This project will seek to use machine learning algorithms and data collected from five different runners to determine the effects of fatigue on runners' form. According to an article titled “The Unescapable Fatigue Effect,” as runners become more fatigued there is a gradual loss in neuromuscular control which has impacts on metrics such as foot strike characteristics. Being able to understand how fatigue effects runners will allow for informed decisions on shoes or orthotics or the ability to prepare training to target weaknesses and avoid injuries (Runscribe 10/27). According to “The Unescapable Fatigue Effect,” it is known that certain metrics are positively affected by run speed when the runner is not fatigued, so this project is predicted to reveal that as the runner becomes fatigued, they will lose control of their form which will been seen in the degradation of certain metrics such foot strike, impact, and pronation excursion (Runscribe 10/27).

The datasets provided are based off five different runners with information from their left and right foot as they run. Each foot of the runner is sampled approximately 2500 times and each step has various data measurements. Some of the main measurements are pace, stride length, step rate, flight ratio, contact time, shock, impact gravitational force, braking gravitational force, type of foot strike, pronation excursion, and max pronation velocity. There is an additional dataset which is more personal to each runner that includes the runner’s age, gender, weight, height, type of shoe worn, medical history, and more miscellaneous information that could potentially contribute to their foot metric results. These two datasets will be used together in model analysis to answer how fatigue impacts running form.

The planned procedures to be used for determining the effects of fatigue on the runner’s form will utilize Grid Search and Ensemble Learning as the main techniques for testing and training the models. Grid Search will be employed to find the best hyper parameters for the Logistic Regression, the Decision Tree, and the Random Forest models. With these hyperparameters, an Ensemble Learning technique will be utilized to combine the three model’s predictions. Then, a majority vote is used to decide the best prediction based on the models. A process diagram for the project is depicted in Figure 1.

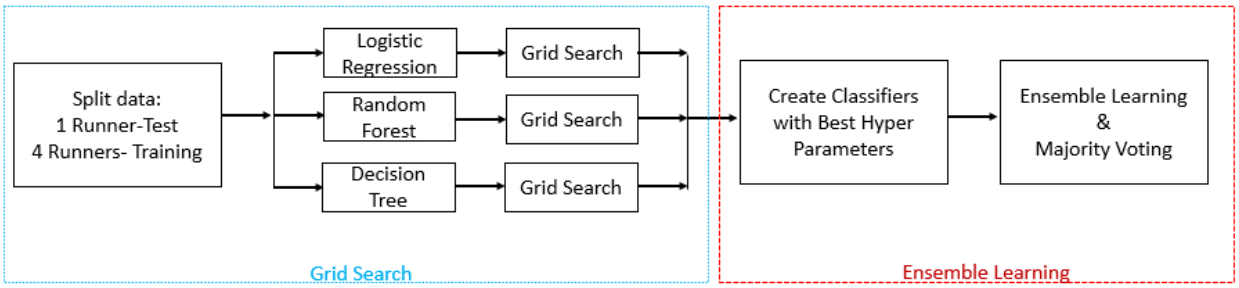


*Figure 1: The figure above depicts the process diagram for the project. The project will be split into two major processes which are defined as the Grid Search process to find the best hyper parameters and the Ensemble Learning process. The latter process utilizes the models and their best hyperparameters to create classifiers. Then, the models are combined, and Ensemble Learning will use the majority vote from all the classifiers to classify the data.*

The data will be split into two major groups- the left foot and the right foot. The data will then be separated further by runners. One runner will be designated as the test for the model, and the remaining four will be used for training. The Grid Search and Ensemble Learning will be performed on both major groups utilizing this test and training split.

The first step of the process is to use Grid Search on three different models to find the best hyperparameters for each model. The models employed will be the Decision Tree, Logistic Regression, and Random Forest models. Each model will be supplied the set of runner’s data and split into test and training sets. A pipeline will be created for each model. Additionally, a range for each parameter will be fed into the Grid Search. The model will then be trained with the data. A Receiving Operating Characteristic plot will be provided. This plot will detail how cross validation improves the model for each fold. The plot also depicts how each fold compares with a perfect performance and a performance of random guesses. Additionally, a plot of the results will be provided and a text file containing the

misclassified samples, classification accuracy, grid search score, and the best parameters will be included with the project.

The second step of the process is to employ the majority voting aspect of the Ensemble Learning technique as a method of using several models to combine the predictions of the models used in the Grid Search. The runner's data will be supplied to different models. Then, pipelines will be created for each of the machine learning models used in the Ensemble Learning. Pipelines will be created for each of the classifier models, as necessary. Next, a ten-fold cross validation will be performed on the data where a Receiver Operating Characteristic plot will be produced. Finally, majority voting is employed to improve the results of the model over one model in particular. A figure consisting of the plots of the different models training results in addition the majority voting plot will be provided.

References

"Metrics." *RunScribe*, https://runscribe.com/metrics/.

"The Unescapable Fatigue Effect." *RunScribe*, 12 Aug. 2017, https://runscribe.com/the-fatigue-effect/.