Revision Exercise #2

MA103 27 OC 2025
Major Patrick Kuiper

Provide a comment on what was done well, and two comments on what could be improved for each selectino

1. **Purpose**

> Clear and concise restatement of the problem asked of them in the prompt. Uses original phrasing. Establishes clear motivation and relevance of the report.
>
> Having safe drinking water in the field is important for both health and being ready for missions. Army field sanitation crews always keep an eye on water buffalos to make sure that Soldiers don't get sick from dirty water while they are training or on duty. The information includes chemical and biological measurements that go along with it, as well as several samples that were tested in a lab to see if the water was safe to drink. This project uses math to find out if a new water buffalo sample from the company is safe. The K-Nearest Neighbors (K-NN) method is used to group the predictor sample based on how similar it is to samples that have been tested before. This makes it clear how people in the field decide whether or not the water is safe.

---

2. **Data Set**

Clearly identifies source, variables, and identifies pertinent variables.

Given the following data from DMI: ID, Initial Height, Initial Weight, Pull-ups (Raw), BBT (Raw), SR (Raw), Crunch (Raw), CBT Initial Push-up, CBT Initial Sit-up, CBT Initial Run, CBT Initial Swim, and injury status (injured or not), the aim is to predict injuries based on baseline testing. For this project, use the CBT Initial Push-up and CBT Initial Run as pertinent variables to construct the graph along with the injury status represented as a binary variable, with '1' showing injury and '0' showing no injury. Lastly, the ID will be used to find each specific candidate. The ethical considerations of this data – Given the fact that DMI approved the release of this information would mean that the data accurately and ethically represents initial scores and injuries of new cadets during CBT. The data limitations are fact that only give a certain number of variables.

3. **Methodology**

Provides logical description of modeling decisions: clearly identifies which variables are explanatory and which is response, identifies assumptions, and articulates how k was selected.

This project utilizes the K-Nearest-Neighbor modelling method to determine whether the predictor record would pass using the data of proximal historical records. Using the provided data set and topics covered in class, it was decided that this was the best method to simultaneously assess two variables and their associated binary categorical result across multiple records. Points were color-coded blue for failure and red for passing, then plotted on a coordinate plane with standardized SDC on the x-axis and standardized APFT runtimesec on the y-axis (Appendix A.2, Table 1). Missing or faulty records were cleaned from the set as necessary for visual representation and were only removed with reasonable cause for justification (Appendix B.3.b, Table 3). All other records were considered feasible and fit for inclusion in the model. Parameters include k, which was determined using the square root of the sample size, and number of records, which totaled 97 after the removal of three unfit points. The K-Nearest Neighbors algorithm implements the Euclidean Distance Formula to calculate the proximity of the predictor record to each historical record. Scores were standardized to make comparisons more realistic between the two events, which occupy different time ranges for completion. This was done by calculating the z-score, which involves subtracting the mean from each data point and dividing the result by the standard deviation. All equations involved in standardization were calculated using Excel and can be found in Appendix B.4. Once calculated, the distance between the standardized scores of the predictor record and each historical record was plugged into the Euclidean Distance Formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (Eq.\,4)$$

where x_1 and x_2 are standardized SDC and APFT runtimesec for the predictor record and y_1 and y_2 are standardized SDC and APFT runtimesec for the i^{th} record. The distances are then ranked from lowest to highest, with the number of closest records used for prediction being determined by k = 10 (App.B.4 Eq.5). For the binary categorical variable Pass/Fail, the most commonly occurring result was chosen as the final prediction for the response variable.

4. **Discussion**

Provides thoughtful interpretation of solution, connecting it back to question asked. Thouroughly discusses constraints and limitations of analysis, including the sensitivity of the k-value.

These findings support the correlation between below-average times for the sprint drag carry and two-mile run and performance on the IOCT. Given the large quantity of passing records in the third quadrant, where negative outliers lie, and the placement of failing records solely in the first quadrant, indicating slower-than-average times, it can be concluded that SDC and APFT runtimesec are reliable indicators of the physical fitness required to pass the AFT. However, outliers in the data suggest that this model may not be perfectly indicative, and even those with above average (ie. Slower) SDC and APFT runtimesec times may be able to pass the IOCT. For instance, one individual with an average 2-mile run and a sprint drag carry z-score of positive 5 still managed to pass the test (Appendix A.2, Figure 1). Furthermore, 37 out of the passing population of 97 records had above-average run times, and 31 had above average (z- score > 0) SDC times, indicating that there may be other variables beyond the endurance and strength assessed in these two events influencing success on the IOCT. However, a majority of passing points still lie in the Quadrants 2 through 4, as depicted in the graph, demonstrating the principal that SDC and APFT rutimesec are both related with IOCT success. With these results in mind, this model has the potential to be used in assessing cadet readiness in relation to the IOCT, with implications for identifying undertrained individuals and prescribing remedial training. This can drastically reduce the uncertainty and anxiety experienced by cadets who need to take the test and can make their preparation more directed and efficient while freeing up time for academics and other activities. SDC and APFT runtimesec may not be perfectly indicative of IOCT readiness, as demonstrated by the significant number of slower-than-average passing records, and should not be used to conclusively mandate physical training or officially declare an individual underprepared. However, they present themselves as powerful tools for pre-IOCT screening which can contribute meaningfully to the array of educational and preparatory strategies already used by cadets and administrators alike.