CS 372: Project 3 Report

Myriah Hodgson

1. *A description of your dataset and where you find it with the link to download the dataset.*

The dataset is provided by 'Responsible Datasets in Context', but it is just an already pre-filtered dataset from the NPS (National Park Service) website, using their query tool. The data contains visitor information about all national parks, such as how many total visitors came to different parks within a given month, as well as the types of visitors (campers, backcountry, etc). The dataset covers years 1979-2023, and the granularity is that each row is specific to a month's total of data. It has 12 features, and approximately 30,000 rows. The data can be found at: https://www.responsible-datasets-in-context.com/posts/np-data/?tab=explore-the-data

2. *A description of the task you are trying to accomplish with the data via machine learning techniques.*

My intention with this project is to predict whether a data entry is during 'High Season' or 'Low Season', depending on whether or not the month is above or below the median TotalVisitation counts for that specific park. Because certain parks are more popular than others, the median and resultant threshold for determining whether a park is in high season or low season is specific to each individual park.

3. *A description of any preprocessing you did to the dataset.*

The dataset I chose to work with was already rather clean. Before any preprocessing, there were already no null values, and all data types that were expected to be numeric were already in integer format. The only somewhat concerning discrepancy in the data was that there were quite a few 0 entries, particularly in the Backcountry feature. However, it is not unlikely that many parks either typically do not support Backcountry hiking or camping, or that it is not available during all seasons at that park. I did drop the 'ParkType' column, because all entries are National Parks. I also utilized groupby and joining functions to specify the threshold for high and low season, specific to each park, in addition to encoding categorical variables as numeric.

4. *A summary of prediction accuracy of all the evaluated machine learning methods in training, validation, and test sets.*

Model	**Training Accuracy**	**Validation Accuracy**	**Test Accuracy**
Logistic	78%	78%	78%
SVM (rbf)	80%	74%	76%
Decision Tree	90%	88%	88%
Random Forest	95%	91%	92%
Neural Network	85%	85%	84%

5. *Write 4-5 sentences summarizing your observations regarding these results, with a comment on the underfitting/overfitting/convergence performance of your models.*

My random forest model performed the best, with the decision tree model performing second best. My random forest model yielded the best test accuracy of 92% on the unseen test set. The decision tree model also performed well, ending at a test accuracy of 88%. I chose to simply look at accuracy because the stakes of model optimization here are really low –– determining whether or not a park is busy does not really need to consider recall more strongly than precision, nor the other way around. In most of these models we see a bit of overfitting, with the exception of the logistic and neural network models, which had convergence in that the training and validation accuracies were the exact same –– and for the neural network, the test accuracy was only slightly lower. Because my training accuracy was higher than test/validation accuracy for the random forest, decision tree, and SVM models, these models were overfitting to the training data. It is also worth noting that the rbf kernel behaved much better than the polynomial kernel for the SVM model.