

# Using ML models to predict the occurrence of invasive species based on habitat preferences

By JanetLou Guan

Algonquin Regional High School, Northborough, MA

## Abstract

Invasive species are posing significant ecological, socio-economic, and human health threats to our society. It is imperative to accurately predict the potential occurrences of invasive species for management to concentrate efforts on prevention, early detection, and swift response. This study aims to employ various ML algorithms to forecast the likelihood of invasive species occurrences based on habitat preferences. The study compared the prediction accuracies of three ML algorithms: Random Forest, Logistic Regression, and Gaussian Naive Bayes. The analysis utilized data collected from twelve lakes located in the Adirondack region of Upstate New York. The outcomes of the study reveal that the Gaussian Naive Bayes model exhibited markedly higher accuracy levels compared to both the random forest and the logistic regression model. These findings highlight the effectiveness of the Gaussian NB model in predicting invasive species occurrences, underscoring its potential as a valuable tool for proactive management and conservation efforts.

Key words: Random Forest; logistic regression; Gaussian naive bayes; machine learning; algorithm; probability; occurrence; habitat preferences

## 1. Introduction

An invasive species is a non-native species whose introduction causes or is likely to cause economic, environmental, or human health harm, or threatens to disrupt the current balanced ecosystem (Executive Order 13751). It can be any kind of living organism: plant, insect, fish, bacteria, fungus, or even an organism's seeds or eggs. It can infiltrate new environments through various means. Many invasive species are introduced into a new region accidentally and a lot of times, by human beings. The era of globalization has significantly amplified both long-distance travel and trade, thereby escalating the incidence of non-native flora and fauna being brought into different ecosystems worldwide.

Invasive species can impact both the native species living within an ecosystem as well as the ecosystem itself. They can change the food web, compete with native organisms for limited resources, cause the extinction of native plants and animals, and reduce biodiversity<sup>1</sup>. (e.g. Vilà et al., 2011), and alter ecosystem functioning<sup>2</sup> (e.g. Pejchar & Mooney, 2009). Numerous studies have shown the negative effect of invasive species on native biodiversity. Some of them concentrated on the genetic level such as how the breeding process (specifically, males and females of two different species were bred together) affected biodiversity<sup>3,4</sup> (Largiadèr 2008; Kumschick et al. 2015). Other studies have demonstrated the negative effect from the ecosystem level<sup>5,6</sup> (Lazzaro et al. 2020; Viciani et al. 2020).

According to the National Wildlife Federation, approximately 42 percent of threatened or endangered species are at risk due to invasive species. In the United States, the annual estimated economic and health-related costs of invasive species have been reported at more than \$21 billion (United States geological survey).

Aquatic invasive species (AIS) have inflicted significant ecological harm on freshwater ecosystems<sup>7,8</sup> (Ricciardi and MacIsaac 2000, Cucherousset and Olden 2011), underscoring the need for a more proactive approach to invasive species management<sup>9,10</sup> (Leung et al. 2002, Pagnucco et al. 2015). The identification and effective management of invasive species relies on an accurate prediction of locations and environment favorable to non-native species for them to survive, establish, reproduce, and spread<sup>11</sup> (Kramer et al. 2017).

However, ecological data are often high dimensional with nonlinear and complex interactions among variables, and with many missing values among measured variables<sup>12</sup> (Sabat-Tomala et al. 2020). Traditional statistical approaches can encounter difficulties in extracting meaningful analyses from such data. Linear statistical methods, such as generalized linear models (GLMs), in particular, may prove inadequate for revealing patterns and relationships that can be unveiled by more advanced techniques<sup>13</sup> (De'ath and Fabricius 2000).

Among the statistical techniques frequently utilized in ecology, classification procedures stand out as some of the most widely adopted, finding applications in tasks such as remote sensing-based vegetation mapping<sup>14</sup> (Steele 2000) and species distribution modeling<sup>15</sup> (Guisan and Thuiller 2005). In recent years, classification trees<sup>16</sup> (Breiman et al. 1984) have gained widespread popularity among ecologists due to their straightforward interpretability, exceptional classification accuracy, and capacity to characterize complex interactions among variables.

The main objective of this study is to construct several different machine learning models and use these models to classify the known species and unknown species (classified as

“other”). I then use these classified labels to analyze the population for each category. Based on the population change in the temporal dimensions data, we will be able to predict whether or not new invasive species have occurred in a particular region. The three models we selected were Random Forest, Logistic Regression, and Gaussian Naive Bayes.

## 2. Results

The prediction accuracy of all three AI algorithms is listed in Table 1 and Table 2.

Table 1: Comparison of results on a dataset including all NaN values

	Random Forest	Logistic Regression	Gaussian NB
Unbalanced	95.20%	36.36%	29.18%
Balanced	35.11%	50.56%	81.23%

Table 2: Comparison of results on dataset excluding all NaN values

	Random Forest	Logistic Regression	Gaussian NB
Unbalanced	99.35%	98.12%	96.14%
Balanced	99.89%	80.80%	95.16%

All three models were trained both unbalanced (with equal weight on all labels) and balanced (with greater weight on non-NaN labels). In each scenario, each was trained first with all NaN values included, and then with all NaN values excluded.

In comparison, including all NaN values, random forest yielded the highest unbalanced accuracy, followed by Logistic regression, then by Gaussian NB. However, random forest yielded the lowest balanced accuracy, whereas Gaussian NB yielded the highest balanced accuracy.

Excluding all NaN values, random forest yielded the highest unbalanced accuracy, followed by Logistic regression, then by Gaussian NB. Random forest also yielded the

highest balanced accuracy, followed by Gaussian NB, then by Logistic regression.

At the same time, occurrences of different species were analyzed according to specific features in order to analyze where in a region each species is most likely to inhabit. Figures 2 and 3 below showed occurrences of each of the four species and the “Other” new species based on distance from shore and depth, respectively. The graphs were created using Python’s Matplotlib and Seaborn libraries. Matplotlib allowed for the creation of various graphs when analyzing the data used in the study. Through Matplotlib, histogram graphs were generated, based on how many times each species appeared with each feature.

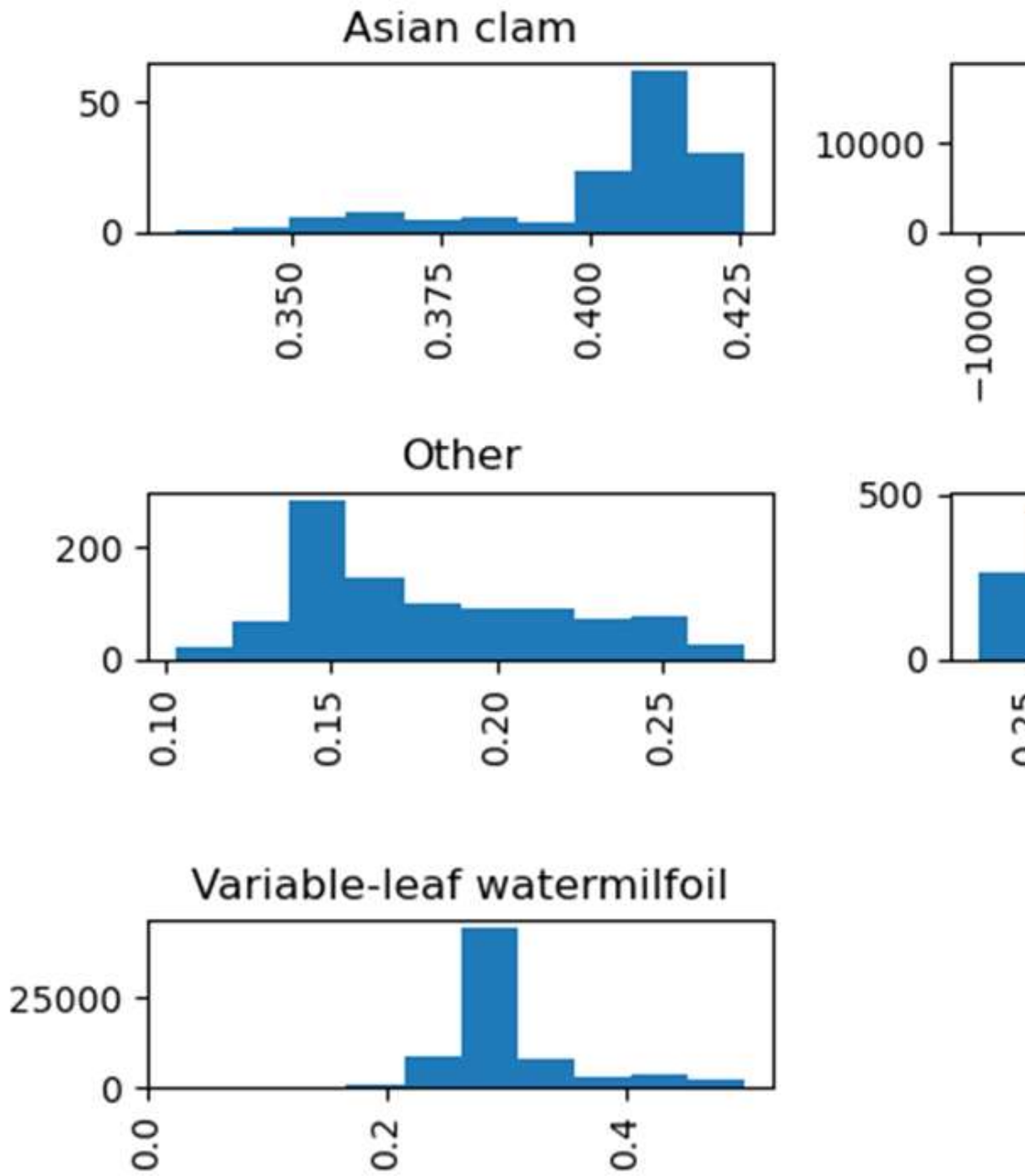


Figure 2. Occurrences of species based on distance from shore (the x-axis represents the distance from shore in meters, while the y-axis represents occurrences of species.)

Figure 2 showed that species Asian Clam were found closest to the shore, with most occurrences appearing less than 0.5 meters from the shore. The Spiny Waterflea had the

most occurrences appearing between 0.25 and 0.3 meters from shore, while most occurrences of other, non-specified species in the data occurred at about 0.15 meters. Most instances of the Eurasian Watermilfoil and the Variable-Leaf Watermilfoil occurred the farthest from shore at about 0.5 meters.

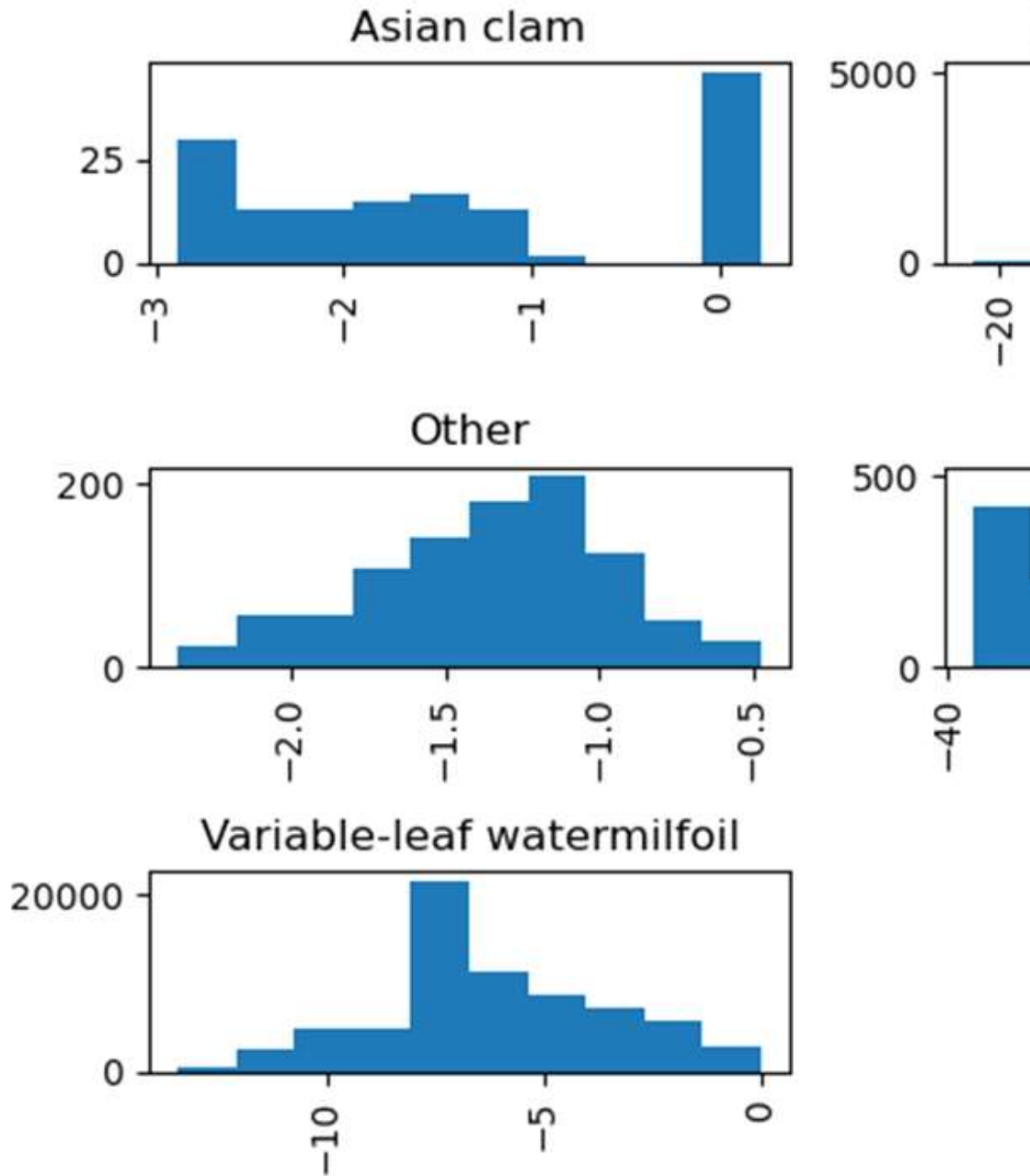


Figure 3. Number of occurrences of species based on depth (the x-axis represents the depth in meters, while the y-axis represents occurrences of species.)

Figure 3 showed that most Asian Clam and spiny waterflea were found around a depth of 0. Most instances of the Eurasian Watermilfoil occurred around a depth of -5 meters, while instances of the Variable-Leaf Watermilfoil occurred at a depth between -5 and -10 meters. Other, non-specified species were found to mostly occur at a depth of around -1 meters.

Seaborn allowed for the creation of scatter plots and area charts, color-coded based on the five species types present in the data (seen in the key) and organized based on the regional features of the data (labeled on the y-axis).

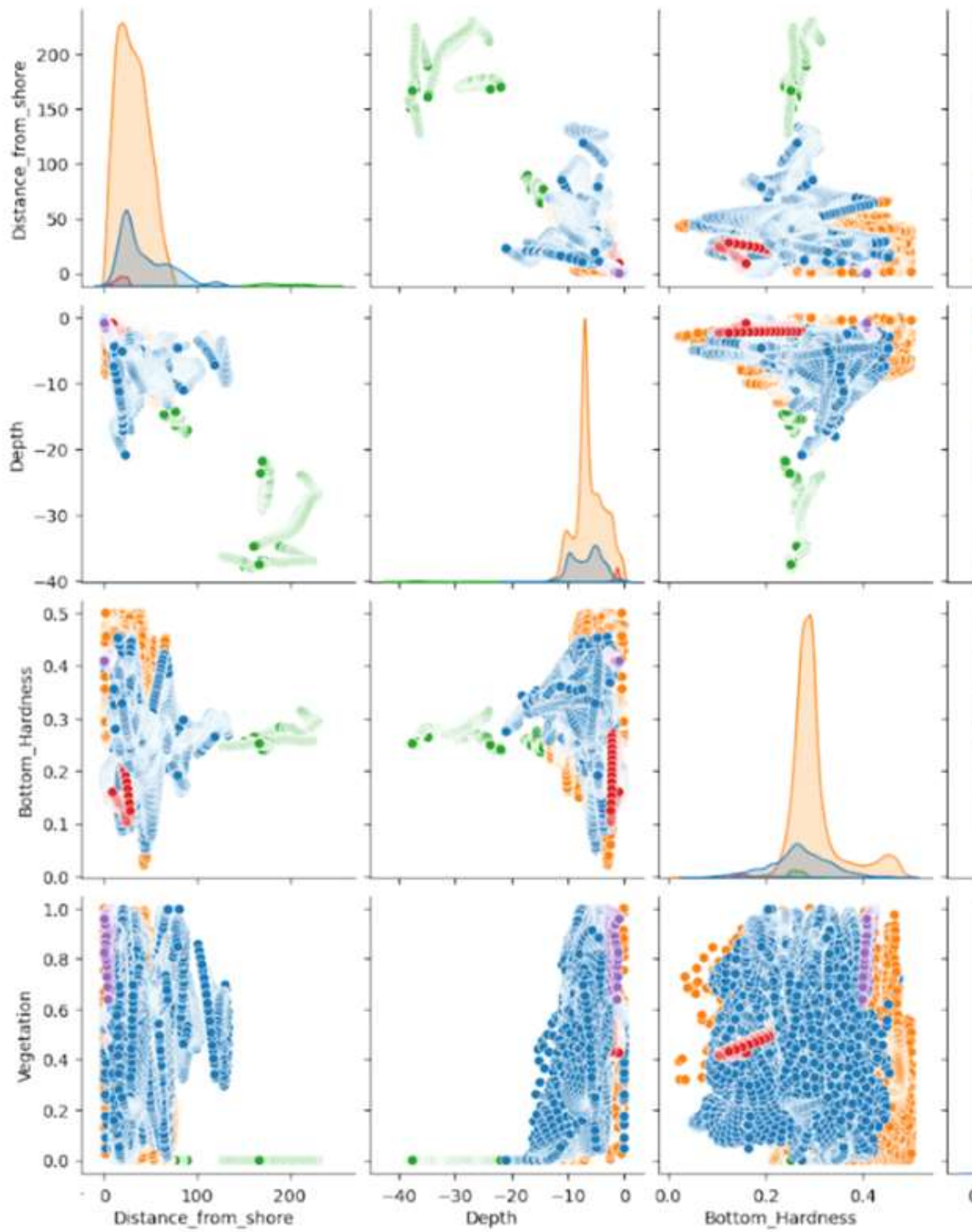




Figure 4. Visual representation using Seaborn (where distance and depth are measured in meters, hardness is measured in Joules/cubic meters, vegetation is measured in volume/cubic meters, and longitude and latitude are measured in degrees.)

From Figure 4 we can observe the relationship between specific variables and occurrences of the five species types observed in the study.

The Eurasian watermilfoil (blue) appeared the most in the figure, and most instances occurred at greater depths. It was also observed to occur closer to the shore, with most instances found less than 100 meters from shore.

The variable-leaf watermilfoil (orange) occurred mostly closer to the shore at areas with greater bottom hardness; however, there were several occurrences further from shore as well. The species appeared to be the most spread out.

The spiny waterflea (green) consistently appeared furthest from the shore compared to the other four species, over 100 meters from shore. It is also found in areas of lower vegetation.

The Asian clam (purple) was seen to appear in areas of high vegetation, as well as closer to the shore. It appeared minimally varied, with occurrences all exhibiting the same features.

Other species (red) seemed to have limited variation in their spread. They mostly appeared under 50 meters from the shore and in areas with depth near 0 and less bottom hardness.

Points on the graphs showing longitude and latitude can be mostly seen along two lines (-75 and -73.5 degrees longitude and 43.4 and 43.8 degrees latitude) due to the twelve lakes being located closely along those longitudes and latitudes. As a result, not much information can be obtained solely by looking at the longitude and latitude.

### **3. Discussion**

Being able to accurately predict the probability of occurrence of invasive species in a particular region is very important for developing efficient management and prevention strategies. In this study, three AI models were used to predict the probability of

occurrence, and their accuracies were compared.

Based on the results gathered from each model, the random forest model performed with the highest unbalanced accuracy at 99.35%, with NaN values excluded, and 95.20% with NaN values included. Logistic regression and Gaussian NB models' unbalanced, NaN-excluded accuracies were slightly lower at 98.12% and 96.14%, respectively; furthermore, their balanced models were also lower at 80.80% and 95.16%, respectively. Thus, the random forest model consistently performed slightly better than both the logistic regression and Gaussian NB models.

Despite this, it is important to note that a model that could predict accurately with NaN values included would be the most useful in forecasting occurrences of invasive species, as there may be numerous locations in a region where no species occurs. Looking at the accuracies of the three models trained with NaN values included, we can see that, while the random forest model predicted a high unbalanced accuracy of 95.20%, its balanced accuracy was much lower at 35.11%. The logistic regression model predicted much lower as well, with a 36.36% unbalanced and 50.56% balanced score, respectively. The Gaussian NB model performed with the lowest accuracy unbalanced at 29.18%; however, it performed with the highest balanced accuracy at 81.23%.

Balanced accuracy score is a further development on the standard accuracy metric where it's adjusted to perform better on imbalanced datasets which is the case for this research. The way it does this is by calculating the average accuracy for each class, instead of combining them as is the case with standard accuracy.

Balanced accuracy scores are better signifiers of accuracy than unbalanced scores as they are less prone to overfitting or underfitting (which occurs when the models become too specialized and make biased predictions favoring a label over others in the dataset). An unbalanced accuracy score with our database, for example, would overfit and put more emphasis on predicting a species such as variable-leaf watermilfoil rather than the Asian clam due to there being far more instances of the former in the data: 73,525 instances of variable-leaf watermilfoil versus 149 instances of Asian clam. Therefore, a balanced accuracy score would be more useful in prediction than an unbalanced one.

As a result, I look at the best-performing model that has the highest balanced accuracy score with all NaN values included. This makes Gaussian NB the ideal method with an 81.23% accuracy among the three tested models when making species predictions based on the locational data provided.













#### **4. Data and Methods**

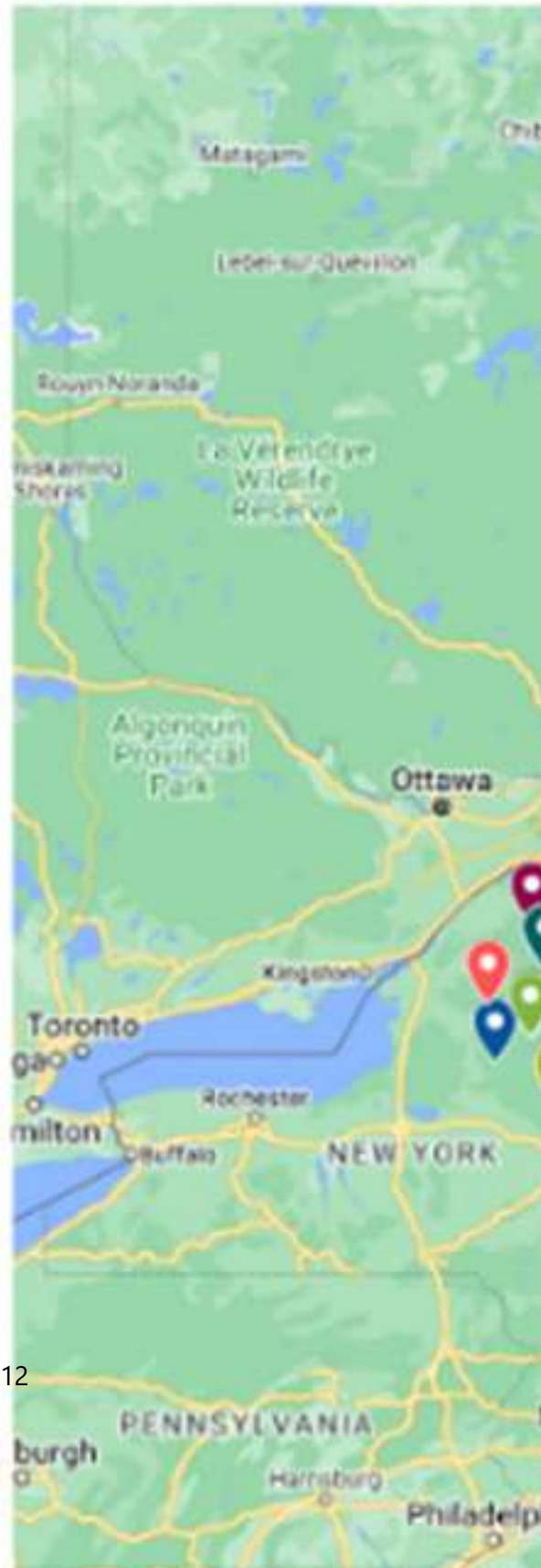
#### **4.1. Area of Interest**

The study area included twelve lakes in the Adirondack region in Upstate New York (Figure 1).

# 12 Lakes in Survey

## 12 Lakes in Survey

-  Hadlock Pond
-  Loon Lake
-  Little Wolf Pond
-  Long Lake
-  Sixth Lake
-  East Stoner Lake
-  Long Pond
-  Clear Pond - Parishville
-  Horseshoe Pond
-  Fifth Lake
-  Courtney Pond
-  Clear Pond - Croghan



**Figure 1.** Area of interest

## 4.2. Species

Four species in the twelve lakes were chosen in this research based on their spreading speed and management agency concerns. These four species include two plants and two invertebrates: Variable Leaf Milfoil, Eurasian Watermilfoil, Spiny water fleas, and Asian clam.

**Variable Leaf Milfoil** (*Myriophyllum Heterophyllum*) (VLM) is a submerged aquatic plant with fine, feather-like leaves whorled around a main stem. VLM is native to the Southeastern and Midwestern United States. It was first observed in New England in Bridgeport, Connecticut in 1932 and is now widely distributed across the New England States. Variable milfoil grows in both still and flowing waters in a variety of substrates at depths from 1 to 5 meters.

Variable leaf milfoil can cause numerous ecological, cultural, and economic impacts. VLM spreads quickly via fragmentation and can easily displace beneficial native aquatic plants. Dense beds of variable leaf milfoil degrade water quality for numerous species of fish and wildlife. In large mats, dissolved oxygen levels can be reduced to zero, making the area completely uninhabitable to game fish. Thick growths of VLM can impede fishing, swimming, and boating, thus indirectly impacting tourism and the economic activity of lake towns.

**Eurasian Watermilfoil** (*Myriophyllum spicatum*) is a submersed, rooted aquatic plant native to Europe, Asia, and northern Africa. It was first reported in the United States in the 1880s<sup>17</sup> (Eiswerth et al. 2000). The plants are rooted at the lake bottom and grow rapidly creating dense beds and canopies. They typically grow in water 1 to 4 meters (3.2 to 13 feet) deep, but have been found in water as deep as 10 m (32.8 ft). Stem densities can exceed 300/m<sup>2</sup> (359/yd<sup>2</sup>) in shallow water.

The widespread of Eurasian Watermilfoil could cause both ecological and economic damage. The introduction of Eurasian watermilfoil can result in native macrophyte diversity and abundance declines. Eurasian watermilfoil beds form dense canopies at the water surface thereby reducing light penetration early in the season before native macrophytes have reached their full growth, shading them out and slowing/reducing growth potential. Milfoil-infested lakes tend to have reduced fish spawning areas and lowered fish growth rates. Besides, the negative impacts on wildlife and fish populations in water bodies with high densities of Eurasian watermilfoil and the difficulty of motor boating and swimming in infested areas result in recreation-oriented financial losses and

the depreciation of shoreline property values.

**Spiny Waterflea** (*Bythotrephes longimanus*) is native to Europe and Asia. The species was unintentionally introduced into the United States' Great Lakes through the discharge of contaminated cargo ship ballast water. They were first discovered in Lake Huron in 1984; established in all of the Great Lakes by 1987<sup>18</sup> (Cullis 1988). Spiny waterfleas live in freshwater habitats and prefer cold temperatures, but can tolerate both brackish and warm water. Spiny waterfleas spread by attaching to fishing lines, downriggers, anchor ropes, and fishing nets and hitch a ride to other bodies of water.

Spiny waterfleas negatively impact native fish populations, aquatic habitats, and sports fishing. Spiny waterflea can clog eyelets of fishing rods and prevent fish from being landed. They also prey on native zooplankton, including *Daphnia*, which are an important food source for native fishes. In some lakes, spiny waterfleas can cause the decline or elimination of some species of native zooplankton.

**Asian Clam** (*Corbicula fluminea*) is native to the fresh waters of eastern and southern Asia. It was likely introduced to the West Coast of North America around 1930, initially assumed to have been imported as a food source for the immigrating Chinese population (USACE ERDC 2007). Live Asian clams were first detected in US waters in 1938 in the Columbia River, Washington; the species quickly spread across the continent and is currently found in 44 states.

The Asian clam is an invasive freshwater clam that prefers sandy lake bottoms and can be found at the sediment surface or slightly buried. Asian clams multiply rapidly and populations can easily reach high densities in freshwater. Asian clams are filter feeders, which means that they take in lake water and strain out algae. At high densities, Asian clams can out-compete other native filter feeders (such as fish, mussels and aquatic insects) for available food. Asian clams have played a role in the decline of many freshwater clams and mussels, reducing native biodiversity. Shells of large populations may also clog intake pipes of power and water facilities.

### 4.3. Data

I used data from Adirondack Research Invasive Species Mapping on LILA BC (Labeled Information Library of Alexandria: Biology and Conservation). The data set contains interpolated lake characteristics data of twelve lakes in the Adirondack region, Upstate New York, including depth, substrate hardness, and vegetation presence. This data is useful for calculating the probability of occurrence of other biological organisms that have habitat preferences related to factors such as water depth, vegetation, and substrate.

I obtained 5,769,518 instances of data, collected from a period of 24 months between 2018 to 2019. Of this data, there were 73,525 instances of Variable-Leaf Watermilfoil, 18,193 instances of Eurasian watermilfoil, 1,627 instances of Spiny Waterflea, 149

instances of Asian clam, and 984 instances of other unidentified species. All remaining instances were NaN values which did not include any species.

In order for the data to be usable in logistic Machine learning algorithms, we preprocessed the raw data and cleaned it by removing all NaN values in the dataset. I also dropped survey dates, removed duplicate data points, and used encoding techniques to convert non-numerical data to numerical data. The data was then randomly split into training and testing sections with 70% of the data used for training and 30% of the data used for testing. By training our models on data that only included instances of species, we were able to obtain predictions with higher accuracy, as the models would predict NaN values had they been trained with NaN values included.

This data set contains the following interpolated lake characteristics data of twelve lakes: Distance from shore (feet), Depth (feet), Bottom Hardness (relative units), Vegetation (relative units, indicating vegetation height), Longitude, and Latitude.

#### **4.4. Algorithms**

I selected three machine learning algorithms and compared their prediction accuracies: Random Forest (RF), Logistic Regression, and Gaussian Naive Bayes. With these models, the features (location input) I used were: Distance from shore, Depth, Bottom Hardness, Vegetation, Longitude, and Latitude. The labels (species output) that the model tried to predict were: Variable-leaf watermilfoil, Eurasian watermilfoil, Spiny waterflea, Asian clam, Other (any other species), and NaN (no species).

##### **Random Forest Classifier**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Random Forest algorithm is a supervised learning algorithm used in classification and regression problems, based upon decision trees. It is a way of constructing and then averaging multiple decision trees, as well as correcting the overfitting that often occurs when using a decision tree because of which, it generally yields more accurate results. Random forests are frequently used in business models due to their ability to make predictions based on a wide range of data.

A decision tree algorithm follows a hierarchical, tree-like model in order to make predictions. Based on significant features, trees can also be split into subtrees. A tree consists of several parts: root, internal, and leaf nodes, as well as branches. Root nodes

are the starting points of the tree. Internal nodes are used to make decisions based on specific criteria and have multiple branches. Leaf nodes are the final output of decisions at the bottom of the tree. Branches are connections in the tree between nodes. In order to make predictions, the model starts from the root node and follows the branches and leaf nodes, making decisions based on criteria in the tree.

### **Logistic Regression:**

Logistic regression is an algorithm based on the logistic function that predicts the relationship between two data factors. This algorithm is particularly useful in predicting two or more values using features that do not have a linear relationship.

It is frequently used in machine learning, especially within the medical field. Previously it has been used in medical scales to assess the severity of a patient's condition, or in detecting the risk of diseases.

The equation for logistic regression is:

$$Y = 1/(1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)})$$

Where Y is the value we are predicting, the x values are the inputs—feature vectors, and each  $\beta$  is a coefficient.

Using the logistic function

$$P(x) = 1/(1+e^{-Y})$$

For multi-target models, variables such as X, Y, etc. are used similarly to how they appear in a linear regression model. The logistic function can be used as an activation function to predict the probability of the label.

There are mainly three types of Logistic Regression:

**Binomial:** Binomial logistic regression has a dichotomous dependent variable. It means there can be only two possible types of dependent variables, such as 0 or 1, Yes or No, etc.

**Multinomial:** Multinomial logistic regression extends the approach for situations where the independent variable has more than two categories. Multinomial logistic can be applied to either ordered or unordered outcomes, such as "cats", "dogs", or "sheep".



**Ordinal:** In contrast with multinomial, ordinal logistic is only for ordered outcomes, such as "low", "Medium", or "High".

### **Gaussian Naive Bayes Classifier**

Gaussian Naive Bayes (GNB) is a classification technique used in Machine Learning based on the probabilistic approach and Gaussian distribution. Gaussian Naive Bayes assumes that each parameter (also called features or predictors) has an independent capacity of predicting the output variable. Bayes' theorem states the following relationship, given class variable  $y$  and dependent feature vector  $x_1$  through  $x_n$

$$P(y|x_1, \dots, x_n) = (P(y) * P(x_1, \dots, x_n|y))/P(x_1, \dots, x_n)$$

Using the naive conditional independence assumption that

$$P(x_i|y) = P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

we can use the Maximum A Posteriori (MAP) estimation to estimate  $P(y)$  and  $P(x_i|y)$ . The former is then the relative frequency of class  $y$  in the training set.

After using the database to train each respective model, results were calculated regarding the accuracy of the model's predictions. The accuracy of the models was calculated using the mean square error (MSE). The MSE is calculated through taking the absolute value of the square of the differences between each predicted value and the correct value, finding the sum, then dividing it by the total number of values.

The equation for the MSE is

Using the MSE, we could find the percentage of our models' predictions that were correct.

My models were tested with unbalanced (sometimes called imbalanced) data. Unbalanced data indicates that the amount of data points accessible for each class varies. For example, if there are two classes, balanced data means 50 percentage points for each class. Slight imbalance is not a concern for most Machine Learning approaches. As a result, if one class has 60% of the points while the other has 40%, there should be no noticeable performance reduction. Only when the imbalanced datasets of machine learning are extreme (i.e. 90% for one class and 10% for the other), would typical optimization parameters or performance metrics be ineffective and require adjustment. The dataset used in this research was considered unbalanced. For example: there were 73,525

instances of Variable-Leaf Watermilfoil, but only 149 instances of Asian clam.

## **5. Conclusion**

The objective of this study has been to forecast the likelihood of invasive species occurrence through the utilization of various AI models. I constructed three models using random forest, logistic regression, and Gaussian NB. In the study, it is evident that AI models serve as robust instruments for accurately predicting the probability of invasive species occurrences. Among the triad of models evaluated, Gaussian NB stands out with the greatest potential as an invaluable resource for steering proactive management and conservation endeavors.

## **6. Limitations and Future Work**

My study is not without limitations:

(1) Machine learning depends on labeled data, but accessing such data in biology and conservation is a challenge. Due to using a single database, there are limitations in the way my models have been trained. For instance, one algorithm may perform better for a smaller dataset than another; and

(2) all my species study areas are confined to the Adirondack region in Upstate NY. For future work, I would apply algorithms in more species and in more geographical conditions with differently distributed features; and

(3) the data was gathered in a two-year time span from 2018 to 2019, making it limited when taking into account how populations change over a longer period of time. Taking population numbers in a small time frame means the models may make unreliable predictions when used to predict species several years from now.

## **Acknowledgments**

I thank Keerthana Gurushankar, a Computer Science PhD student at Carnegie Mellon University, for her mentorship, invaluable advice, and constant guidance.

## References

1. M. Vilà, J. L. Espinar, M. Hejda, P.E. Hulme, V. Jarošík, J.L. Maron, P. Pyšek, Ecological impacts of invasive alien plants: A meta-analysis of their effects on species, communities and ecosystems. *Ecology Letters*. **14**, 702–708 (2011).  
<https://doi.org/10.1111/j.1461-0248.2011.01628.x>
2. L. Pejchar, & H. A. Mooney, Invasive species, ecosystem services and human well-being. *Trends in Ecology and Evolution*. **24**, 497–504 (2009).  
<https://doi.org/10.1016/j.tree.2009.03.016>
3. C. R. Largiadèr, Hybridization and introgression between native and alien species. In: Nentwig W. (eds) Biological Invasions. *Ecological Studies (Analysis and Synthesis)*. **193**. Springer, Berlin, Heidelberg. (2008) doi:[https://doi.org/10.1007/978-3-540-36920-2\\_16](https://doi.org/10.1007/978-3-540-36920-2_16). [Crossref], [Google Scholar]
4. S. Kumschick, M. Gaertner, F. Vilà, J.M. Essl, P. Jeschke, A. Pyšek, S. Ricciardi, T.M. Bacher, J.T.A. Blackburn, T. Dick, P.E. Evans, I. Hulme, A. Kühn, J. Mrugała, W. Pergl, D. M. Rabitsch, A. Richardson, M. Sendek. Winter, Ecological impacts of alien species: quantification, scope, caveats, and recommendations. *BioScience*. **65**, 55–63 (2015).
5. L.R. Lazzaro, G. Bolpagni, R. Buffa, M. Gentili, A. Lonati, A.T. Stinca, R. Acosta, et al., Impact of invasive alien plants on native plant communities and Natura 2000 habitats: State of the art, gap analysis and perspectives in Italy. *Journal of Environmental Management*. **274**, 111–140 (2020). doi: <https://doi.org/10.1016/j.jenvman.2020.111140>. [Crossref], [PubMed], [Web of Science ®], [Google Scholar]
6. D.M. Viciani, D. Vidali, R. Gigante, M. Bolpagni, A.T.R. Villani, R. Acosta, et al., A first checklist of the alien-dominated vegetation in Italy. *Plant Sociology*. **57(1)**, 29–54 (2020). doi: <https://doi.org/10.3897/pls2020571/04>. [Crossref], [Google Scholar]
7. A. Ricciardi, and H. J. MacIsaac, Recent mass invasion of the North American Great

- Lakes by Ponto-Caspian species. *Trends in Ecology & Evolution*. **15**, 62–65 (2000).
8. J. Cucherousset, and J. D. Olden, Ecological impacts of nonnative freshwater fishes. *Fisheries*. **36**, 215–230 (2011).
9. B.D. Leung, M. Lodge, D. Finnoff, J. F. Shogren, M. A. Lewis, and G. Lamberti, An ounce of prevention or a pound of cure: Bioeconomic risk analysis of invasive species. *Proceedings of the Royal Society of London Series B: Biological Sciences*. **269**, 2407–2413 (2002).
10. K.S. Pagnucco, G. A. Maynard, S. A. Fera, N. D. Yan, T. F. Nalepa, and A. Ricciardi, The future of species invasions in the Great Lakes-St. Lawrence River basin. *Journal of Great Lakes Research*. **41**, 96–107 (2015).
11. A. M. Kramer, G. Annis, M.E. Wittmann, W. L. Wittmann, R. Chadderton, S. Edward, D.M. Lodge, L. Mason, Lacey, D. Beletsky, C. Riseng, J.M. Drake. Suitability of Laurentian Great Lakes for invasive species based on global species distribution models and local habitat <https://doi.org/10.1002/ecs2.1883> (2017)
12. A.Sabat-Tomala, E. Raczko, and B. Zagajewski. Comparison of support vector machine and Random Forest Algorithms for invasive and expansive species classification using airborne hyperspectral data. *Remote Sensing*. **12 no. 3**, 516 (2020).  
<https://doi.org/10.3390/rs12030516>
13. G. De'ath, and K. E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*. **81**, 3178–3192 (2000).
14. B.M. Steele, Combining multiple classifiers: an application using spatial and remotely sensed information for land cover mapping. *Remote Sensing of Environment*. **74**:545–556 (2000).
15. A. Guisan, and W. Thuiller, Predicting species distribution: offering more than simple habitat models. *Ecology Letters*. **8**, 993–1009 (2005).
16. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and regression trees. *Wadsworth and Brooks/Cole*, Monterey, California, USA. 1984.
17. M.E. Eiswerth, S.G. Donaldson, and W.S. Johnson, Potential environmental impacts and economic damages of Eurasian watermilfoil (*Myriophyllum spicatum*) in western Nevada and northeastern California. *Weed Technology*. **14(3)**, 511-518 (2000).
18. K.I. Cullis, and G.E. Johnson, First evidence of the cladoceran *Bythotrephes cederstroemi* Schodler in Lake Superior. *Journal of Great Lakes Research*. **14(4)**, 524-525 (1988).

## **Appendix A**

Dataset used in this research come from surveys for the following twelve lakes in Upstate NY :

**Lake Name**

**County**

**Township**

Loon Lake,	Warren County,	Piercefield
Hadlock Pond,	Washington County,	Fort Ann
Little Wolf Pond,	Franklin County,	Tupper Lake
Long Lake,	Oneida County,	Forestport
Sixth Lake,	Hamilton County,	Inlet
East Stoner Lake,	Hamilton County,	Arietta
Long Pond,	St. Lawrence County,	Piercefield
Clear Pond	St. Lawrence County,	Parishville
Horseshoe Pond,	St. Lawrence County,	Piercefield
Fifth Lake,	Hamilton County,	Inlet
Clear Pond	Lewis County,	Croghan
Courtney Pond,	Essex County,	North Hudson

## **Appendix B**

### **Technical skills**

The AI model utilized in the study was created using the Python programming language and related Python libraries. The model was coded in JupyterLab and the libraries used in the project were Pandas, Scikit-Learn, Matplotlib, and Seaborn.

Pandas is a Python data analysis library. The data referenced in the project was stored in a comma-separated values (CSV) file, which was then able to be read by importing and calling the Pandas library.

Scikit-Learn is a Python machine learning library that features multiple classification, regression, and clustering algorithms. It allows for the easier training and integration of various machine learning algorithms, which were used in the study to train the various models.

Matplotlib is a Python visualization library that allows for the embedding of various plots. It is used alongside a dataset in order to provide 2D and 3D visualizations of the data in graphs such as line plots, histograms, contour plots, and scatter plots.

Seaborn is another Python visualization library for plotting statistical graphs based on data. It offers a visualization between variables through various types of plots such as

distribution and relational plots.