

Tarea 6 – Introducción a Data Science

Support Vector Machines y Naïve Bayes

Profesor: Brian Keith Norambuena, brian.keith@ucn.cl.

Ayudante: Matías Salas Villanueva, matias.salas@alumnos.ucn.cl.

Objetivos

1. Conocer las ventajas y desventajas de Support Vector Machine y Naïve Bayes.
2. Seleccionar modelos de clasificación en base a métricas de evaluación.

Entregables

- Un cuaderno interactivo *Jupyter Notebook* con todo el código fuente, resultados, y análisis.
- El cuaderno interactivo debe utilizar apropiadamente celdas de *markdown* y de código.

Fecha de Entrega

Viernes 14 de noviembre a las 23.59 por Campus Virtual con el formato indicado en la sección de Consideraciones de Evaluación y Administrativas.

Requerimientos Mínimos

Se describen a continuación los requerimientos mínimos para esta tarea.

1. Análisis Exploratorio [2 puntos]:

- Visualice la correlación absoluta de cada característica con la variable objetivo (Class) en un gráfico de barras. **(+1 punto)**
- Visualice la distribución para las 5 características con más varianza en cada clase. **(+1 punto)**

2. Support Vector Machine [2 puntos]:

- Utilizando una separación del 70%/30% del conjunto de datos en entrenamiento y prueba respectivamente, utilice el modelo [SVC](#) para ajustar un modelo de Support Vector Machine. **(+1 punto)**
- Optimice y reporte los siguientes hiperparámetros; kernel: ('linear', 'rbf', 'poly') y C: (0.1, 1, 10). **(+1 punto)**

3. Naïve Bayes [2 puntos]:

- A partir del conjunto de datos obtenido, analice y confirme si existe independencia. **(+1 punto)**
- Utilizando una separación del 70%/30% del conjunto de datos en entrenamiento y prueba respectivamente, utilice el modelo [GaussianNB](#) para ajustar un modelo de Naïve Bayes. **(+1 punto)**

4. Selección de Modelo [1 punto]:

- Genere una tabla en donde compare: modelo, F1-score, recall y precision (clase 1). **(+0.5 puntos)**
- En base a la métrica recall, seleccione el mejor modelo y fundamente su decisión. **(+0.5 puntos)**

Descripción del conjunto de datos

Se debe hacer uso del conjunto de datos proporcionado *creditcard_Tarea6.csv* para la ejecución de esta tarea, a continuación, una breve descripción de cada columna:

- **id**: Identificador único para cada transacción.
- **V1-V28**: Características anónimas que representan varios atributos de la transacción (por ejemplo, tiempo, ubicación, etc.), ya estandarizadas.
- **Amount**: El monto de la transacción, sin estandarizar.
- **Class**: Etiqueta binaria que indica si la transacción es fraudulenta (1) o no (0).

Consideraciones de Evaluación y Administrativas

- El **formato de entrega** será el siguiente “NombreApellido1_NombreApellido2.ipynb” (sin tildes) mediante plataforma de campus virtual.
Ejemplo: **PedroRojas_JosePerez.ipynb**
- A partir de las 00:00 hrs del día siguiente del día de entrega comenzará un descuento de 10 décimas **por cada hora** de retraso, Ejemplo: Si el límite de entrega es el 29 de agosto a las 23:59 hrs y la tarea es entregada el 30 de agosto a las 01:00 hrs, tiene derecho a nota máxima 6.0.
- La escala de evaluación será al 60%.
- Para recibir el puntaje completo debe asegurarse de que el cuaderno pueda ser ejecutado de corrido, de lo contrario se aplicará la **mitad del puntaje**.
- Los análisis de resultados y correspondientes conclusiones serán evaluados según la claridad de sus ideas y su presentación (escritura y formato).
- **Si utiliza Inteligencia Artificial generativa** (e.g., ChatGPT) durante su trabajo, por favor incluya todas sus *prompts* al final del cuaderno interactivo como documentación o entregue un anexo con esta información. No habrá penalización por el uso de estas herramientas como apoyo, pero no deberían ser utilizadas para resolver la tarea por ustedes.
- Los Créditos Extras serán puntos extras que se sumarán al final del semestre a tareas que no tengan el puntaje completo, comenzando por la tarea con menor puntaje. Los créditos extra sobrantes se anularán si el estudiante tiene puntaje máximo en todas las tareas.