

Modelos y Simulación II: Proyecto del Curso

Prof. Julián David Arias Londoño
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Colombia
julian.arias1@udea.edu.co

Abstract—Las actividades propuestas para el desarrollo del proyecto buscan que cada uno de los grupos de estudiantes presenten todo el diseño, análisis y simulación de un sistema de predicción basado en técnicas de aprendizaje de máquina; describiendo el problema y su contexto en términos del estado del arte, especificando cada una de las etapas del desarrollo del trabajo, los modelos con sus respectivas restricciones, la metodología de validación, los resultados de las simulaciones y las conclusiones obtenidas.

Index Terms—Modelos controlados por datos, aprendizaje de automático.

I. EJERCICIOS

A. Parte I: Comprensión del problema de ML

- 1) Describa de manera clara el problema de predicción que está abordando, su campo de aplicación y explique si corresponde a un problema de clasificación o de regresión.
- 2) Enumere las variables incluidas como entrada al sistema y la o las variables a predecir. Explique claramente el tipo de codificación de cada variable y, si la base de datos cuenta con **valores faltantes**, explique cómo se llenaron los vacíos en cada caso.
- 3) Realice una búsqueda de al menos 4 artículos que hayan abordado el mismo problema de aprendizaje que Uds están trabajando. Incluya, en la medida de lo posible, trabajos que hayan empleado la misma base de datos. Describa brevemente:
 - ¿Qué técnica(s) de aprendizaje usan en los artículos?
 - ¿Qué metodología de validación usaron?
 - ¿Cuáles fueron los resultados obtenidos en cada uno de los trabajos citados?

Se recomienda buscar trabajos en las bases de datos: www.sciencedirect.com y www.ieeexplore.org. También se pueden buscar trabajos en la base de datos <http://link.springer.com>, pero se debe tener en cuenta que el acceso que tiene la Universidad es mucho más limitado para dicha base de datos. Incluir preferiblemente artículos publicados en revista no en congresos o conferencias. Se recomienda utilizar el buscador <https://scholar.google.com> para encontrar artículos que hayan citado la base de datos seleccionada. **No utilice más de una página del informe para esta descripción.**

B. Parte II: Entrenamiento y Evaluación de los Modelos

- 4) Incluya una sesión dentro de su informe con el nombre *Experimentos*, en la cual describa la metodología de

validación usada y la base de datos que está usando para llevar a cabo el proyecto, incluyendo la fuente de la base de datos como referencia, el número de muestras, variables, la distribución de muestras por clase, etc. Si la base de datos está desbalanceada, deben considerar el uso de técnicas de submuestreo y sobremuestreo inteligente, además de usar estrategia de validación apropiada.

En esta sección deben especificar las medidas de desempeño que usarán para evaluar el sistema, indicando la medida principal. En caso de que utilicen medidas diferentes a las vistas en clase, deben incluir las definiciones correspondientes que permitan al lector comprender la medida.

- 5) Si su problema es de clasificación debe evaluar, como mínimo, el uso de los siguientes modelos de predicción y documentar los resultados de las simulaciones realizadas:
 - Análisis discriminante Cuadrático
 - Ventana de Parzen (Método kernel)
 - Gradient Boosting Tree
 - Redes Neuronales Artificiales
 - Máquinas de Soporte Vectorial
- 6) Si su problema es de regresión debe evaluar, como mínimo, el uso de los siguientes modelos de predicción y documentar los resultados de las simulaciones realizadas:
 - Regresión múltiple
 - Ventana de Parzen
 - Random Forest
 - Redes Neuronales Artificiales
 - Regresión por Vectores de Soporte con kernel RBF.

Para cada modelo **se deben mostrar tablas** que indiquen el resultado obtenido para los diferentes parámetros evaluados durante la fase de validación. Las tablas deben mostrar las medidas de desempeño seleccionadas y listadas en la sección *Experimentos*, así como su correspondiente intervalo de confianza.


Finalmente, se debe incluir una tabla adicional con los resultados en el conjunto de test, de los modelos seleccionados como mejores (el mejor conjunto de hiperparámetros) para cada uno de los tipos de modelos evaluados.

C. Parte III: Selección/Extracción de Características

- 7) Realice un análisis individual de cada una de las características, a partir de medidas de correlación y del índice de Fisher (según sea el caso). Identifique de acuerdo con este análisis, ¿cuáles son las características candidatas a ser eliminadas?
- 8) Realice selección de características por el método de búsqueda secuencial ascendente o descendente y evalúe nuevamente en los 3 mejores modelos evaluados. Para realizar este punto cada grupo debe decidir la función criterio a usar en el algoritmo de selección, justificando su decisión. Incluya una tabla con los resultados, indicando el porcentaje de reducción alcanzado y las demás medidas de desempeño. **Recuerde incluir en el informe cuál fue el criterio de selección usado y porqué.**
- 9) Realice extracción de características por el método PCA y evalúe nuevamente en los 3 mejores modelos de predicción evaluados. Cada grupo debe decidir el criterio para seleccionar el número de componentes justificando su decisión. Incluya una tabla con los resultados, indicando el porcentaje de reducción alcanzado y las demás medidas de desempeño.
- 10) Incluya una sección final de discusión en la cuál analice los resultados obtenidos y los compare con los resultados de los artículos consultados en el punto 2.

Nota. El informe se debe presentar en formato IEEE (Similar al usado en esta guía), las plantillas IEEE para Word o \LaTeX pueden ser descargadas del siguiente enlace. Procure realizar una redacción coherente y haga un buen uso de la referenciación (consulte las normas IEEE para incluir las citas en la bibliografía). El informe puede tener máximo una longitud de **10 páginas**. Cada punto del informe tiene un valor diferente así:

Punto	Valor
1,2,3 y 4	1
5 ó 6	1
7	0.5
8	0.7
9	0.8
10	0.5
Sustentación	0.5
Total	5.0

Los scripts y/o notebooks desarrollados para el proyecto deben ser subidos a un repositorio  y el enlace incluido en el informe. El repositorio debe incluir un archivo README que explique claramente cómo deben ser ejecutados los scripts para reproducir los resultados.

Se darán puntos adicionales si el problema abordado corresponde a un paradigma de múltiples salidas o de múltiples instancias. También se darán puntos adicionales si el grupo desarrolla una aplicación web que implemente toda el pipeline del procesamiento de información y consuma el modelo. La idea es poder entregarle al modelo una muestra (real o artificial) y obtener una predicción.

El grupo que no presente sustentación no tendrá nota. Para la sustentación cada grupo debe preparar un video de máximo 15 minutos, subirlo a Drive e incluir el enlace en el informe. En el video deben presentar el contexto del problema, describir la base de datos, el diseño experimental y los mejores resultados encontrados en las simulaciones. Si realizaron el proceso de despliegue, pueden mostrar su funcionamiento. Es importante que los miembros del equipo discutan los hallazgos, comparen sus resultados entre los abordajes con y sin reducción de dimensión, además de compararlos con los consultados en el estado del arte y finalmente, den conclusiones sobre el alcance de los resultados obtenidos en el proyecto. **La nota de la sustentación es individual para cada uno de los miembros del equipo, de acuerdo con sus aportes a la claridad de la exposición y a la calidad de la discusión presentada.**