UNIDAD DE INGENIERÍA EN COMPUTACIÓN
Centro Central
Inteligencia Artificial
Profesor: Kenneth Obando Rodríguez

TEC | Tecnológico de Costa Rica

Trabajo Corto (%) Fecha de Entrega: 12 de mayo, 2025

Proyecto de Machine Learning

1 Introducción

Este trabajo tiene como objetivo poner en práctica el uso de herramientas computacionales para la ingesta, limpieza, transformación y exportación de los datos. El software sugerido es Python (Cuaderno Jupyter o Google Colab). En todos los casos se deben aplicar las mejores prácticas estudiadas en el curso de Inteligencia Artificial.

Se debe aplicar la metodología CRISP-DM para resolver el caso de minería de datos. Esta metodología se autodescribe como *"un modelo de analítica de datos organizado como un proceso jerárquico*. Producto de este proceso, se obtiene una documentación ordenada, con claridad en los hallazgos obtenidos y con un orden lógico del proceso".

Los datos de la actividad deben ser propuestos por los estudiantes a partir de bases de datos públicas que se pueden encontrar en plataformas como Kaggle o del Banco Mundial.

2 Descripción de la Actividad

1. Fase de entendimiento del negocio
   (a) Determinar los objetivos del negocio
      i. Objetivos del negocio
      ii. Criterios de éxito (en términos del negocio)
   (b) Determinar los objetivos de minería de datos
      i. Objetivos de minería de datos
      ii. Criterios de éxito (desde la perspectiva de minería de datos)
2. Fase de entendimiento de los datos:
   (a) Recopilación inicial de datos
      i. Lista de fuentes de datos requeridos
      ii. Método de acceso (para cada fuente de datos indicar si se obtiene de Internet, de un sistema interno, etc.)
      iii. Descripción de los datos (describa las principales variables a utilizar, indique también cantidad de registros y variables)

iv. Exploración de los datos (al menos 3 variables importantes, puede incluir estadísticos y análisis de la distribución de los datos, visualizaciones, etc). Utilice su creatividad para obtener información valiosa de los datos con los que está trabajando.

v. Calidad de datos (debe constar vía código la revisión de aspectos de calidad de datos para todas las variables a utilizar)

3. Fase de preparación de los datos:

   El objetivo de esta fase es preparar los datos en vista al entrenamiento de los modelos de estimación. Esta sección debe describir las etapas utilizadas en Python. Puede presentar visualizaciones y tablas que demuestren el resultado de los procesamientos.

   (a) Selección de los datos

   (b) Limpieza de los datos

   (c) Construcción de nuevos datos (atributos) (Opcional y dependiendo de los datos elegidos)

   (d) Transformaciones aplicadas a los datos

4. Fase de modelado: Esta fase describe las etapas del entrenamiento y evaluación. Se debe entrenar cuatro modelos (algoritmos), puede ser de regresión o clasificación.

   Debe constar los diferentes experimentos con sus respectivos parámetros con el fin de seleccionar el mejor modelo con búsqueda de hyperparámetros, (puede usar herramientas específicas para ello como GridSearchCV, Optuna o MLFlow) . Para ello, no olvide separar los datos en *training* y *testing* para la evaluzación de los mismos, y aplicar CrossValidation como técnica para la selección tomando en cuenta que el modelo escogido tenga una relación adecuada entre *Bias* y *Variance*.

   Además, se debe constatar que los modelos no están realizando un sobreajuste.

5. Criterio de Selección y Conclusiones: En esta sección debe analizar porqué seleccionó el modelo y qué nueva información agrega a los datos (variables más importantes, etc).

En los archivos puede encontrar el documento ”crisp-dm.pdf” como documento de referencia sobre la metodología CRISP-DM. Como podrá notar, la metodología incluye más secciones no incorporadas en la especificación del proyecto.

3 Referencias

- CRISP-DM: `https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-54bad20bbc7938991bf34f86dde0babfbd2d5a72`
- IBM-CRISP-DM: `https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf`

4 Evaluación

La evaluación del proyecto se realizará en base a los siguientes criterios, con un total de 100 puntos posibles. Cada criterio tiene asignado un rango de puntos según el nivel de cumplimiento observado en el trabajo del estudiante.

4.1 Entendimiento del Negocio (10 puntos)

- **Objetivos del negocio claramente definidos** (2.5 puntos): Se evalúa si los objetivos del negocio están claramente identificados y descritos.
- **Criterios de éxito del negocio** (2.5 puntos): Se evalúa si se han establecido y descrito claramente los criterios de éxito del proyecto en términos del negocio.
- **Objetivos de minería de datos** (2.5 puntos): Se evalúa la claridad y la alineación de los objetivos de minería de datos con los objetivos del negocio.
- **Criterios de éxito de la minería de datos** (2.5 puntos): Se evalúa si los criterios de éxito desde la perspectiva de la minería de datos están bien definidos y son medibles.

4.2 Entendimiento de los Datos (20 puntos)

- **Recopilación y descripción de datos** (10 puntos): Se evalúa la completitud y la precisión en la recopilación y descripción de los datos.
- **Exploración y calidad de datos** (10 puntos): Se evalúa la profundidad de la exploración de datos realizada y la adecuada revisión de la calidad de los datos.

4.3 Preparación de los Datos (20 puntos)

- **Limpieza y selección de datos** (10 puntos): Se evalúa la efectividad de los procesos de limpieza y selección de los datos.
- **Construcción de atributos y transformaciones** (10 puntos): Se evalúa la creatividad y la adecuación de los nuevos atributos construidos y las transformaciones aplicadas a los datos.

4.4 Modelado (30 puntos)

- **Selección de modelo y experimentación** (20 puntos): Se evalúa la justificación de la selección del modelo y la calidad de los experimentos realizados. Debe probar con al menos 4 distintos algoritmos.
- **Evaluación y validación del modelo** (10 puntos): Se evalúa la adecuación de las técnicas de evaluación y validación aplicadas, incluyendo la utilización de CrossValidation.

4.5 Criterio de Selección y Conclusiones (20 puntos)

- **Análisis y justificación del modelo seleccionado** (10 puntos): Se evalúa la profundidad del análisis y la solidez de la justificación para la selección del modelo.

- **Aporte de nueva información** (10 puntos): Se evalúa la capacidad del proyecto para generar nueva información relevante a partir de los datos, incluyendo el análisis de las variables más importantes.