Carter Rogers

Prof. Amir Jafari

Intro to Data Mining

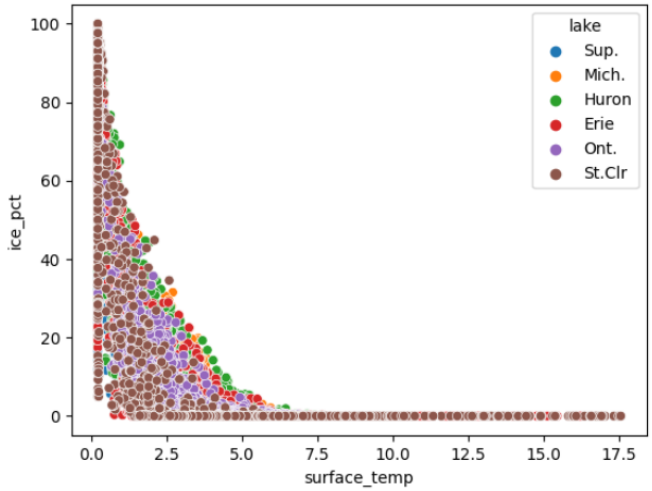
06 December 2021

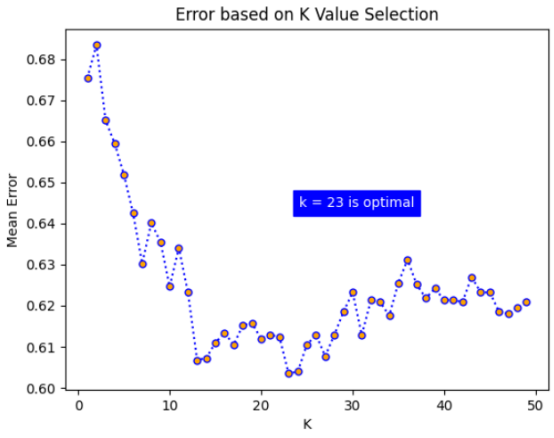
Individual Final Report

Our goal for this project was to utilize data mining techniques and algorithms to study the physical properties of the Great Lakes. We collected data from NOAA on the physical properties of the Great Lakes and cleaned it in preparation for modeling. Then, we used classification and regression modeling techniques to describe the properties of the Great Lakes. Finally, we developed simple GUIs to summarize the results of the modeling process in an intuitive manner.

I was responsible for the classification modeling for the project. I chose to use K-Nearest Neighbors to classify data points into their respective lakes based on ice concentration and surface temperature in order to gain a further understanding about the physical properties of the Great Lakes. KNN is a popular classification modeling technique that uses data points in close proximity around a given observation in order to assign a classification to that observation. KNN excels with data of lower dimensionality and is relatively easy to comprehend. Euclidean distance is the most common metric used to evaluate proximity, which becomes very computationally expensive with higher-dimensional data (Joby). Since we are only using two features to classify our observations, KNN is a good fit for our modeling purposes. The only tuning required for a KNN model is the choice of k-value. In order to find a suitable k-value for the model, I evaluated the error for each model built off of k-values from 1 to 50, where error is defined as the percentage of incorrectly classified observations. The ideal k-value ended up being $k = 23$, with the final KNN model built off of $k = 23$ yielding an accuracy percentage of 39%.

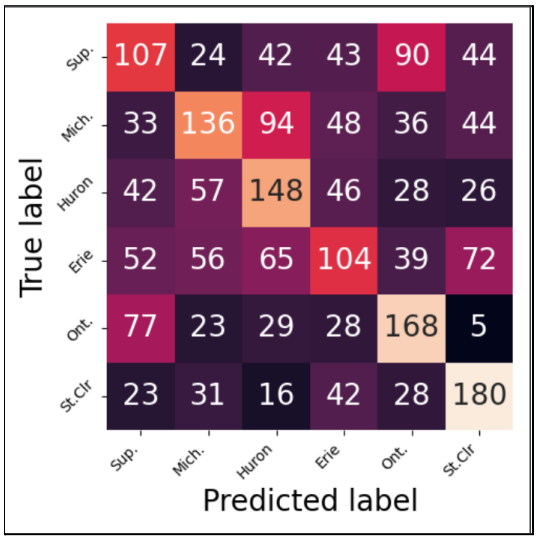
I was primarily responsible for the classification modeling and the development of the GUI for the classification model, as well as significant portions of the group report and the presentation slides. Once I obtained the cleaned data from Rhys, I performed the necessary preprocessing which consisted of filtering out NaN values, renaming and reordering columns, and very basic EDA. The most useful EDA plot I was able to obtain portrays the surface temperature plotted with the ice concentration, colored by the lake from which the observation was recorded. After that, I normalized the features and split the data into training and test sets. Then, I began to implement the KNN classifier model by fixing the k-value at k = 5. I was dissatisfied with the result of this first model, so I used a loop to compare error values for models built using k = 1 to k = 50. By plotting the error for each k-value, I observed that the optimal k-value that minimizes the error of the model is k = 23. This k-value was used to construct the final model that scored an accuracy percentage of 39%. After finishing the modeling component, I built a simple GUI that allows the user to input a k-value and outputs the accuracy percentage of





the model that corresponds to that k-value. This serves as a simple way for the user to understand how the choice of k affects the accuracy of the model.

The results of the model can be visualized using a confusion matrix that compares the true vs. predicted labels. We can see that our model struggled with classifying data points between Lake Ontario and Lake Superior. The model excelled at classifying data points from Lake St. Clair, which is technically not one of the five Great Lakes but was included in the data set. Overall, the model was 39% accurate in classifying the data. This is fairly low, but it is important to consider that a naive model of randomly classifying the data would be about 16.7% accurate. Our model significantly outperforms this theoretical naive model.



From the results of the model, we can infer several aspects of the physical features of the Great Lakes. Since the model struggled to distinguish between Lake Superior and Lake Ontario, we can infer that these two Great Lakes have similar physical properties, and thus the model had difficulty distinguishing between the observations gathered from them. On the other hand, the model performed well in classifying points from Lake St. Clair. This is a notable result because Lake St. Clair is far smaller than the other Great Lakes, so we can infer that it's physical features are much different than the other lakes in the data. The model likely classified these points with greater accuracy because of its relative uniqueness. In the future, we could improve upon the accuracy of the model by ensembling with other classification techniques such as Support Vector

Machine or Gradient Boosting. We could also build a more robust and complex GUI that further helps the user understand the modeling process and visualize the results in an intuitive manner. Percentage of code found/copied from the Internet: $(30 - 15)/(30 + 70) * 100 = 15\%$

Works Cited

Beeton, Alfred M.. "Great Lakes". Encyclopedia Britannica, Nov 23rd, 2020, https://www.britannica.com/place/Great-Lakes. Accessed 3 December 2021.

Joby, Amal. “What Is K-Nearest Neighbor? A ML Algorithm to Classify Data.” *Learn Hub*, 19 July 2021, https://learn.g2.com/k-nearest-neighbor.