UNIVERSIDAD TECNOLÓGICA METROPOLITANA
del Estado de Chile

Tarea N° 6

Minería de datos

“Aplicación Redes Bayesianas y SVM”

Sección: 412

Docente: Pablo Figueroa

Integrantes:

- Ricardo Castillo
- Diego Moya
- Jaime Orellana

Fecha: 17 / 11 / 2025

1. Metodología

Para esta actividad se utilizó el mismo conjunto de datos **banco.csv**, compuesto por 5.000 registros y 11 variables explicativas asociadas al perfil del cliente (edad, experiencia laboral, ingreso, tamaño de familia, gasto en tarjeta de crédito, nivel educacional, monto de hipoteca, valores de inversión, posesión de certificado, uso de banca en línea y uso de tarjeta de crédito). La variable objetivo corresponde a **prestamo**, codificada como 0 (sin préstamo) y 1 (con préstamo), con un desbalance moderado: 4.520 casos sin préstamo (90,4%) y 480 con préstamo (9,6%).

Como preprocesamiento, las variables numéricas se convirtieron explícitamente a tipo numérico y los valores faltantes se imputaron mediante la mediana de cada columna. Posteriormente se eliminaron los registros con valores faltantes en la variable objetivo. Luego, los datos se particionaron en un conjunto de entrenamiento (70%) y un conjunto de prueba (30%), utilizando train_test_split con stratify=y para mantener la proporción de clases. De este modo se entrenaron los modelos sobre 3.500 observaciones y se evaluaron sobre 1.500.

Para cumplir el objetivo de la tarea, se implementan **Máquinas de Vectores de Soporte (SVM)** con **tres kernels distintos**:

- SVM lineal (kernel="linear").
- SVM con kernel RBF (gaussiano, kernel="rbf").
- SVM con kernel polinomial (kernel="poly", grado 3).

Dado que SVM es sensible a la escala de las variables, cada modelo se construyó como un **Pipeline** con dos etapas, **estandarización de los predictores mediante StandardScaler**, y **entrenamiento del clasificador SVM.**

La evaluación se realizó en dos niveles:

- Validación cruzada estratificada de 5 pliegues (5-fold CV) sobre el conjunto completo (X, y), empleando StratifiedKFold para preservar el desbalance de clases en cada fold.

- Evaluación final en el conjunto de prueba (hold-out) definido previamente (30% de los datos).

En ambos casos se utilizó la exactitud (accuracy) como métrica principal, complementada con el error de clasificación (1 − accuracy) y las métricas por clase (precisión, recall y F1-score) en el conjunto de pruebas.

2. Resultados

2.1. Resultados de validación cruzada (5-fold)

En primer lugar, se aplicó validación cruzada estratificada de 5 pliegues para cada uno de los tres kernels. La **Tabla 1** resume la **accuracy promedio** y la **desviación estándar** en los cinco folds.

Tabla 1: Resultados de validación cruzada para los modelos SVM (5-fold).

Modelo	Accuracy promedio CV	Desviación estándar
SVM lineal	0,9526	0,0037
SVM RBF	0,9734	0,0052
SVM polinomial	0,9736	0,0043

Los resultados muestran que:

- El **SVM lineal** obtiene una accuracy promedio cercana a 0,95, con baja variabilidad entre pliegues.
- Los modelos **SVM RBF** y **SVM polinomial** alcanzan una accuracy promedio alrededor de **0,97**, superando al modelo lineal y manteniendo también una desviación estándar pequeña.

Esto sugiere que la relación entre las variables explicativas y la variable objetivo **no es estrictamente lineal**, y que modelos más flexibles (RBF y polinomial) se adaptan mejor a la estructura de los datos.

2.2. Resultados en el conjunto de prueba (hold-out)

A continuación, se evaluó cada modelo SVM sobre el conjunto de prueba independiente (30% de los datos, 1.500 observaciones). En la **Tabla 2** se presenta la accuracy y el error de clasificación para cada kernel.

Tabla 2: Resultados de los modelos SVM en el conjunto de prueba.

Modelo	Accuracy en test	Error de clasificación
SVM lineal	0,9553	0,0447
SVM RBF	0,9747	0,0253
SVM polinomial	0,9753	0,0247

Además de la accuracy global, es relevante observar el comportamiento sobre la **clase minoritaria** (clientes con préstamo = 1). A partir de los reportes de clasificación:

- **SVM lineal**
    - Precisión (clase 1): 0,91
    - Recall (clase 1): 0,59

El modelo identifica correctamente solo alrededor del **59%** de los clientes que sí tienen préstamo.

- **SVM RBF**
    - Precisión (clase 1): 0,96
    - Recall (clase 1): 0,77

Mejora tanto la accuracy global como la capacidad para **detectar la clase con préstamo**, elevando el recall a un 77%.

- **SVM polinomial**
    - Precisión (clase 1): 0,96
    - Recall (clase 1): 0,78

Aquí se presenta el mejor desempeño sobre la clase minoritaria, con un recall cercano al 78%, manteniendo una accuracy global similar a la del modelo RBF.

En términos prácticos, los kernels RBF y polinomial logran no solo un **menor error de clasificación total**, sino también una **mejor detección de los clientes con préstamo**, que es la clase menos representada y generalmente de mayor interés para el análisis.

2.3. Comparación global de los SVM

Integrando los resultados de la validación cruzada y del conjunto de prueba, se observa que:

- El **SVM lineal** tiene buen rendimiento general, pero se queda por debajo de las variantes no lineales tanto en accuracy como en recall de la clase positiva.

- Los modelos **SVM RBF** y **SVM polinomial** presentan **accuracies muy similares** (≈ 0,97–0,98) y una mejor capacidad para identificar a los clientes con préstamo.

- El **SVM polinomial** obtiene ligeramente la mejor combinación de accuracy en test (0,9753) y recall para la clase con préstamo (0,78), aunque la diferencia con RBF es pequeña.

Para integrar los resultados de la validación cruzada y del conjunto de prueba, se construyó una tabla comparativa con las métricas más relevantes de cada modelo SVM: accuracy promedio en Cross Validation, accuracy en el conjunto de prueba y recall de la clase minoritaria (préstamo = 1).

Tabla 3. Comparación global de los modelos SVM.

Modelo	Accuracy promedio CV	Accuracy en test	Recall clase préstamo = 1 (test)
SVM lineal	0,9526	0,9553	0,59
SVM RBF	0,9734	0,9747	0,77
SVM polinomial	0,9736	0,9753	0,78

A partir de la **Tabla 3** se observa que el SVM lineal presenta el rendimiento más bajo de los tres modelos, tanto en la validación cruzada como en el conjunto de prueba, con accuracies en torno al 95%. En cambio, los modelos con kernels RBF y polinomial alcanzan accuracies cercanas al 97% en ambos escenarios, lo que indica un mejor ajuste a la estructura de los datos.

La diferencia es aún más clara al analizar el recall de la clase préstamo = 1, que corresponde a la clase minoritaria. El SVM lineal solo logra identificar correctamente cerca del 59% de los clientes que tienen préstamo, mientras que los modelos RBF y polinomial elevan este valor a aproximadamente 77% y 78%, respectivamente. Esto significa que los kernels no lineales no solo reducen el error global, sino que también mejoran de forma importante la detección de la clase positiva, que suele ser la de mayor interés en este tipo de aplicaciones.

En conjunto, los resultados de la validación cruzada y del conjunto de prueba coinciden en que los SVM con kernels no lineales (RBF y polinomial) ofrecen un desempeño superior al modelo lineal, tanto en términos de exactitud total como en la identificación de los clientes con préstamo.

3. Conclusiones

Se aplicaron Máquinas de Vectores de Soporte (SVM) con tres kernels distintos (lineal, RBF y polinomial) sobre el conjunto de datos bancarios, utilizando estandarización de variables y validación cruzada estratificada de 5 pliegues para evaluar cada modelo.

La validación cruzada mostró que los modelos con kernels **RBF** y **polinomial** obtienen una **accuracy promedio cercana a 0,97**, mientras que el SVM lineal se mantiene alrededor de 0,95. Esto sugiere que la relación entre las variables del problema y la variable objetivo no es puramente lineal y se beneficia de fronteras de decisión más flexibles.

En el conjunto de prueba, los SVM RBF y polinomial alcanzan **errores de clasificación cercanos al 2,5%**, frente a un 4,5% aproximadamente del SVM lineal. Además, el **recall de la clase préstamo = 1** aumenta desde un 59% en el modelo lineal hasta valores cercanos al 77–78% en los modelos no lineales, lo que implica una mejor detección de la clase minoritaria.

Considerando tanto la validación cruzada como el desempeño en el conjunto de prueba, los SVM con kernels **RBF** y **polinomial** se posicionan como las opciones más adecuadas para este conjunto de datos, ya que combinan alta exactitud global con una mejor capacidad para identificar a los clientes que poseen un préstamo.