.txt
SVM	SVD	200	lineal	-	0.31	SVM_SVD_200_lineal.txt
SVM	SVD	200	RBF	-	0	SVM_SVD_200_RBF.txt
SVM	SVD	400	lineal	-	0.24	SVM_SVD_400_lineal.txt
SVM	SVD	400	RBF	-	0	SVM_SVD_400_RBF.txt
SVM	SVD	1000	lineal	-	0.22	SVM_SVD_1000_lineal.txt
SVM	SVD	1000	RBF	-	0	SVM_SVD_1000_RBF.txt
SVM	tSNE	1	lineal	10	0	SVM_tSNE_1_lineal_10.txt
SVM	tSNE	1	lineal	30	0	SVM_tSNE_1_lineal_30.txt
SVM	tSNE	1	lineal	50	0	SVM_tSNE_1_lineal_50.txt
SVM	tSNE	1	RBF	10	0	SVM_tSNE_1_RBF_10.txt
SVM	tSNE	1	RBF	30	0.05	SVM_tSNE_1_RBF_30.txt
SVM	tSNE	1	RBF	50	0	SVM_tSNE_1_RBF_50.txt
SVM	tSNE	2	lineal	10	0	SVM_tSNE_2_lineal_10.txt
SVM	tSNE	2	lineal	30	0	SVM_tSNE_2_lineal_30.txt
SVM	tSNE	2	lineal	50	0	SVM_tSNE_2_lineal_50.txt
SVM	tSNE	2	RBF	10	0	SVM_tSNE_2_RBF_10.txt
SVM	tSNE	2	RBF	30	0.05	SVM_tSNE_2_RBF_30.txt
SVM	tSNE	2	RBF	50	0.03	SVM_tSNE_2_RBF_50.txt
SVM	tSNE	3	lineal	10	0	SVM_tSNE_3_lineal_10.txt
SVM	tSNE	3	lineal	30	0	SVM_tSNE_3_lineal_30.txt
SVM	tSNE	3	lineal	50	0	SVM_tSNE_3_lineal_50.tx
SVM	tSNE	3	RBF	10	0	SVM_tSNE_1_RBF_10.txt
SVM	tSNE	3	RBF	30	0	SVM_tSNE_1_RBF_30.txt
SVM	tSNE	3	RBF	50	0	SVM_tSNE_1_RBF_50.txt