CS 491/591
Introduction to Machine Learning

Assignment 2

Due by 11:59PM, Friday, **December 1**

NOTE: Assignment is to be done in your assigned teams. Do not look at another team's answer and do not allow another team to look at your answer. Doing so will result in charges of academic misconduct. Email your submission to me (atkison@cs.ua.edu). Only one submission is required per team.

Cardiovascular diseases are a global health crisis, claiming approximately 17.9 million lives each year, which represents 31% of all deaths worldwide. A primary outcome of these diseases is heart failure, a condition that our lab focuses on predicting and understanding through a dataset of 12 key indicators known to influence mortality rates.

The majority of heart diseases can be averted by addressing lifestyle-related risk factors, such as smoking, poor diet, physical inactivity, and excessive alcohol use, with comprehensive public health strategies. Moreover, for individuals who already have or are at high risk of cardiovascular diseases—due to hypertension, diabetes, hyperlipidemia, or other established conditions—early detection and management are crucial.

In this context, machine learning models emerge as powerful tools for early intervention, capable of analyzing complex medical data to identify those at high risk. In lab 2, you will engage with medical records from patients who have suffered heart failure, employing this very technology. The dataset you will work with includes a range of predictors related to cardiovascular health risks and outcomes. Through this analysis, you'll learn how predictive models can assist in the early detection and management of heart disease, potentially saving lives.

Objective:

To apply and compare three machine learning models — Random Forest Classifier, K-Nearest Neighbor Classifier, and Decision Tree Classifier — to predict the mortality outcome of heart failure from clinical data.

Dataset Description:

The dataset you'll be exploring in this lab contains comprehensive clinical records from 299 patients. It encompasses several key health indicators that have been recorded to assess the severity and outcome of heart failure in each patient. Here's a detailed breakdown of what's included in the dataset:

- **Age**: The age of the patient in years, which is a continuous variable and can be a significant factor in heart health.
- **Anaemia**: A binary variable indicating the absence (0) or presence (1) of anaemia in the patient.
- **Creatinine Phosphokinase (CPK)**: The level of the CPK enzyme in the blood (mcg/L), a marker that can indicate stress on the heart muscle.
- **Diabetes**: A binary variable indicating whether the patient has diabetes (1) or not (0).
- **Ejection Fraction**: The percentage of blood leaving the heart at each contraction, indicating heart function.
- **High Blood Pressure**: A binary variable that flags if the patient has high blood pressure (1) or not (0).
- **Platelets**: The number of platelets in the blood (kiloplatelets/mL), which are critical for blood clotting.
- **Serum Creatinine**: The level of serum creatinine in the blood (mg/dL), indicative of kidney function, which is related to cardiovascular health.
- **Serum Sodium**: The level of sodium in the blood (mEq/L), with imbalances potentially linked to heart failure.
- **Sex**: The biological sex of the patient (male: 1, female: 0).
- **Smoking**: A binary variable that indicates whether the patient smokes (1) or does not smoke (0).
- **Time**: The follow-up period (days) during which the patient's health was monitored.
- **Death Event**: The target variable, which indicates whether the patient deceased (1) or survived (0) during the follow-up period.

Lab Tasks

1. **Dataset Familiarization:**
   - Examine the dataset to understand the features and the target variable.
   - Generate descriptive statistics to gain insights into the dataset.
2. **Data Preprocessing:**
   - Clean the dataset by addressing any missing or inconsistent data.
   - Perform necessary data transformations such as encoding categorical variables and normalizing numerical features.
3. **Model Implementation:**
   - Implement the **Random Forest Classifier**, **K-Nearest Neighbors Classifier**, and **Decision Tree Classifier**.
   - Ensure you understand the parameters being used and how they affect the model.
4. **Model Training:**
   - Split the dataset into training and testing sets to evaluate the performance of your models.
   - Train each model using the training set and make predictions on the testing set.
5. **Evaluation and Comparison:**
   - **Accuracy Assessment:**
     - Utilize the **accuracy_score** function to calculate the accuracy of the **Random Forest**, **K-Nearest Neighbors**, and **Decision Tree models**. This will provide a straightforward measure of each model's overall performance.
     - Discuss the relevance of accuracy in a clinical setting, considering the potential consequences of false positives and false negatives in predicting heart failure mortality.
   - **Confusion Matrix Analysis:**
     - Apply the **confusion_matrix** function to obtain the confusion matrices for each model. This will yield a more nuanced view of the models' performance, showing how many predictions fall into each category of true positives, true negatives, false positives, and false negatives.
     - Analyze the confusion matrices to assess the models' abilities to correctly identify positive (death event) and negative (no death event) cases. Emphasize the significance of each type of error in a medical context.
   - **Model Comparison:**
     - Compare the accuracy scores and confusion matrices side by side for each of the three models. This comparison will help determine which model is most effective for the task at hand.
     - Encourage discussion on why certain models may have performed better or worse, considering factors such as the models' complexity, the nature of the data, and the balance between precision and recall.
6. **Reporting:**
   - Compile your findings into a report that includes your preprocessing steps, model performance metrics, and your analysis of the results.
   - Discuss the potential real-world application of each model and how they might assist medical professionals in predicting patient outcomes.

7. Deliverables:

- A completed Google Colab notebook (.ipynb) file, comments, and analysis.
- A lab report summarizing your methodology, results (includes plots generated from the Google Colab), and personal reflections.