Title: Kawasaki Disease Midterm Report

By: Alyshondria Hicks

For: Predictive Modeling & Analytics 566

Date: F 10/17/25

Assignment: Midterm Report (Assignment 3)

**Project Description**

Kawasaki Disease (KD), also known as mucocutaneous lymph node syndrome, is an acute, self-limited inflammatory disorder that primarily affects children under the age of five, of which the cause for it is unknown. Characterized by persistent fever and systemic vasculitis, KD can result in serious cardiovascular complications such as coronary artery aneurysms if not promptly treated. While diagnosis is largely clinical, predicting the severity and progression of the disease remains a major challenge due to variability in patient response and the lack of specific biomarkers. Machine learning (ML) techniques provide a powerful framework for modeling complex clinical and laboratory data to predict patient outcomes and disease trajectories. This project aims to apply ML methods to support the diagnosis of clinical outcomes in Kawasaki Disease—such as the risk of coronary artery involvement or treatment resistance—thereby enabling more personalized and timely medical interventions.

**Data Description**

This dataset, obtained from the Kaggle repository, comprises 2,752 patient records and 15 distinct features. The features are categorized into four primary domains:

* **Demographic Information** — *Patient ID, Date of Diagnosis, Age at Diagnosis, Gender, Ethnicity, and Location*
* **Clinical Characteristics** — *Fever Duration, Symptoms, Laboratory Test Results, and Echocardiography Findings*
* **Treatment Variables** — *Treatment Approach*
* **Outcome and Prognostic Indicators** — *Clinical Outcome, Complications, Follow-up Visits, and Long-Term Effects*

This structured dataset enables the exploration of demographic, clinical, and treatment-related factors in relation to patient outcomes and prognostic trends.

**Methods and Analysis**

A series of ML classification algorithms were implemented to model the clinical outcomes of KD. The classifiers included K-Nearest Neighbors (KNN), Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machines (SVM) with both linear and non-linear kernels. Prior to model training, feature selection was performed to remove less informative variables, and the dataset was partitioned into training (80%) and testing (20%) subsets. Initial evaluations indicated that the KNN classifier achieved the highest performance, with an overall accuracy of approximately 40%. To further assess model robustness and generalizability, a stratified k-fold cross-validation approach was applied. Under this validation framework, the Decision Tree classifier demonstrated the most consistent performance, achieving nearly 40% across accuracy, precision, recall, and F1-score metrics.

**Evaluation**

The overall low accuracy across all models suggests that the dataset’s clinical and demographic features may lack strong discriminative power for predicting KD outcomes or that additional feature engineering and data balancing are required. Future work should focus on integrating higher-dimensional data and exploring advanced algorithms, including ensemble learning and deep neural networks, to improve predictive accuracy and clinical relevance.