ESCUELA DE POSGRADO

MAESTRIA EN INFORMÁTICA

PUCP

Proyecto de curso

Curso: Aprendizaje automático: Teoría y aplicaciones (INF648)
Docentes: Dr. César Beltrán Castañón
Mag. César Olivares Poggi

El proyecto de curso permitirá que los estudiantes de postgrado apliquen su experiencia en el diseño, implementación y evaluación de algoritmos de aprendizaje de máquina. Para su elaboración se conformará grupos de hasta tres alumnos cada uno. El objetivo principal de la actividad será la aplicación de mejores prácticas de investigación en aprendizaje automático en una tarea de clasificación. Se requerirá el uso de al menos 4 métodos de clasificación o regresión de su elección y evaluar su rendimiento en una tarea común, así como establecer una comparación con una línea base, de preferencia reportada previamente en la literatura.

El proyecto tendrá un peso del 40% de la nota final y constará de los siguientes entregables:

Entregable	**Fecha límite de entrega**	**Puntaje**
1. Propuesta de proyecto	20 de mayo	2 puntos
2. Primera parte del informe escrito (Introducción, estado del arte y diseño del experimento), en formato IEEE, tamaño A4, máximo 3 páginas de extensión	3 de junio	5 puntos
3. Código y/o scripts (Jupyter Notebooks) con la experimentación realizada	17 de junio	6 puntos
4. Informe final	24 de junio	7 puntos

Propuesta de proyecto

La propuesta de proyecto deberá incluir lo siguiente (máximo una hoja A4):

- Nombre del proyecto
- Miembros del equipo
- Conjunto de datos a utilizar
- Objetivo del proyecto
- Artículos científicos relevantes
    - Al menos 5 artículos que hayan utilizado este conjunto de datos anteriormente o directamente relevantes por el tipo de problema a abordar. Escoger los más relevantes para el proyecto.
- Propuesta tentativa de modelos de clasificación a utilizar
- Declaración de trabajo grupal (según formato de la *Directiva y normas para la elaboración de trabajos grupales*)

Conjuntos de datos

Cada equipo deberá elegir uno de los siguientes 6 conjuntos de datos. Cada conjunto de datos podrá ser elegido por un máximo de 2 equipos. En caso el grupo desee elegir algún otro conjunto de datos, éste deberá ser autorizado por el profesor a cargo a más tardar 3 días antes de la fecha de entrega de la propuesta de proyecto.

1. **[Clasificación] Insurance Company Benchmark (COIL 2000) Data Set** - This data set used in the CoIL 2000 Challenge contains information on customers of an insurance

company. The data consists of 86 variables and includes product usage data and socio-demographic data.
https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+%28COIL+2000%29

2. **[Clasificación] Myocardial infarction complications Data Set** - Prediction of myocardial infarction complications.
https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications

3. **[Clasificación] Productivity Prediction of Garment Employees Data Set**- This dataset includes important attributes of the garment manufacturing process and the productivity of the employees which had been collected manually and also been validated by the industry experts.
https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees

4. **[Clasificación] Online Shoppers Purchasing Intention Data Set** - Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

5. **[Regresión] Online News Popularity -** This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).
https://archive.ics.uci.edu/dataset/332/online+news+popularity

6. **[Regresión] Superconductivty Data -** Two files contain data on 21263 superconductors and their relevant features. The goal here is to predict the critical temperature based on the features extracted.
https://archive.ics.uci.edu/dataset/464/superconductivty+data

7. **[Regresión] SGEMM GPU kernel performance –** Estimate running times for multiplying two 2048 x 2048 matrices using a GPU OpenCL SGEMM kernel with varying parameters (using the library 'CLTune').
https://archive.ics.uci.edu/dataset/440/sgemm+gpu+kernel+performance

8. **[Regresión] Steel Industry Energy Consumption -** The data is collected from a smart small-scale steel industry in South Korea. Goal is to predict energy consumption.
https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption

Informe del proyecto

El informe del proyecto deberá explicar claramente el objetivo del estudio, trabajos previos sobre el problema, dificultades en la tarea de clasificación o regresión, elección de los algoritmos, estrategias de combinación de modelos, estimación de la tasa de error, etc.

El informe debe ser presentado en formato IEEE, tamaño A4, máximo 8 páginas de extensión (ver plantillas abajo en la sección Recursos).

El informe debe incluir la siguiente información:

- **Introducción**
  - Presentación del problema general sobre el que versará el trabajo y cómo se integra dentro del campo del aprendizaje automático
  - Objetivo del estudio
  - Organización del informe (secciones)
- **Estado del arte**

  - Breve síntesis del aporte que otros artículos científicos han realizado para este problema, aunque no necesariamente con el conjunto de datos seleccionado.
  - Breve síntesis del aporte que al menos 2 artículos científicos que han usado el conjunto de datos seleccionado, así como de los resultados obtenidos en ellos.
- **Diseño del experimento**
  - Descripción del conjunto de datos
    - Número y tipo de características (binarias, discretas, continuas, etc.).
    - Número de muestras en los conjuntos de entrenamiento y prueba. En caso aplique, número de muestras por clase.
    - Estadística descriptiva y visualización de los datos
  - Metodología
    - De ser el caso, estrategia para el manejo de datos faltantes.
    - Selección y extracción de características.
    - Selección y justificación de la medida de calidad.
    - Algoritmos que serán empleados y estrategia para su ajuste.
    - Estrategia de validación a emplear para el ajuste de hiperparámetros.
- **Experimentación y resultados**
  - Línea base: Reproducción de resultados reportados en un artículo científico anterior.
  - Evaluación del rendimiento de los modelos ensayados.
  - Comparación de línea base y resultados propios.
- **Discusión**
  - Interpretación de los resultados obtenidos.
  - Identificación y visualización de ejemplos en los que tienen dificultad los modelos ensayados. ¿A qué se podría atribuir?
  - ¿Cómo podría ser mejorado su sistema?
- **Conclusiones y trabajos futuros**

Código y/o scripts (GitHub y Google Colab)

- El código será trabajado colaborativamente en GitHub, de manera que se pueda verificar los aportes hechos por cada uno de los integrantes del curso.
- El lenguaje a utilizar será Python. Si el equipo desea trabajar en un lenguaje de programación diferente, consultarlo previamente con el profesor.
- Como entorno se puede usar Google Colab, Azure ML, Databricks u otras plataformas semejantes.
- Se recomienda utilizar MLFlow u otra herramienta semejante para registrar la experimentación y guardar los modelos entrenados.
- Opcional: utilizar librerías para optimización de parámetros de entrenamiento, tales como Optuna o Hyperopt.
- Se deberá asignar nombres representativos a los archivos, de manera que se pueda identificar su orden relativo y el propósito de cada uno.
- No hay restricciones para tomar *como base* código tomado de otras fuentes, siempre y cuando *se cite debidamente* la fuente y se realice las *adaptaciones* que requiera el propio trabajo.
- El código deberá estar mínimamente comentado, siempre en español. Se ignorará cualquier comentario en otro idioma.
- Asimismo, se ignorará cualquier código simplemente copiado cuya fuente no haya sido citada, y se asignará el puntaje mínimo al grupo en el entregable 3.

Recursos

- Directiva y normas para la elaboración de trabajos grupales
  http://files.pucp.edu.pe/homepucp/uploads/2018/02/07084342/DIRECTIVA-Y-NORMAS-PARA-LA-ELABORACION-DE-TRABAJOS-GRUPALES.pdf

- Buscador de literatura académica
  https://scholar.google.com.pe/

- Acceso a bases de datos electrónicas – PUCP
  https://biblioteca.pucp.edu.pe/recursos-en-linea/bases-de-datos

- GitHub Student Pack (inscribirse usando direcciones @pucp.**edu**.pe)
  https://education.github.com/pack

- Git y Github | Curso Práctico de Git y Github Desde Cero
  https://www.youtube.com/watch?v=HiXLkL42tMU

- Formato IEEE (MS Word y LaTeX)
  https://www.ieee.org/conferences_events/conferences/publishing/templates.html

- Editor colaborativo LaTeX en línea
  https://www.overleaf.com/