Group Project W207

Section 6

Group 3 Final Project Baseline Submission

Title: **Forest Cover Type Prediction**

Kaggle Link: **https://www.kaggle.com/c/forest-cover-type-prediction/overview**

Team Members: **Savita Chari, Tymon Silva, Blake Bormes, Andrew Beckerman**

Git Repo : https://github.com/savitaChari/W207-Final-Project-Group3_Section6

1. Exploratory data analysis & Split the data into train/dev/test
    - Link to the Notebook for EDA
2. Metric of evaluation and why you chose it
    - Link to the Notebook for Metric Evaluation
3. A simple machine learning technique & Evaluate results
    - Link to the Notebook for ML & Evaluation

Explain how you will evaluate any challenges

Potential challenges could include our scaling methodology affecting binary features to have negative values. We will evaluate this challenge by attempting to normalize our data to have values between 0 and 1 and also by converting continuous features to be binary and not transforming our binary features. With these changes, we will see which has the best effect on our model performance.

An additional challenge could be training the neural network model with too many epochs. This could cause extended computation times for our model, so we will adjust for this by starting with lower numbers of epochs as we fine tune other model hyper-parameters.

We may encounter overfitting, so we can use our test data to ensure the accuracy of the test data improves at a similar rate to the accuracy of our training data. We also can evaluate if changing the ratio of training and test data from our original dataset will help reduce overfitting.

Lastly, differing class distributions between the training and test set could cause poor model performance, so stratifying the dataset to ensure class distribution is identical within both training and test sets will adjust for this.

Briefly describe what you still plan to do

Our team plans to also test a K Nearest Neighbors model to understand which model type, Neural Network or KNN, best predicts forest covers. As time permits, we will test additional models as well.

In order to improve both our KNN and Neural Network models, our team will leverage different feature engineering techniques and fine tune model hyperparameters by:

- Scaling our data by subtracting by the mean and dividing to ensure our data points are scaled equally
- Normalizing our data to ensure all data points are between 0 and 1
- Converting continuous features to be binary to ensure all features (binary and continuous) are treated equally by the model
- Adding additional layers to our neural network, adjusting the learning rate or epochs, using softmax or sigmoid, changing the loss function or optimizer, and changing other neural network hyperparameters
- Adjusting the K value, smoothing factor/ alpha, and other KNN hyperparameters

Lastly, our team will use our metrics of evaluation to determine the best model, hyperparameters, and feature engineering techniques to predict forest cover types.