Methodology

The assignment provided data collected from six accelerometer and gyroscope sensors attached to the lower limb of eight subjects while walking on different terrains. Using this provided data the task focused on predicting the terrain by identifying it as one of four classes – solid ground, downward stairs, upward stairs, and grass. The data primarily suffered from class imbalance and differential sampling rates of attributes and respective labels. The mismatched sampling rates were dealt by mapping the closest y label to every x sample containing the attributes. Hence equalizing the number of x and y samples.

The seven models considered (Weighted Random Forest Classifier, Random Forest with random over sampling, Random Forest with random under sampling, Random Forrest with SMOTE and random under sampling, Random Forest with SMOTE tomek, Random Forest with ADASYN, Gradient Boost with SMOTE and random under sampling) were picked to course correct the specific shortcomings due to class imbalance combined with a potent classifier.

As random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset, we could effectively use a weighted average RF to target the moderation of a specific class with higher number of data points by allowing for its lesser influence using a lower weight. RF is also effective in calculating variable importance measures which aided in directing a variable’s overall influence on prediction.

An alternative to using a weighted RF was to use a simple RF and then compensate the class imbalance using random under and over sampling techniques. Random under sampling randomly eliminated data points of the majority classes to match that of the minority class. Hence reducing the strength of each class to that of the minority class giving a balanced data set. On the contrary random over sampling duplicated data points of the minority class to match the class strength of the majority class. Hence giving a balanced data set with each class strength equal to that of the majority class.

While random over sampling balanced out the data set it did not add to any information of the minority class trend. SMOTE is an approach to addressing imbalanced datasets by oversampling the minority class and augmenting the data using new instances synthesised from existing samples. The new instances are generated as a convex combination of two nearby instances using a point on the approximate path between the two, adding to the contribution of the minority class. While the minority class was oversampled using SMOTE the majority class was under sampled randomly, hence we were able to arrive at a strength for each class which was somewhere between the strength of the majority and minority class. Hence giving a balanced data strength with a robust class strength.

Though random under sampling was effective in reducing the data points of the majority class the data points to be eliminated are selected at random, this came at a risk of losing valuable information. SMOTE-tomek sampling improved on this by using the tomek-link algorithm which eliminated those data points from the majority class whose nearest neighbour was a point in the minority class. Hence balancing out the data by removing points from the majority class that were near the classification boundary and were susceptible to reflect certain degree of ambiguity.

Although SMOTE is an effective oversampling technique the synthetic samples were generated at random from the minority class. As opposed to ADASYN that used a density distribution as a criterion to automataically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions. This replicated more minority samples in a denser feature space, hence providing a balanced data set by contributing to data points in accordance with the minority class trend.

As opposed to RF, Gradient boost also serves as an operative classifier as it offers an ensemble boosting method that improves on the predecessor’s prediction with each iteration. For every instance in the training set, it calculates the residuals for that instance, Once it has done this, it builds a new Decision Tree that actually tries to predict the residuals that were previously calculated. Hence moving towards a more accurate prediction with each iteration and aggregating several weak learner predictions into a robust output prediction.

As each model in and of itself targets a specific property to maximise prediction accuracy an ensemble of these models resulted in a more coherent and thorough prediction. In order to obtain a rounded training set, while at the same time having a large enough test data set we split the provided data in a ratio of 70:30.