Machine Learning with Accelerometer and GPS Data
Deliverable #5

Data Files

- **exploratory/kenya_oct_15_data_processed.csv**
  - **Batch_id:** ID number for this batch
  - **time:** timestamp of this datapoint
  - **Longitude:** longitude of the location
  - **Latitude:** latitude of the location
  - **x:** vertical jolt (up and down, measured in meters per second squared) Vertical jolt is calculated as the incremental difference of vertical acceleration between timestamp t and t-1.
  - **y:** horizontal jolt (left and right, measured in meters per second squared) Horizontal jolt is calculated as the incremental difference of horizontal acceleration between timestamp t and t-1.
  - **z:** forward jolt (forward and back, measured in meters per second squared) Forward jolt is calculated as the incremental difference of forward acceleration between timestamp t and t-1.
  - **velocity:** traveling speed of the vehicle
  - **x_raw:** vertical acceleration without natural gravity (up and down, measured in meters per second squared)
  - **y_raw:** horizontal acceleration (left and right, measured in meters per second squared)
  - **z_raw:** forward acceleration (forward and backward, measured in meters per second squared)
  - **label:** label of road hazards (0: no road hazard; 1: speed bumps; 2: potholes)
  - **x_ratio_speed:** ratio of vertical acceleration divided by traveling speed
  - **x_ratio_z:** ratio of vertical acceleration divided by forward acceleration

R Scripts:

- **exploratory/kenya_oct_15_data_processed.Rmd**
  - Calculate **x_ratio_speed (**ratio of vertical acceleration divided by traveling speed)
  - Calculate **x_ratio_z** (ratio of vertical acceleration divided by forward acceleration)
  - Write a new data file **exploratory/kenya_oct_15_data_processed.csv**
- **data/find_speedbump.R**
  - Designed to label the following data files:
    - **data/los_angeles_1.csv**

- **data/los_angeles_2.csv**
- **data/los_angeles_3.csv**
- **data/los_angeles_4.csv**
- **data/los_angeles_5.csv**

- Write the following new data files:
    - **sklearn_Models/speedbumps_1.csv**
    - **sklearn_Models/speedbumps_2.csv**
    - **sklearn_Models/speedbumps_3.csv**
    - **sklearn_Models/speedbumps_4.csv**
    - **sklearn_Models/speedbumps_5.csv**
- Calculate the mean of vertical acceleration
- Calculate the standard deviation of vertical acceleration
- Given the exploratory analysis on the unlabeled data files, this R script labels any data points with 5 standard deviations or more away from the mean of vertical acceleration.

Python Scripts:

- **sklearn_Models/sklearn_CVGrid.py**
    - Parameter estimation using grid search with cross validation
    - sklearn.model_selection.GridSearchCV Model 1
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.868
        - StdDev F1 score: 0.108
        - Learning rate = 0.01
        - Minimum sample leafs = 1
        - N estimators = 150
        - Minimum sample split = 6
        - Max features = 4
- **sklearn_Models/sklearn_CVRandomized.py**
    - Randomized search on hyper parameters
    - Model with Rank 1:
        - Mean validation score: 0.868 (std: 0.108)
        - Hyper-parameters: {'learning_rate': 0.01, 'min_samples_leaf': 1, 'n_estimators': 150, 'max_features': 4, 'min_samples_split': 6, 'max_depth': None}
    - Model with Rank 2:
        - Mean validation score: 0.852 (std: 0.099)

        - Hyper-parameters: {'learning_rate': 0.1, 'min_samples_leaf': 1, 'n_estimators': 150, 'max_features': None, 'min_samples_split': 2, 'max_depth': None}
    - Model with rank: 2
        - Mean validation score: 0.852 (std: 0.099)
        - Hyper-parameters: {'learning_rate': 0.1, 'min_samples_leaf': 1, 'n_estimators': 150, 'max_features': 3, 'min_samples_split': 2, 'max_depth': None}
    - Model with rank: 2
        - Mean validation score: 0.852 (std: 0.099)
        - Hyper-parameters: {'learning_rate': 0.1, 'min_samples_leaf': 1, 'n_estimators': 150, 'max_features': 3, 'min_samples_split': 6, 'max_depth': 10}
    - Model with rank: 2
        - Mean validation score: 0.852 (std: 0.099)
        - Hyper-parameters: {'learning_rate': 1, 'min_samples_leaf': 1, 'n_estimators': 100, 'max_features': None, 'min_samples_split': 2, 'max_depth': None}
    - Model with rank: 2
        - Mean validation score: 0.852 (std: 0.099)
        - Hyper-parameters: {'learning_rate': 0.1, 'min_samples_leaf': 1, 'n_estimators': 150, 'max_features': 3, 'min_samples_split': 2, 'max_depth': 5}
- **sklearn_Models/sklearn_MLPClassifier.py**
    - Multi-Layer Perceptron (neural network) Classifier model
    - Data: Los Angeles
    - sklearn.neural_network.MLPClassifier Model 1
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.760686122
        - StdDev F1 score: 0.1537659455829
        - Median F1 score: 0.77350427350
        - IQR F1 score: 0.1333334
        - Skewness F1 score: -0.17764724445877225
    - sklearn.neural_network.MLPClassifier Model 4
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.7709969
        - StdDev F1 score: 0.1402254

        - Median F1 score: 0.80000000
        - IQR F1 score: 0.13333334
        - Skewness F1 score: -0.17764724445877225

- **sklearn_Models/sklearn_DTClassifier.py**
    - Decision Tree Classifier model
    - Data: Los Angeles
    - sklearn.tree.DecisionTreeClassifier Model 1
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.89899393769
        - StdDev F1 score: 0.118201302569
        - Median F1 score: 0.9230769231
        - IQR F1 score: 0.0921034
        - Skewness F1 score: -0.9312941130701623
    - sklearn.tree.DecisionTreeClassifier Model 5
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.84655181
        - StdDev F1 score: 0.10527561
        - Median F1 score: 0.866071428
        - IQR F1 score: 0.138907
        - Skewness F1 score: -0.489563
- **sklearn_Models/sklearn_RFClassifier.py**
    - Random Forest Classifier model
    - Data: Los Angeles
    - sklearn.ensemble.RandomForestClassifier Model 1
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.744434045
        - StdDev F1 score: 0.160948970
        - Median F1 score: 0.75
        - IQR F1 score: 0.231867
        - Skewness F1 score: -0.391874
    - sklearn.ensemble.RandomForestClassifier Model 5
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.74717207
        - StdDev F1 score: 0.172979588
        - Median F1 score: 0.769230769

        - IQR F1 score: 0.222222222222
        - Skewness F1 score: -0.98346783
- **sklearn_Models/sklearn_GBClassifier.py**
    - Gradient Boosting Classifier model
    - Data: Los Angeles
    - sklearn.ensemble.GradientBoostingClassifier Model 1
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.8221567168
        - StdDev F1 score: 0.0.120072183
        - Median F1 score: 0.0.857142856
        - IQR F1 score: 0.16923077234
        - Skewness F1 score: -0.621688196
    - sklearn.ensemble.GradientBoostingClassifier Model 2
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.3518205293
        - StdDev F1 score: 0.13588175
        - Median F1 score: 0.33333333
        - IQR F1 score: 0.1246923817
        - Skewness F1 score: 0.76200239
    - sklearn.ensemble.GradientBoostingClassifier Model 3
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.8399761
        - StdDev F1 score: 0.11177313
        - Median F1 score: 0.857142857
        - IQR F1 score: 0.20003817
        - Skewness F1 score: -0.66963835
    - sklearn.ensemble.GradientBoostingClassifier Model 4
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.8191443628
        - StdDev F1 score: 0.104556871
        - Median F1 score: 0.80000000
        - IQR F1 score: 0.16734588
        - Skewness F1 score: -0.8265433
    - sklearn.ensemble.GradientBoostingClassifier Model 5
        - Features: speed, Z-accel, Z-jolt

        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.84139751
        - StdDev F1 score: 0.10140479
        - Median F1 score: 0.84210526
        - IQR F1 score: 0.237768
        - Skewness F1 score: -0.1524893
- **sklearn_Models/sklearn_Logistic.py**
    - sklearn.linear_model.LogisticRegression Model 1
    - Data: Los Angeles
        - Features: speed, X-accel, Y-accel, Z-accel, Z-jol
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.32698412
        - StdDev F1 score: 0.07853534
        - Median F1 score: 0.333333333
        - IQR F1 score: 0.1222243222
        - Skewness F1 score: -0.5984130
    - sklearn.linear_model.LogisticRegression Model 2
        - Features: speed, Z-accel, Z-jolt
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.480952380
        - StdDev F1 score: 0.131621666
        - Median F1 score: 0.5
        - IQR F1 score: 0.0888888888889
        - Skewness F1 score: 0.063076006
- **sklearn_Kenya/sklearn_DTClassifier.py**
    - Decision Tree Classifier model
    - Data: Kenya
    - sklearn.tree.DecisionTreeClassifier Model
        - Features: x, y, z
        - Labels: speedbump (1 = yes, 0 = no)
        - Average F1 score: 0.0353807517853
        - StdDev F1 score: 0.0152980191865
        - Median F1 score: 0.0266666666667
        - QR F1 score: 0.0261269020959
        - Skewness F1 score: 1.0654257915650123
        - Zero F1 score: 0
- **sklearn_Models/sklearn_GBClassifier.py**
    - Gradient Boosting Classifier model
    - Data: Kenya

  - sklearn.ensemble.GradientBoostingClassifier Model
    - Features: x, y, z
    - Labels: speedbump (1 = yes, 0 = no)
    - Average F1 score: 0.0273425270958
    - StdDev F1 score: 0.00373470331891
    - Median F1 score: 0.0274024024024
    - IQR F1 score: 0.00262891734411
    - Skewness F1 score: 0.5085889688574291
- **tensorflow_Kenya/Kenya_RNN.py**
  - Data: Kenya
  - Keras Sequential LSTM Model
    - Features: x, y, z
    - Layers: 12
    - - 2s - loss: 0.0277 - mean_squared_error: 0.0337 - val_loss: 0.0185 - val_mean_squared_error: 0.0185
    - [0.018500001268249109]
    - MAE: 0.0185000012682
    - RMSE: 0.136014705087