Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores
Unidad Morelia

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES UNIDAD MORELIA

Tecnologías para la información en ciencias

Introducción a la ciencia de datos

Métricas y validación cruzada

Araceli Romero

424057545

Lunes 21 de Abril 2025

Índice

1. Introducción

El objetivo de este proyecto es clasificar países en tres categorías de ingreso (Bajo, Medio, Alto) según su Producto Interno Bruto per cápita (GDP), utilizando modelos de aprendizaje supervisado basados en Máquinas de Soporte Vectorial (SVM). Se evaluaron tres variantes de SVM con diferentes kernels (Lineal, Polinomial y RBF) mediante validación cruzada de 5 partes, considerando métricas como exactitud (accuracy), precisión (precision), sensibilidad (recall) y F1-score. El análisis se realizó en Python utilizando la biblioteca scikit-learn.

2. Limpieza de Datos

Primero los datos con valores nulos se reemplazaron con 0, se eliminaron registros con GDP menor a 50 (outliers) y valores inconsistentes principalmente en la columna ‘Population’ (se quitaron los registros del mundo, de la Unión Europea y los que no tenían habitantes). Las variables ‘Area’ y ‘Population’ se convirtieron a float e int respectivamente, para facilitar el procesamiento de los datos.

3. Procesamiento de Datos

Posteriormente se definieron 3 categorías según el GDP:

- Bajo: $50 \leq \text{GDP} \leq 5{,}000$
- Medio: $5{,}000 < \text{GDP} \leq 25{,}000$
- Alto: $\text{GDP} > 25{,}000$

Para definir las variables de entrenamiento y de prueba se consideró:
X: Se utilizaron características como área, tasa de desempleo, deuda externa, usuarios de internet, aeropuertos, entre otras. De este conjunto se excluyó 'GDP (USD)', 'GDP_Categoria', e 'Internet Users', ya que las primeras correspondían a la variable Y (conjunto a predecir) y la de 'Internet Users' se vio que perjudicaba la predicción ya que el porcentaje de usuarios era otra columna, por lo que se decidió eliminar.

Y: La categoría de GDP (GDP_Categoria).

División de datos: 80% entrenamiento, 20% prueba.

4. Modelos Evaluados

Se implementaron tres variantes de SVM con diferentes kernels. Enseguida se presenta cada uno de ellos con los parámetros que lograron el mejor desempeño.

1. SVM Lineal

- Parámetros optimizados: C (regularización).
- Mejor configuración: C = 7.
- Exactitud (CV): 76.8%

2. SVM Polinomial

- Parámetros optimizados: C y degree (grado del polinomio).
- Mejor configuración: C = 13, degree = 2.
- Exactitud (CV): 73.5%

3. SVM RBF

- Parámetros optimizados: C y gamma (coeficiente del kernel).
- Mejor configuración: C = 110, gamma = 0.0007.
- Exactitud (CV): 75.7%
-

5. Resultados

- **Kernel Poly**

```
Best parameters: {'svm__C': 13, 'svm__degree': 2, 'svm__kernel':
'poly'}
Best cross-validation score: 0.7348348348348348

Classification Report:
              precision    recall  f1-score   support

        Alto       0.78      0.82      0.80        17
        Bajo       0.83      0.45      0.59        11
       Medio       0.68      0.83      0.75        18

    accuracy                           0.74        46
   macro avg       0.76      0.70      0.71        46
weighted avg       0.75      0.74      0.73        46
```

```
Test set accuracy: 0.739
```

- **Kernel lineal**

```
Best parameters: {'svm__C': 7, 'svm__kernel': 'linear'}
Best cross-validation score: 0.7680180180180181

Classification Report:
              precision    recall  f1-score   support

        Alto       0.83      0.88      0.86        17
        Bajo       0.78      0.64      0.70        11
       Medio       0.68      0.72      0.70        18

    accuracy                           0.76        46
   macro avg       0.77      0.75      0.75        46
weighted avg       0.76      0.76      0.76        46

Test set accuracy: 0.761
```

- **Kernel RBF**

```
Best parameters: {'svm__C': 110, 'svm__gamma': 0.0007,
'svm__kernel': 'rbf'}
Best cross-validation score: 0.7566066066066066

Classification Report:
              precision    recall  f1-score   support

        Alto       0.83      0.88      0.86        17
        Bajo       0.89      0.73      0.80        11
       Medio       0.74      0.78      0.76        18

    accuracy                           0.80        46
   macro avg       0.82      0.80      0.80        46
weighted avg       0.81      0.80      0.80        46

Test set accuracy: 0.804
```

Observaciones:

- `cross-validation score`: El SVM Lineal obtuvo el mejor `desempeño` (76.8%), seguido por RBF (75.7%) y Polinomial (73.5%).
- `Test set accuracy`: El SVM RBF obtuvo el mejor desempeño (80.4%), seguido del SVM Polinomial (76.1%) y el lineal (73.9%)

- El SVM RBF obtuvo los mejores resultados de precision, recall y f1-score, para las categorías Alto, Bajo y Medio.

6. Experimento adicional

Se probó cambiar `StandardScaler` por `MinMaxScaler` y se vió que el accuracy subió a 82.6 (de RBF). MinMaxScaler normaliza el rango del atributo a una escala de 0 a 1.

```
Best parameters: {'svm__C': 100, 'svm__gamma': 0.01,
'svm__kernel': 'rbf'}
Best cross-validation score: 0.7510510510510511

Classification Report:
              precision    recall  f1-score   support

        Alto       0.88      0.88      0.88        17
        Bajo       0.89      0.73      0.80        11
       Medio       0.75      0.83      0.79        18

    accuracy                           0.83        46
   macro avg       0.84      0.81      0.82        46
weighted avg       0.83      0.83      0.83        46

Test set accuracy: 0.826
```

7. Conclusiones

Este proyecto exploró la capacidad de un modelo de machine learning, específicamente el algoritmo Support Vector Machine (SVM), para predecir la categoría del PIB de un país basándose en características como el área, población o el porcentaje de número de usuarios de internet. **El mejor modelo resultó ser el SVM con kernel RBF, con parámetros C=100 y gamma=0.0007, utilizando una normalización de los atributos (MinMaxScaler)** para clasificar países por nivel de ingreso Bajo, Mediano o Alto, considerando la información de las columnas incluidas en la tabla de países.

Podría ser interesante incluir más características de un país, por ejemplo, el promedio de años escolares, los porcentajes de vacunación, la producción de patentes y artículos científicos, etc.

También sería interesante utilizar un muestreo estratificado, ya que no hay la misma cantidad de ejemplos de cada clase.