This project aims to build a machine learning model that predicts a runner's 5K race time for a specific event. The model uses past race performances along with race location and environmental conditions to produce a single predicted time that accounts for the course and race-day factors. By using a large dataset and supercomputing resources, the project focuses on improving prediction accuracy by including realworld variables that strongly affect performance.

5K race times depend on more than just fitness. Weather conditions such as temperature and humidity, as well as course features like elevation and surface type, can have a noticeable impact on results. Many existing prediction methods either ignore these factors or apply very simple adjustments. This project uses machine learning to better capture these influences and generate more accurate and personalized predictions.

The dataset includes historical 5K times, race locations, environmental conditions on race day, and available course characteristics. The data is cleaned and prepared by handling missing values, normalizing numerical features, and encoding categorical information. Several models are tested, starting with basic approaches and moving to more advanced ones. Supercomputing resources are used to speed up training and evaluate multiple models efficiently.

So far, the dataset has been collected and organized, and initial data cleaning and feature selection have been completed. Baseline models have been implemented, setting the stage for further model improvement and evaluation.