Tarea de Investigación Final – Minería de Datos I

UNIVERSIDAD NACIONAL
SECCIÓN REGIONAL HUETAR NORTE Y CARIBE
ESCUELA DE INFORMÁTICA
BACHILLERATO EN INGENIERÍA EN SISTEMAS DE INFORMACIÓN (BA-INFORM)

Objetivo

Realizar un análisis exhaustivo de un dataset de elección libre (clasificación o regresión) del UCI Machine Learning Repository o Kaggle, aplicando el ciclo completo de Minería de Datos, desde la comprensión del negocio hasta la evaluación de modelos predictivos y la comunicación de resultados.

Dataset: De libre elección (Clasificación o Regresión)

El estudiante deberá seleccionar un dataset libre de UCI ML Repository o Kaggle que presente un problema de clasificación o regresión de complejidad media.

- Sugerencia: Dataset de House Prices (Regresión) o Credit Card Fraud Detection (Clasificación Desbalanceada).

Actividades a desarrollar en este trabajo de investigación:

N.º	Tema Aplicar	Contenido	Gráficos Sugeridos
1.	Comprensión del Negocio (Problema)	Definir claramente el objetivo del análisis (la pregunta de negocio o la predicción a lograr), el contexto del dataset (fuente, año, industria) y la variable objetivo (qué se busca predecir).	*Ninguno*
2.	Comprensión de los Datos y Análisis Exploratorio de Datos (EDA)	Descripción de las variables (tipo, rango, valores únicos). Análisis de la calidad (valores faltantes, *outliers*). Distribución de la variable objetivo y las características principales.	Histogramas (Distribución de variables numéricas); Gráficos de Barras (Distribución de categóricas); Boxplots (*Outliers*); Mapa de Calor (Correlación entre variables).
3.	Preprocesamiento y Transformación de Datos	Detallar las técnicas de limpieza (manejo de faltantes, *outliers*), codificación de variables categóricas (ej. One-Hot, Label Encoding) y escalamiento/normalización de variables numéricas. Mencionar la división Train/Test.	*Código y Tablas de Transformación*
4.	Diseño y Desarrollo del Modelo Predictivo	Selección e Implementación de al menos dos (2) algoritmos de Machine Learning (ej. Regresión Lineal/Logística, Árboles de Decisión, Random Forest, SVM). Justificación de la elección de los algoritmos(Consenso).	*Código de Modelado*
5.	Evaluación y Validación del Modelo	Presentar las métricas de rendimiento pertinentes al problema (Clasificación: Accuracy, F1-Score, AUC ROC; Regresión: MAE, MSE,). Aplicar Validación Cruzada (K-Fold CV) para medir la robustez y evitar el sobreajuste.	Curva ROC (Clasificación); Gráfico de Dispersión (Predicción vs. Real - Regresión); Gráficos de Residuos (Regresión).
6.	Comunicación de Resultados (Conclusión)	Resumen de los hallazgos clave. Interpretación de la mejor métrica. Conclusión sobre si se logró el objetivo de negocio. Recomendaciones de mejora (ej. Feature Engineering, *Hyperparameter Tuning*).	Tabla Comparativa de Métricas de los modelos; Gráfico de Importancia de Características.

Evaluación

N.º	Criterio de Evaluación	Descripción	%
1.	**Comprensión del Negocio (Problema)**	Claridad y precisión en la definición del objetivo, contexto y la variable a predecir.	**10%**
2.	**Comprensión de los Datos y EDA**	Calidad del análisis exploratorio, identificación de problemas de datos y uso de visualizaciones descriptivas apropiadas.	**20%**
3.	**Preprocesamiento y Transformación**	Correcta aplicación de técnicas de limpieza, codificación y escalamiento. Justificación de los métodos usados.	**20%**
4.	**Diseño y Desarrollo del Modelo**	Selección de algoritmos pertinentes y correcta implementación de al menos dos modelos.	**15%**
5.	**Evaluación y Validación del Modelo**	Uso correcto de métricas de rendimiento y aplicación de Validación Cruzada (K-Fold CV). Análisis de *overfitting*.	**20%**
6.	**Comunicación de Resultados (Conclusión)**	Resumen claro, interpretación de la mejor métrica, cumplimiento del objetivo y presentación de recomendaciones.	**15%**
Total			**100%**